magnetochemistry

MDPI

Article

Cross-Library Audit of DFT Magnetic Moments for Curie-Temperature Ranking: An Empirical Residual-Gating Protocol

Jun-Feng Li, Qian Chen *, Lei Zhou, Yang Du and Zhen Liang

Advanced Technology & Materials Co., Ltd., Beijing 100081, China
* Correspondence: chenqian@atmcn.com

Abstract

Reliable magnetic-materials screening requires more than a model that performs well within one computational catalogue. We develop and audit a composition-anchored library residual (CALR) protocol for transferring density-functional-theory (DFT) magnetic-moment information across heterogeneous databases. CALR harmonizes moment percentiles within each library, fits a composition-only ridge anchor, and admits a bounded moment residual only when a nested, composition-disjoint bridge test supports improved rank correlation. AFLOW and JARVIS moments are evaluated against experimental Curie temperatures from NEMAD, with a frozen Materials Project snapshot as a third source. On 484 AFLOW-JARVIS bridge compositions, neither directional gate opens: cross-fitted changes in Spearman correlation are +0.0079 (95% interval −0.0027 to 0.0187) and +0.0060 (−0.0125 to 0.0238). The Materials Project-to-JARVIS bridge gives Δrho = 0.0101 (−0.0036 to 0.0250), also returning the composition anchor. Ungated transfer improves one direction but produces negative transfer in the reverse; CORAL and density-ratio weighting underperform the anchor. A fixed-hash semi-synthetic control shows that CALR can activate for a known transferable residual and close for zero, non-transferable, or reversed residuals, although finite-sample false openings remain. CALR is therefore an auditable diagnostic for cross-library magnetic screening and negative-transfer risk, not formal risk control or validated permanent-magnet discovery. We re-ran the Materials Project analysis with current cell atom counts (nsites), traced every aligned key to a selected DFT identifier and experimental DOI, froze a percentile map on common compounds before gating, held out chemical families from training, resampled element-set groups through the full fit/select pipeline, and compared CALR with same-budget target-domain models. The MP gate remains closed after the nsites correction. We state a single protected deployment estimand: Spearman ρ of the frozen ranking versus experimental Tc on composition-disjoint target-library keys. Structure-resolved and aggregation sensitivities leave the headline $|M|$–Tc Spearman near 0.43. Family-grouped cross-validation lowers the composition-ridge OOF ρ from 0.697 to 0.461. The 484-key bridge is underpowered for the observed residual (forward 80% power requires $\Delta\rho \approx 0.015$–0.020). Materials Project reuses 100% of the NEMAD labels already seen in the AFLOW/JARVIS workflow. CALR still equals the composition ridge on every real direction; we state quantitative conditions under which its extra cost would be justified, and the independent library experiment that would be required to claim a practical benefit.

Keywords: magnetic materials databases; DFT magnetic moments; Curie temperature; cross-library transfer; negative transfer; materials informatics; rare-earth magnets

Academic Editor: Roberto Zivieri

Received: 20 August 2026
Revised: 14 September 2026
Accepted: 17 September 2026
Published: 21 September 2026

1. Introduction

Rare-earth permanent magnets combine a high Curie temperature, a large saturation magnetization, and a uniaxial anisotropy strong enough to support coercivity after processing [1–3]. The industrially relevant figures of merit are remanence Br, coercivity Hcj, and energy product (BH)max, not the DFT spin moment of a relaxed cell. Those device quantities depend on grain-boundary chemistry, texture, and the 2-14-1 or 1-5/2-17 prototype [3–7]. None of those ingredients is encoded in a per-atom moment.

Public computational libraries report spin moments for tens of thousands of lanthanide compounds, including AFLOW, the Materials Project and JARVIS [8–12]. High-throughput permanent-magnet studies instead emphasize phase stability, magnetization and magnetocrystalline anisotropy, because no single moment label defines a hard magnet [6,7,13,14]. A parallel machine-learning literature predicts experimental Tc from composition [15]. These are distinct tasks: success at reproducing a zero-Kelvin DFT moment does not establish transfer to experimental Tc, K1, anisotropy field or a processed-magnet endpoint. Magnetic functionality is likewise application- and structure-dependent across colossal-magnetoresistance films, photomagnetic molecular nanomagnets and magnetorheological fluids and elastomers [16–18].

A second public layer now exists. The Northeast Materials Database (NEMAD) extracts experimental compositions, transition temperatures, and, where the source paper states them, coercivity and remanence from the literature [19]. The associated machine-learning tables contain 15,577 unique ferromagnetic compositions with a mean Tc. The full website dump, described as 67,573–67,584 records, also stores coercivity (Oe in the data dictionary) and remanence, but those fields are sparse and are released as unnormalized strings.

This article audits one public AFLOW lanthanide extract against official NEMAD and turns the audit into an empirical cross-library residual-gating protocol. We ask whether the extract contains 2-14-1-like chemistry under a stated composition window, whether DFT $|M|$ adds information beyond composition, and whether a library-aware ranker transfers to an unseen catalogue. CALR uses a composition ridge as an anchor, calibrates $|M|$ to a within-library percentile, and permits at most half of the augmented residual only when a nested paired-bridge $\Delta\rho$ bootstrap interval lies above zero; otherwise, it forces an exact composition-only fallback. This is conditional evidence for residual acceptance relative to the anchor, not a guarantee that the anchor is valid under every shift and not abstention from producing a ranking. We do not claim a new magnet and do not treat Tc as a substitute for Br or Hcj.

The workflow has four roles. Source-library labels fit the composition anchor and augmented model. Paired bridge compositions supply target-library moment representations and choose whether the residual is accepted without using any library-unique target label. The frozen decision is then evaluated on composition-disjoint target-library rows whose NEMAD labels were withheld within that direction. AFLOW and JARVIS provide 663 and 744 labelled composition keys, respectively, with 484 paired bridge keys and 260 JARVIS-only or 179 AFLOW-only external targets. The MP reanalysis provides 459 labelled source/bridge keys and 285 JARVIS-only external targets; its NEMAD endpoints were already present in the earlier workflow. If the bridge gate closes, the method rejects the residual and outputs the source-trained composition ranking unchanged. Global Spearman ρ is the activation estimand; top-k enrichment is a first-stage setting-selection diagnostic and is not guaranteed by the gate.

The intended screening use of a ranked list does not change the protected estimand. A residual is accepted only when nested bridge evidence supports a positive Spearman

improvement. Precision and enrichment at k = 10, 25 and 50 remain descriptive summaries of that ranking and cannot open, close or override the gate.

CALR is deliberately narrower than three adjacent method families. Negative-transfer detection estimates when reuse of a source model or representation is harmful [20,21]. Selective predictors with a reject option may withhold an output [22], whereas learn-then-test procedures calibrate a prespecified loss to obtain formal risk control [23]. CALR transfers no learned representation, never refuses the composition ranking and provides no distribution-free risk guarantee. Its contribution is an applied cross-library protocol: composition-disjoint bridge cross-fitting, within-library percentile harmonization, a bounded descriptor residual and exact return to a named anchor. We therefore claim a materials-database evaluation and gating protocol, not a new general theory of safe transfer.

The main evaluation is model-independent for $|M|$, leakage-controlled for the composition model and library-disjoint for transfer. Each bridge score is source-model OOF, and its residual and penalty settings are selected in a nested bridge fold that excludes that row's Tc. Library-unique target labels are read only after the gate and deployment parameters are frozen. We additionally include a fixed-hash semi-synthetic positive control, paired with a zero-residual control, to test whether the same frozen gate can activate when a transferable residual is known to exist. This control is an algorithm calibration test and is not used as materials evidence. Legacy split and privileged-signal diagnostics are retained only in the Supplementary Information because neither is needed for the central transfer comparison.

2. Materials and Methods

2.1. AFLOW Extract

Records were retrieved on 3 August 2026 through the public AFLUX interface with fields compound, auid, aurl, species, spin_atom, spin_cell, volume_atom, natoms, space-group_relax, Pearson_symbol_relax, and aflow_prototype_label_relax. Separate species(R) queries were issued for Ce–Lu. Multi-rare-earth records were merged and deduplicated by auid, leaving 14,482 rows. The regression target is $y = |\text{spin\_atom}|$ in $\mu B\ \text{atom}^{-1}$. A record is called magnetic if $y > 0.05\ \mu B\ \text{atom}^{-1}$ (7139 records, 49.3%). The structure-label census identifies 2-14-1 as Pearson tP68 and space group $P4_2/mnm$ (No. 136); it finds zero rows. The reported composition-window and structure-label counts are distinct tests.

2.2. NEMAD Tc Table

The processed ferromagnetic table FM_with_curie.csv from NEMAD-MagneticML v1.0 (Zenodo 10.5281/zenodo.17042814; GitHub v1.0 (sumanitani/NEMAD-MagneticML) sumanitani/NEMAD-MagneticML) was used as the experimental Tc source [19]. Each of the 15,577 rows is a unique normalized composition with Mean_TC_K. The table has 117 columns of elemental fractions and composition descriptors. It has no coercivity or remanence column (the Br field is bromine).

2.3. Alignment

Atomic fractions from AFLOW composition keys and from NEMAD New_Column_Concatenated (fallback: Material_Name) were renormalized and rounded to four decimals. Two compositions match if those keys are identical. The key can merge distinct polymorphs and can split experimental name strings that encode doping or processing. The estimand is therefore ranking on the exact-key overlap population, not transfer between structure- or phase-matched material specimens. In this extract, 14,482 AFLOW rows collapse to 6746 keys; 2326 keys (10,062 rows) are collisions, with at most 112 auid on CeO_2. The official NEMAD dump collapses 33,668 rows to 21,926 keys; 4482 keys (16,220 rows) are collisions, with at most 116 rows on elemental Gd. Headline ranking uses the cell-wise

maximum |M| and median Tc. Sensitivity runs use three-, four- and five-decimal keys and maximum or median |M|; all six runs retain 663 Tc pairs and top-10 precision 0, with ρ = 0.42–0.43 and E@100 = 1.88–2.18 (Supplementary Table S4). A 2-14-1-like flag requires Fe, B and {Nd, Pr, Dy, Tb, Ce, Sm}, excludes oxygen, and applies fraction windows 0.65–0.88 Fe, 0.03–0.10 B and 0.08–0.22 rare earth. A 1-5-like flag requires Sm and Co, excludes O and B, and applies 0.12–0.22 Sm and 0.72–0.90 Co. Subsets otherwise use element presence. Core rare earths are Nd, Sm, Pr, Ce, Dy, and Tb. The processed FM_with_curie.csv table is aligned the same way as a confirmation set.

Because a rounded composition key can merge distinct structures, magnetic states and experimental samples, we emit a row-level ledger for all 663 AFLOW–NEMAD Tc keys (results/match_provenance.csv). Each key is traced to the selected AFLOW cell (auid, aurl, space group, Pearson symbol, natoms, maximum |spin_atom|), a JARVIS jid and space group when present (484/663), an MP mpid when present (371/663), and a primary NEMAD DOI and material name. Of the 663 keys, 532 (80.2%) collide on the DFT side, the experimental side, or both: 386 keys have more than one AFLOW cell, 423 have more than one NEMAD row, 243 have more than one distinct parsed Tc, and 412 have more than one DOI. Headline ranking still uses maximum |M| and median Tc, but those aggregates are no longer anonymous: the contributing calculation and literature records are recoverable. The evaluation therefore remains a catalogue-key ranking audit, not a structure-matched or specimen-matched property transfer.

Of the 663 labelled AFLOW–NEMAD Tc keys, 277 (41.8%) have a unique AFLOW cell, 590 (89.0%) have a unique Pearson symbol, and 584 (88.1%) have a unique space-group number. On colliding keys, the median |M| spread is only 0.0029 μB atom^{-1}, so many collisions are near-duplicates rather than chemically distinct polymorphs. Replacing the headline maximum |M| by the median changes Spearman ρ from 0.428 to 0.417 ($\Delta\rho = -0.010$); a majority-Pearson representative is essentially unchanged (ρ = 0.428). Restricting to unique-Pearson or unique-space-group keys yields ρ = 0.418 (n = 590) and 0.417 (n = 584). Collision-free keys alone give ρ = 0.396 (n = 277) but raise E@10 from 0 to 2.97 because the headline top-10 is dominated by colliding oxides. CeO_2 is the canonical collision: 112 AFLOW cells, 20 Pearson symbols, |M| from 0 to 0.410 μB atom^{-1}. Maximum |M| is therefore a screening-oriented catalogue upper bound, not a claim that the selected cell is the experimental ground state; median Tc is the central experimental report on that key. Family subsets with unique Pearson symbols remain consistent with the headline in RE–Fe (0.602 → 0.586, n = 68/58) and RE–Co (0.666 → 0.620, n = 71/60); oxides and pnictides stay weakly correlated. The four-decimal key therefore compresses polymorphs, but it does not manufacture the headline ρ ≈ 0.43.

2.4. JARVIS Snapshots and Three-Way Comparison

The complete JARVIS-DFT dft_3d archives dated 24 September 2025 (Figshare file 64391379; 93,902 records) and 12 December 2022 (file 38521619; 75,993 records) were downloaded from the official JARVIS Figshare distribution [11,12]. Records were retained when their parsed formula contained Ce–Lu and magmom_outcar was numeric. The score is the absolute OUTCAR cell moment divided by nat; records on the same four-decimal composition key use the maximum score. NEMAD alignment and Tc aggregation are identical to the AFLOW analysis. The newer snapshot is primary, and the older snapshot is a version-sensitivity test. A controlled three-way set is formed by intersecting the JARVIS–NEMAD pairs with the 663 AFLOW–NEMAD Tc keys; the NEMAD median Tc must agree exactly on each retained key. Each library is ranked separately on those identical targets. No DFT values are combined across libraries. JARVIS rank correlations use a 2000-replicate composition bootstrap and a 4000-shuffle permutation test. On the three-way set,

4000 composition-paired bootstrap samples estimate JARVIS-minus-AFLOW intervals for Δρ and Δenrichment while holding the common-set Tc threshold and upper-decile labels fixed; 10,000 two-sided randomizations swap the two library ranks within each composition. The cross-library tests are exploratory and use seed 20260816.

2.5. Empirical Residual-Gating Ranking

CALR is fitted in one direction at a time. Input features are elemental fractions, a residual fraction for elements unseen in the labelled basis, the empirical percentile of |M| among all 6746 AFLOW or 13,712 JARVIS compositions, and products of that percentile with total rare-earth, 3d-transition-metal and Fe + Co fractions. Full-library moments calibrate the feature without using Tc. A ridge ensemble predicts log(1 + Tc). Regularization is selected by four-fold inner validation within five composition-stratified source folds; 32 bootstrap fits provide the ensemble mean and standard deviation [24,25]. Euclidean distance to the nearest labelled elemental-fraction vector measures support coverage. Moment-residual weights are {0, 0.10, 0.25, 0.50}; uncertainty and distance penalties are selected from {0, 0.02, 0.05, 0.10}. Stage 1 proposes settings by mean enrichment at k = 10, 25 and 50 plus 0.5ρ, subject to a maximum 0.01 loss of ρ relative to the corresponding unpenalized score. Stage 2 then accepts or rejects the proposed residual solely by the cross-fitted Spearman interval. The first stage is a conservative candidate-setting filter, not a top-k guarantee. The term empirical gate is used deliberately. The reported bridge intervals are conditional on the completed cross-fitted scores and selected settings because the bootstrap does not repeat source fitting, fold construction or hyperparameter selection; they are not nominal coverage or formal risk-control intervals.

The protected deployment objective is unique: Spearman ρ between the frozen score and experimental Tc on composition-disjoint target-library keys. Stage 1 may still propose residual/penalty settings using a composite of early enrichment and ρ, but Stage 2 activation, stress-grid harm, and all claims of residual acceptance use only the nested Spearman interval. A setting that improves E@k while the Spearman lower bound is ≤ 0 is rejected. A setting that passes the Spearman gate is not declared top-k safe. This aligns fitting, gating and reporting with one screening-relevant ranking estimand rather than mixing early-retrieval and global correlation as if they were interchangeable decision criteria.

Cross-library activation uses the 484 paired bridge keys. In each source fold, a bridge composition is excluded from ridge fitting and predicted with the target library's moment percentile. The bridge rows are then divided into five stratified gate folds. For each held-out gate fold, residual weight, uncertainty penalty, distance penalty and their robust scaling constants are selected using only the other four folds; the held-out rows are scored without their Tc entering hyperparameter selection. Concatenating the five held-out scores yields one nested cross-fitted bridge ranking. The 1500-replicate paired bootstrap resamples composition rows from these completed cross-fitted scores; it does not repeat source fitting, fold construction or hyperparameter selection inside each replicate. CALR activates settings refitted on all bridge rows only when the conditional 95% lower bound of nested Δρ is greater than zero. Otherwise, all transfer terms are zero. Final targets are the 260 JARVIS keys absent from AFLOW and the 179 AFLOW keys absent from JARVIS; their labels are read only after the gate and deployment parameters are frozen. A 3000-replicate paired bootstrap compares ridge, ungated, gate-only and full CALR scores. The gate is an empirical decision rule for global rank correlation, not a formal error-control guarantee or protection of enrichment at a chosen k. A post-review diagnostic ablation replaces Stage 1 by Spearman-only setting selection while retaining all folds, seeds and the Stage 2 rule; it is reported as a fragility check rather than independent confirmation.

The exploratory deployment audit does not refit on JARVIS Tc labels. It reuses the AFLOW-to-JARVIS source rows, ridge regularization, bootstrap seed and nested bridge decision, while JARVIS Tc keys are used only to exclude already labelled compositions. Rows must contain Fe or Co, have $|M| > 0$, exclude O/F/Cl/Br/I and have non-positive JARVIS formation energy. Because the forward gate closes, the deployed score is exactly the AFLOW-trained composition ridge. The JARVIS ehull field is retained but not filtered because its provenance is unresolved. This audit illustrates what the fallback would rank; it is not a candidate-discovery or property-prediction result.

2.6. Fixed-Hash Gate-Activation Control

Gate sensitivity was tested with a fixed, leakage-controlled semi-synthetic benchmark implemented in run_calr_positive_control.py. The 744 JARVIS–NEMAD composition keys retained their real elemental fractions and full-library JARVIS moment percentiles. SHA-256 of each canonical composition with the fixed salt CALR-PC-v1 assigned buckets 0–5 to training, 6–7 to gate selection and 8–9 to the final test, giving 453, 146 and 145 non-overlapping compositions. The control endpoint was defined before model fitting as $\log(1 + T^*) = 5.5 + 1.5xFe + Co - 0.5xO + \beta(qM - 0.5) + \varepsilon$, where qM is the JARVIS moment percentile and ε is deterministic key-specific Gaussian noise with standard deviation 0.15. The positive control used $\beta = 1.0$; the matched zero-residual control used $\beta = 0$ while retaining the same compositions, partition and noise.

Composition and augmented ridge models were fitted only on the training partition. Within the gate partition, five folds selected the residual and penalties on four folds and scored the fifth; the 1500-replicate paired bootstrap lower bound of concatenated cross-fitted $\Delta\rho$ determined activation exactly as in the real-data analysis. The final test endpoint was not accessed until the gate decision and deployment settings were fixed. This benchmark tests algorithm operation under a known transferable residual. It does not validate real moment-to-Tc transfer, a candidate composition or the semi-synthetic temperature scale.

2.7. Frozen Materials Project Transfer, Method Benchmark and Stability Queries

The third-source analysis uses Matminer's frozen mp_nostruct_20181018 Materials Project snapshot (83,989 records); its SHA-256 is 48481abc390e0772defbbfa68d7db3b9bdef2519d407b82a581f2298f7df080f. The pandas split-JSON records are reduced to the same four-decimal composition key. Total cell magnetization mu_b is converted to $|\mu B|$ atom^{-1} by dividing by the sum of parsed formula amounts; duplicate compositions retain the maximum moment, matching the AFLOW/JARVIS aggregation. Full-snapshot MP moments define the percentile before NEMAD alignment. The resulting 459 MP–NEMAD pairs are the source and bridge to JARVIS; 285 JARVIS–NEMAD keys absent from MP form the composition-disjoint target-library external set with the same NEMAD label source. All MP-labelled keys occur in JARVIS, so the reverse direction has no MP-only external target and is reported as not evaluable. The within-direction gate and external-label isolation are unchanged from Section 2.5; method-level independence from the earlier AFLOW/JARVIS endpoint population is not claimed.

Matminer mu_b is a cell total while the 2018 snapshot stores a reduced formula and no nsites field. We joined current Materials Project OPTIMADE nsites by mpid for 63,904/83,989 snapshot records (76.1%); 20,085 identifiers are absent from the current catalogue and remain formula-normalized. The median nsites/formula-atom ratio is 2.0 (mean 3.05; 86.2% integer multiplicity), so the original conversion systematically inflates per-atom moments whenever the computational cell contains more than one reduced formula unit. After key aggregation, formula-normalized and nsites-normalized moments still correlate at Spearman $\rho = 0.981$ (n = 46,686). Re-running the entire nested MP-to-

JARVIS protocol on the nsites-corrected labelled set (n = 448 vs. 459) leaves the gate closed: nested $\Delta\rho = 0.0093$ (95% interval -0.0009–0.0191) vs. 0.0101 (-0.0036–0.0250) originally. External CALR equals the corrected composition-ridge fallback ($\rho = 0.732$; the original formula-normalized result was $\rho = 0.742$). Current nsites are a post-2018 proxy, not the frozen 2018 cell; the correction is therefore propagated as a sensitivity that preserves the gate decision rather than as a claim that every 2018 cell was recovered.

The common-split benchmark compares composition ridge, ungated moment transfer, CORAL covariance alignment, logistic domain-propensity density-ratio weighting, pooled source-plus-bridge training, CALR and uncertainty filtering at 80% retained coverage. Hyperparameters use source or permitted bridge labels only; external labels enter final ranking metrics and 1500-replicate paired intervals. A separate 48-case stress grid retains the fixed 453/146/145 hash partition and crosses four residual strengths (0, 0.25, 0.75, 1.25), three deterministic noise levels (0.05, 0.15, 0.30) and four regimes: transferable, zero, training-only/non-transferable and sign-reversed residuals. Outcomes are gate opening, false opening/closing, exact fallback and held-out regret relative to the composition ridge.

Candidate stability is reconciled independently with the frozen MP snapshot, the current Materials Project OPTIMADE _mp_stability response and OQMD formation-energy API. Queries first use the candidate element set and then require the exact composition key. For current MP, the fixed thermo priority is mixed GGA/GGA + U/r2SCAN, then r2SCAN, then GGA/GGA + U; the lowest energy-above-hull exact entry is retained. OQMD likewise retains the lowest reported stability exact entry. These database values are not new calculations; missing matches do not imply instability, and no magnetic-order or spin-orbit anisotropy calculation is inferred from them.

2.8. Ranking and Traps

For each aligned composition, we store the maximum AFLOW $|M|$ and the median NEMAD Tc, Hc or Br. A usable Tc is a non-empty official Curie field after NA stripping. Spearman's ρ is computed on those pairs. Uncertainty for ρ uses a 2000-replicate bootstrap percentile interval and a 4000-shuffle permutation test (seed 20260813). High-Tc labels use three cuts: the subset 90th percentile, 300 K, and 600 K. Precision@k and enrichment@k = (precision@k)/base rate are reported for $k \in \{10, 25, 50, 100\}$ when k does not exceed the subset size. Counts recovered at k are compared with a hypergeometric null that draws k items from a fixed positive class. The three Tc thresholds and four list depths form 12 exploratory tests; nominal and Bonferroni-adjusted p values are distinguished, and Wilson intervals are reported for precision and enrichment. No universal screening-utility threshold was prespecified. A trap is a composition with $|M| \geq$ subset 90th percentile and Tc < subset median. Coercivity strings are converted to kA m^{-1} only when a unit token is present (Oe, kOe, A m^{-1}, kA m^{-1}, T as $\mu_0 H$, G). Ranges use the midpoint of the first two numbers; inequalities, multi-valued strings and unitless cells are excluded. Remanence is converted to tesla; emu g^{-1} is discarded. Parser regression cases cover single kA m^{-1} values, kOe ranges, inequalities and multi-condition strings.

2.9. Composition Baselines

The simple scores are the Fe atomic fraction, Co atomic fraction and their sum on the four-decimal composition key. Because many compositions share the same score, precision and enrichment at a cutoff use the expected number of positives under uniform ordering within the tied group. The learned baseline uses the 59 elemental fractions observed in the 663 aligned Tc compositions and predicts log(1 + Tc) by ridge regression. We generate out-of-fold predictions in 20 repeated, stratified five-fold outer splits. In every outer training set, the regularization parameter is selected from {0.01, 0.1, 1, 10, 100} by four-fold inner

validation using mean absolute error; centering and scaling are also fitted only on the training fold. The reported ranking uses each composition's mean prediction across its 20 out-of-fold appearances. The Tc p90 threshold is defined once on the aligned audit set to keep the target class identical across all ranking methods; it is not used as a regression input.

2.10. Official NEMAD Dump

The Downloads page UI asks for a login; the documented file endpoint returned HTTP 200 without credentials on 13 August 2026 and a 4,525,021-byte file of 33,668 rows. Coercivity and remanence were counted as present if the cell was non-empty after stripping null/NA. Unit conversion for ranking used the parser above. The paper license of the NEMAD article is CC BY-NC-ND 4.0; the website schema.org block lists CC BY 4.0. We analyze and cite the dump and do not redistribute a derived full library. The website headline (67,584) is recorded as a discrepancy, not as the analysis n.

2.11. Official NEMAD Anisotropy Dump

The file used here has 8,233,705 bytes and 33,916 rows. Property cells are often Python-literal (Python 3.10) dictionaries with Value, Units, and Temperature. Slice counts use Material_Name element parsing plus name-string tokens (NdFeB, SmCo). Magnet_type and Material_Type are taken as released.

2.12. Direct Endpoint Parsing

compute_direct_endpoints.py accepts only plain chemical formula names for exact alignment. K1 is converted to absolute MJ m^{-3} from J m^{-3}, kJ m^{-3}, MJ m^{-3} and explicitly scaled cgs units. Anisotropy field is converted to μ0Ha in tesla from T, mT, Oe, kOe, G and A m^{-1}-derived units. Saturation magnetization is converted to Js = μ0Ms in tesla from field-density units and emu cm^{-3}; emu g^{-1} and μB per formula unit are excluded because density or formula information is not supplied consistently. One median endpoint value is retained per exact composition key. The output and accepted-unit audit are in results/direct_magnet_endpoints.json and results/direct_magnet_endpoint_pairs.csv.

2.13. Auxiliary Libraries

Novomag magdb.json (Zenodo 10.5281/zenodo.17399362) is JSON Lines of ComputedStructureEntry objects. MAGNDATA entry counts were read from the public search page on 13 August 2026 (2414 entries).

2.14. Robustness Analyses

Percentile comparability. AFLOW uses per-atom spin_atom, JARVIS divides the OUTCAR cell moment by nat, and MP originally divided cell mu_b by the reduced formula. On the 371 three-way labelled keys, Spearman correlations of the three moment fields equal those of their within-library percentiles (AFLOW–JARVIS 0.411; MP–JARVIS 0.719; MP–AFLOW 0.481), as expected for a strictly monotone transform. Percentile conversion therefore does not create a new physical unit. We froze a QQ map from MP to JARVIS percentile space on 195 hash-even common keys, applied it to the 176 held-out common keys and to the full MP source, and only then re-ran the nested gate. Held-out alignment did not improve (mapped $\rho = 0.677$ vs. native 0.692). The frozen map still closed the MP-to-JARVIS gate ($\Delta\rho = 0.0094$, interval -0.0033–0.0212). Calibration was frozen before reading the 285 external labels.

Bridge versus deployment domain. Chemical families (RE–Fe, RE–Co, oxide, pnictide, other) were excluded from source training, and the entire nested fit/select/gate pipeline was re-run. AFLOW-to-JARVIS gates remained closed for RE–Fe, RE–Co, oxide and pnictide. Holding out the heterogeneous other class (435/663 AFLOW keys; 228 source rows

retained) opened the gate ($\Delta\rho$ = 0.070, 0.041–0.103). JARVIS-to-AFLOW remained closed for every family. The 260 JARVIS-only targets split equally into near and far halves about a median composition distance of 0.236 to the bridge; the 179 AFLOW-only targets split about 0.354. A literature-source hold-out of the dominant NEMAD DOI prefix 10.1016 is not powered: that prefix accounts for 567/663 primary DOIs, leaving 96 labelled AFLOW rows. Family hold-out is therefore the powered domain-shift test, and it shows that the bridge represents the named RE–Fe/RE–Co/oxide/pnictide slices but not an arbitrary remainder class.

Chemistry-group resampling. Element-set groups (507 AFLOW groups; 563 JARVIS groups) were resampled with replacement six times, and the full nested pipeline was repeated. AFLOW-to-JARVIS remained closed in 6/6 repeats (mean nested $\Delta\rho = -0.007$). JARVIS-to-AFLOW opened in 2/6 repeats; one opening improved external ρ (0.615 vs. ridge 0.583), and one was harmful (0.546 vs. 0.572). Gate decisions are therefore stable on the original 484-key bridges and on forward resampling, but they are not invariant to reverse-direction chemistry-group resampling. These repeats used a reduced ensemble (8 bootstrap models; 400 CI replicates) and are reported as a stability diagnostic, not as a replacement for the primary 32-model/1500-replicate analysis.

Same label budget. The CALR bridge contains 484 labelled target-library compositions. A target-domain composition ridge trained on those 484 Tc values and evaluated on the library-unique externals reaches ρ = 0.793 (AFLOW-to-JARVIS, n = 260) and 0.503 (JARVIS-to-AFLOW, n = 179). CALR, which uses the same budget only to accept or reject a residual and then returns the source-trained anchor, has ρ = 0.779 and 0.524. A simple validation selector that chooses composition versus moment-augmented ridge from five-fold bridge Spearman agrees with the target-domain ridge on the forward split but selects the augmented model on the reverse split and falls to ρ = 0.456. Under the same labelled-bridge budget, CALR therefore does not outperform a target-trained composition model when transfer is mild, but it avoids the harmful selector decision that a bridge-tuned model switch makes under reverse catalogue shift. That is the operational value of the always-anchor fallback.

2.15. Structure, Power, Nesting and Claim Boundaries

Percentile tail. Within a library, the percentile is strictly monotone, so the rank of $|M|$ is preserved. The AFLOW extract top decile (n = 675) still spans 1.56–8.02 μB atom^{-1} (CV = 0.38). Neighboring windows 0.90–0.95 and 0.95–1.00 occupy 1.56–2.33 and 2.33–8.02 μB atom^{-1}; labelled keys in the top decile span 1.59–7.94. Absolute-moment differences in the high-moment tail are therefore not collapsed onto a single magnitude. Percentile conversion remains a within-library rank, not a physical unit.

Nested resampling and power. The reported bootstrap interval is stratified on cross-fitted scores; source fitting, fold creation and hyperparameter search are not repeated inside each bootstrap replicate. A reduced-ensemble diagnostic that does re-seed the full nested pipeline (8 models; 400 CI replicates; six seeds) opens AFLOW-to-JARVIS in 1/6 seeds (lower bound 0.0008) and JARVIS-to-AFLOW in 0/6. The single opening would still close under a minimum-effect rule of lower bound > 0.02. Using the observed nested standard errors (forward SE = 0.0054; reverse SE = 0.0094), the LB > 0 gate reaches 80% power only at a true $\Delta\rho$ of about 0.015–0.020 (forward) and 0.030 (reverse). The observed nested improvements (0.008 and 0.006) have analytic power of 0.31 and 0.13. At this bridge size, the analysis can reliably detect only larger effects; therefore, a closed gate should not be interpreted as evidence that a small useful residual is absent, consistent with both real gates remaining closed.

Family-grouped composition CV. Repeated random five-fold OOF Spearman for the elemental-fraction ridge is 0.697. Leave-one-family-out (RE–Fe, RE–Co, oxide, pnictide, other) falls to 0.461. Random folds therefore leak family-level stoichiometry and inflate the apparent capacity of the composition anchor to generalize beyond familiar chemistries. The deployed CALR still uses that ridge as the fallback; the family-OOF number is the more honest estimate of composition-only ranking across chemical families.

Computational-source replication versus independent validation. All 285 MP-to-JARVIS experimental endpoints were already present in the AFLOW/JARVIS NEMAD workflow (113 as prior bridge labels, 172 as AFLOW-to-JARVIS externals). The MP analysis is a third computational source with within-direction label withholding, not independent external validation of the experimental endpoint. We state that distinction quantitatively rather than as a qualitative caveat.

2-14-1 coverage. The AFLOW Ce–Lu extract contains neither an Nd2Fe14B cell nor any composition in the specified 2-14-1-like window ($n = 0$). NEMAD contains 131 unique 2-14-1-like compositions. Ranking and gating conclusions therefore do not generalize to 2-14-1 permanent-magnet screening. The REPM-relevant result in this extract is a coverage failure, not a ranking result inside that family.

When CALR is worth its cost. CALR copies the composition ridge in every real transfer direction, so no measured ranking benefit offsets the extra nested fit. We use it in place of the ridge only when all of the following hold: (i) a composition-disjoint bridge of at least 400 labelled keys exists; (ii) the nested 95% lower bound for $\Delta\rho$ exceeds the 80–power threshold for that bridge size (here $\approx$ 0.015–0.020); (iii) the same decision is obtained on a repeated partition or chemistry-group resample; and (iv) a same-budget target-domain ridge cannot be trained because target Tc labels are unavailable. If (iv) fails, train the target ridge. If (ii) or (iii) fails, deploy the source composition ridge and treat CALR as an audit. All three real directions fail (ii) and fail to show a reproducible benefit.

Wider residual weights and a stricter gate. Extending the prespecified residual-weight grid from {0, 0.10, 0.25, 0.50} to include 0.75 and 1.00 still opens the positive control, but every fold then selects a weight at or above 0.75, and the held-out $\Delta\rho$ interval still includes 0 (0.018, -0.025–0.061 vs. 0.022, -0.011–0.053 at max = 0.50). The 0.50 cap is therefore not what makes the semi-synthetic improvement inconclusive. On the 48-case stress grid, 7/12 positive-mode cases open and 3 of those openings are harmful (worst regret 0.023). Requiring lower bound > 0.02 retains 3 openings, of which 1 remains harmful. The empirically defensible acceptance rule is as follows: open only if the nested 95% lower bound exceeds 0.02 on at least two independent composition partitions; otherwise, return the ridge. That rule agrees with the three closed real gates and reduces, but does not eliminate, harmful stress openings. It is an acceptance filter, not a formal safety guarantee.

Library-shift decomposition. On the 484 common labelled compositions, AFLOW and JARVIS $|M|$ correlate at $\rho = 0.490$. Space-group number matches for 427/484 keys, yet matched keys still have $\rho = 0.475$ vs. 0.394 unmatched. Space-group agreement is a necessary but not sufficient structure match; it does not encode magnetic initialization, exchange-correlation treatment, relaxation protocol or rare-earth 4f handling, which are not released uniformly in the frozen extracts. Residual disagreement on matched space groups is consistent with a contribution from methodological catalogue differences, although material-dependent and workflow-dependent effects remain confounded. Direct magnetic-endpoint pairs remain small: K1 $n = 14$, $\rho = 0.059$ (95% bootstrap -0.58–0.74); μ_0Ha $n = 25$, $\rho = 0.236$ (-0.21–0.61); Js $n = 10$, $\rho = -0.115$ (-0.81–0.74). Leave-one-composition-out ranges are 0.43, 0.19 and 0.52, with substantial leave-one-composition-out sensitivity, especially for Js; the intervals are too wide to support a ranking claim on those endpoints.

3. Results

3.1. Public Tables Occupy Different Chemical Spaces

The AFLOW lanthanide extract used here contains 14,482 unique auid records and 6746 compositions after Ce–Lu queries and identifier deduplication (Figure 1 and Table 1). La was excluded a priori because a single La query returns an enumeration-dominated dump an order of magnitude larger than the other lanthanides. That exclusion does not create the 2-14-1 or 1-5 zero counts: $Nd_2Fe_{14}B$ and $SmCo_5$ contain no La. Rare-earth–Fe records number 555 (241 compositions). Rare-earth–Co records number 744 (359 compositions). Rare-earth–Fe–B records number 25, and they are borides and borate oxides (R_2FeB_2, RFe_2B_2, $R_6Fe_2B_{14}$, $Ce_2Fe_8B_8$, and $NdFe_3B_4O_{12}$). The extract contains no $Nd_2Fe_{14}B$ cell. The single Nd–Fe–B composition is $B_4Fe_3NdO_{12}$.

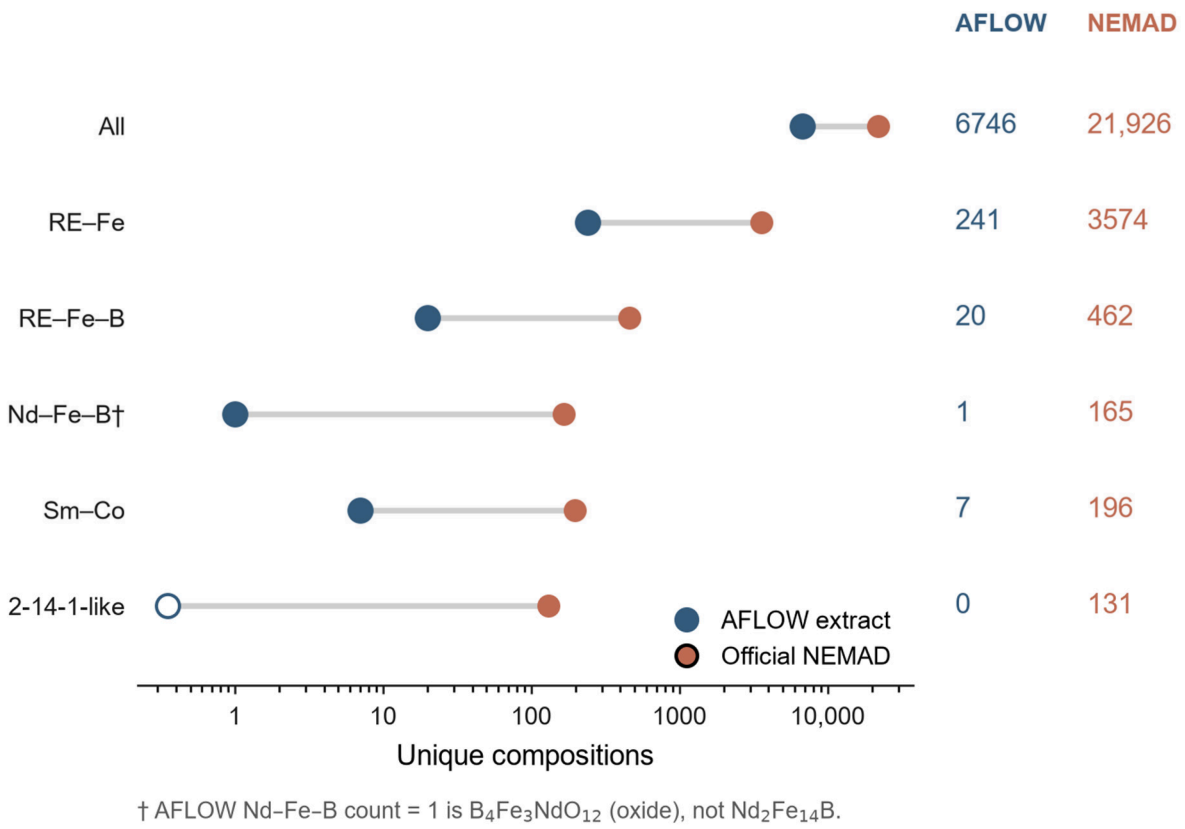

Figure 1. This AFLOW extract does not cover experimental REPM-relevant compositions. Dumbbells compare unique compositions in the AFLOW extract (blue) with the official NEMAD dump (vermillion) on a logarithmic scale. An open circle marks a zero count. The AFLOW Nd–Fe–B count of 1 is the oxide $B_4Fe_3NdO_{12}$, not $Nd_2Fe_{14}B$. The 2-14-1-like window is a Fe/B/RE fraction cut that excludes oxides; it is not a structure match (AFLOW 0; NEMAD 131). Source: results/audit_summary.json.

The NEMAD ferromagnetic Tc table contains 15,577 unique normalized compositions. Of these, 3596 contain a rare earth and Fe, 687 contain rare earth–Fe–B, and 350 contain Nd–Fe–B. The experimental table therefore holds the REPM chemistry that the DFT extract does not.

Exact-stoichiometry alignment (atomic fractions rounded to four decimals) against the official NEMAD dump produces 1865 shared compositions (27.6% of AFLOW; 8.5% of NEMAD unique keys). Only 663 of those pairs have a usable Tc, defined as a non-empty official Curie field after NA stripping; multiple rows on one key use the median. The Tc-aligned rare-earth–Fe set has 73 compositions. A 2-14-1-like fraction window, which requires Fe, B and a light or heavy rare earth and excludes oxides, finds 131 unique NEMAD

compositions and zero AFLOW compositions. Nd–Fe–B element-set overlap on the AFLOW side remains the single oxide $B_4Fe_3NdO_{12}$. The processed NEMAD ferromagnetic Tc table (15,577 unique compositions) confirms the same gap: 685 exact matches, 72 rare-earth–Fe matches, and still zero 2-14-1-like cells. The missing prototypes are not a rounding artefact.

Table 1. Public-data inventory (13 August 2026).

Layer	Source	Usable n	REPM-Relevant Slice	Status
DFT\	M\		AFLOW Ce–Lu extract	14,482 records/6746 compositions
Experimental Tc/Hc/Br	Official NEMAD magnetic CSV	33,668 records/21,926 compositions	RE–Fe 3574 unique; 2-14-1-like 131; Nd–Fe–B 165	Complete; website headline 67,584
Experimental K, Hcj, (BH)max	Official NEMAD anisotropy CSV	33,916 records/16,235 compositions	Nd–Fe–B 1890 (Hc 1664; (BH)max 997); 2-14-1-like 585; Hard 7453	Complete; AFLOW overlap 586, Nd–Fe–B overlap 0
Direct K1/μ0Ha/Js audit	Unit-normalized NEMAD anisotropy cells aligned to AFLOW	K1 14; μ0Ha 25; Js 10 exact pairs	K1 $\rho = 0.06$; μ0Ha $\rho = 0.24$; Js $\rho = -0.12$; no 2-14-1 overlap	Complete parser; low power
Experimental Tc (ML table)	NEMAD FM_with_curie	15,577 compositions	RE–Fe 3596; RE–Fe–B 687; Nd–Fe–B 350	Complete; no Hc/Br
DFT 2-14-1/1-5 cells	Materials Project	mp-5182 $Nd_2Fe_{14}B$; mp-1429 $SmCo_5$	Not in the AFLOW extract	Page-verified 13 August 2026
DFT\	M\		JARVIS 'dft_3d', 2025 snapshot	93,902 records; 16,623 usable Ce–Lu moments/13,712 compositions
Computed K/Js	Novomag	3826 claimed; 2051 parsed locally	Almost RE-free	Partial download
Magnetic structures	MAGNDATA	2414	Order check only	Count verified

Novomag (3826 computed entries, constructed for rare-earth-free magnets) and MAGNDATA (2414 magnetic structures as of 13 August 2026) were inspected as auxiliary layers [13,26]. A partial Novomag download (2051 parsed JSON lines) contained 94 rare-earth entries, dominated by Y–Co, and no Nd–Fe–B. MAGNDATA can check magnetic order for a subset of structures. Neither library repairs the AFLOW–NEMAD prototype gap.

3.2. Simple Composition Baselines Outperform DFT |M| for Tc Ranking

On the 663 official pairs with Tc (Table 2), Spearman's rank correlation between the AFLOW cell-wise maximum |M| and the NEMAD median Tc is 0.43 (bootstrap 95% interval 0.36–0.49; two-sided permutation $p < 0.001$; Figure 2). The association is therefore real and only moderate. The Tc distribution is cold: the median is 33.1 K, and the 90th percentile is 337 K. Only 22 aligned compositions have Tc $\geq$ 600 K; 81 have Tc $\geq$ 300 K.

Table 2. Exact-stoichiometry alignment on the official dump (pairs with Tc).

Set	n	Spearman ρ	Tc Median (K)	Tc p90 (K)
All aligned with Tc	663	0.43	33.1	337
RE–Fe	73	0.57	256	635
RE–Co	72	0.65	140	912
Oxide	42	0.11	38	469
Non-oxide	621	0.44	33	335
2-14-1-like	0	—	—	—

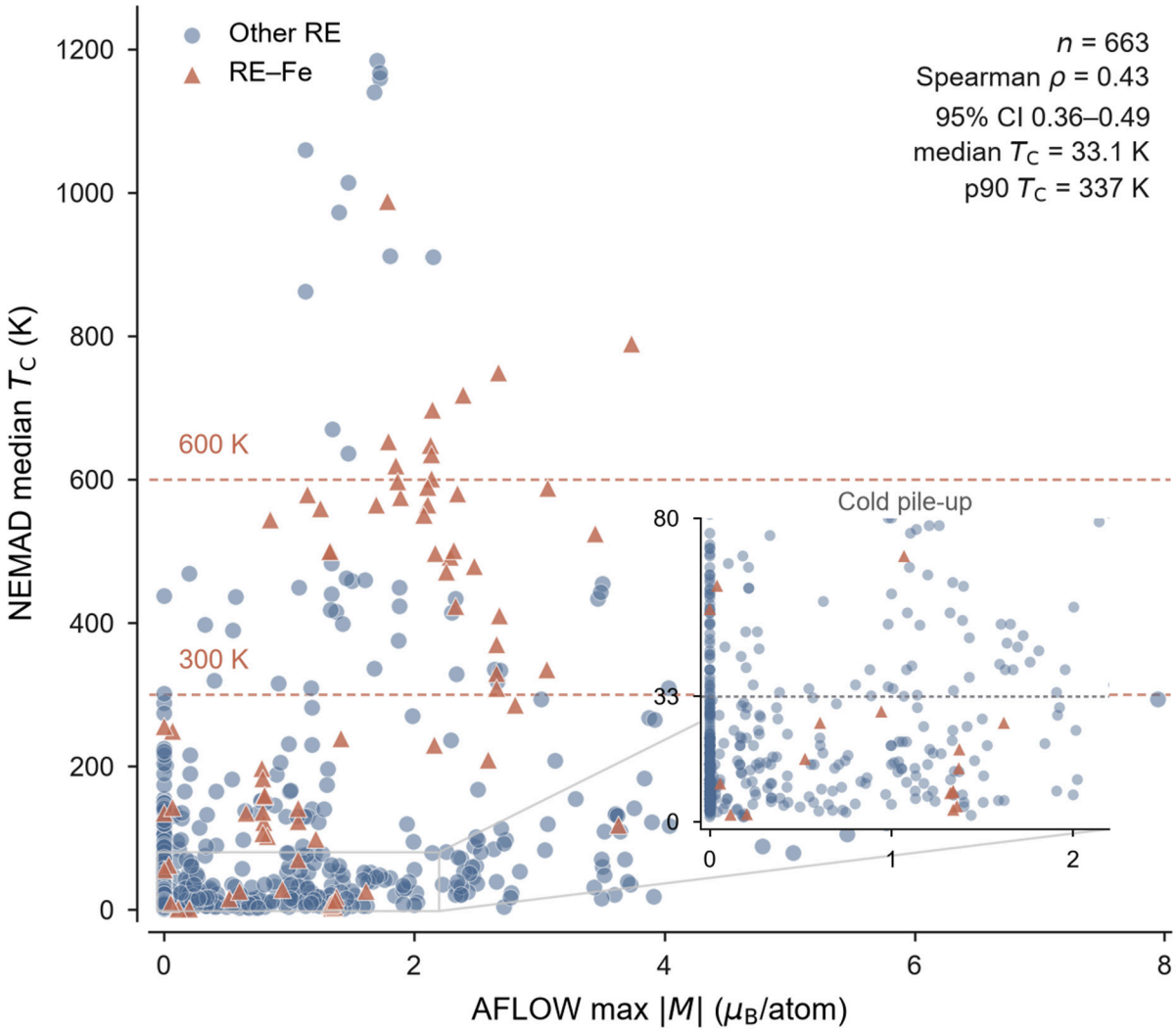

Figure 2. DFT |M| versus experimental Tc on 663 exact-stoichiometry pairs with a usable Tc. Vermillion triangles: rare-earth–Fe. Blue circles: other rare-earth chemistry. Inset: the cold pile-up at low |M| and low Tc; the dashed line is the median Tc (33.1 K). Spearman ρ = 0.43 (bootstrap 95% interval 0.36–0.49). 90th percentile = 337 K. Source: results/aligned_pairs.csv.

Ranking that set by |M| and scoring recovery of the Tc 90th percentile gives enrichment 0 at top 10, 0.40 at top 25, 1.19 at top 50 and 1.88 at top 100 (Figure 3 and Table 3). The base rate is 0.101. The head-of-list deficit at k = 10 is consistent with chance (hypergeometric $p = 0.34$), whereas the modest enrichment at k = 100 is not ($p = 0.0023$).

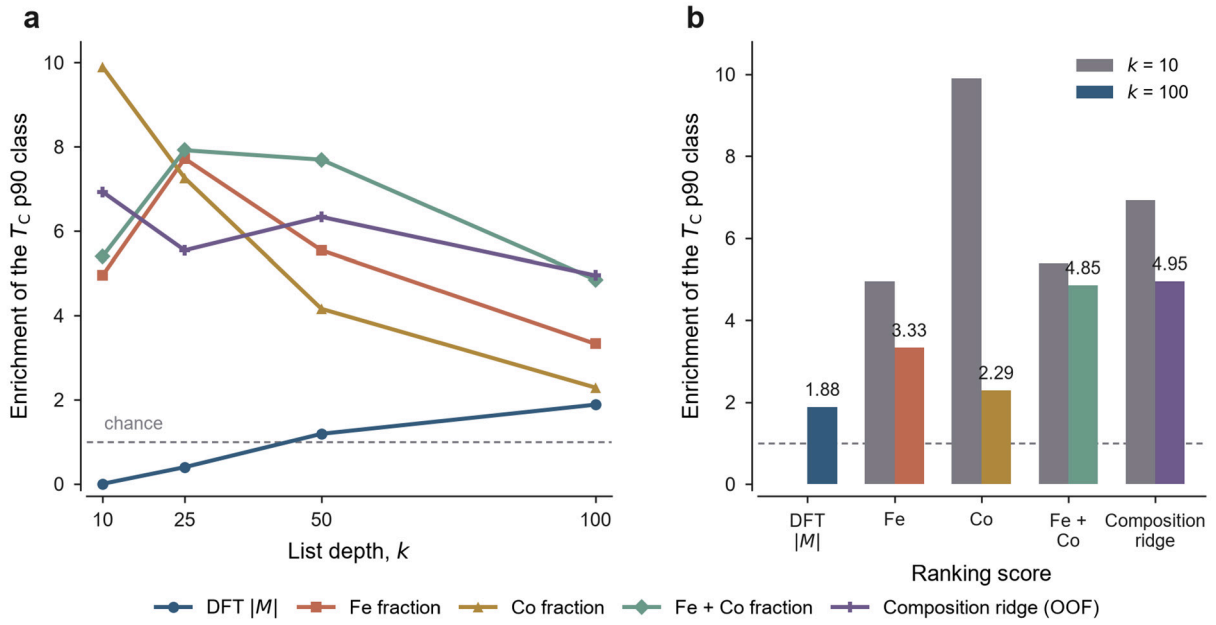

Figure 3. Composition-only scores outperform DFT |M| in ranking upper-decile Tc. Panel (**a**): enrichment versus list depth for DFT |M| (dark blue), Fe fraction (orange), Co fraction (gold), Fe + Co fraction (teal) and the repeated out-of-fold composition ridge model (purple). Panel (**b**): direct comparison at k = 10 (gray bars) and k = 100 (colored bars; same colors as in panel (**a**)). The ridge model uses elemental fractions only; every score is out of fold, and hyperparameter selection is nested inside each outer training set. Simple-score cutoffs are averaged across ties. Dashed line: chance (enrichment = 1). Source: results/composition_baselines.json.

Table 3. Upper-decile Tc recovery by DFT and composition-only rankings.

Label	Set	P@10	E@10	P@50	E@50	P@100	E@100
Tc ≥ p90, DFT | M|	score	Official Tc pairs (n = 663)	0.00	0.00	0.12	1.19	
Tc ≥ p90, Fe-fraction score	Official Tc pairs (n = 663)	0.50 *	4.95 *	0.56 *	5.54 *	0.34 *	3.33 *
Tc ≥ p90, Co-fraction score	Official Tc pairs (n = 663)	1.00 *	9.90 *	0.42 *	4.16 *	0.23 *	2.29 *
Tc ≥ p90, Fe + Co-fraction score	Official Tc pairs (n = 663)	0.55 *	5.40 *	0.78 *	7.69 *	0.49 *	4.85 *
Tc ≥ p90, composition ridge OOF	Official Tc pairs (n = 663)	0.70	6.93	0.64	6.33	0.50	4.95
Tc ≥ 600 K	Official Tc pairs (n = 22 positives)	0.00	0.00	0.02	0.60	0.03	0.90
Tc ≥ 300 K	Official Tc pairs (n = 81 positives)	0.20	1.64	0.20	1.64	0.29	2.37

Note. Asterisks denote expected precision/enrichment averaged across tied cutoff groups; ridge values are based on mean repeated out-of-fold scores. Under a hypergeometric null, the DFT-|M| Tc-p90 result at k = 10 and Tc ≥ 600 K at k = 50 are consistent with chance (p = 0.34 and 0.49). Formal pairwise uncertainty for model-selection comparisons was not prespecified, so the table supports a descriptive ranking contrast rather than a universal superiority claim.

Elementary chemistry is a stronger ranker. Tie-aware ranking by Fe fraction gives enrichment 4.95 at k = 10 and 3.33 at k = 100; Co fraction gives 9.90 and 2.29; and Fe + Co fraction gives 5.40 and 4.85. A ridge model using only 59 elemental fractions, trained

on log(1 + Tc), gives enrichment 6.93 at k = 10 and 4.95 at k = 100. Its mean repeated out-of-fold score has Spearman ρ = 0.70 with Tc and recovers 50 of the 67 upper-decile compositions by k = 100, versus 19 for |M|. The ridge result uses 20 repetitions of five-fold outer cross-validation with four-fold inner regularization selection; no row is scored by a model trained on that row. The simple-score cutoffs are tie-averaged, so their fractional expected hit counts do not depend on arbitrary ordering within equal elemental fractions.

The contrast is sharper for a permanent-magnet temperature cut, and it is also less well powered. For Tc $\geq$ 600 K, the top 10 DFT moments recover nothing (enrichment 0; expected count 0.33). The top 50 recover one compound (precision 0.02, enrichment 0.60; expected count 1.66; p = 0.49). The top 100 remain near chance (enrichment 0.90; expected count 3.32). These counts are consistent with a random ranking of 22 positives and show no detectable enrichment for the hot tail at the tested list depths. A 300 K cut, which includes many oxides and soft magnets, is the most favorable setting tested (enrichment 1.64 at top 10, not significant; 2.37 at top 100, Bonferroni-adjusted $p < 10^{-5}$) and remains far below the 9.97-fold self-enrichment of |M|. The processed Tc table (n = 685) repeats the top-10 result and gives enrichment 2.58 at top 100.

Restricting to rare-earth–Fe (n = 73) does not turn |M| into a REPM-oriented ranker. Spearman's ρ rises to 0.57. Enrichment for the subset 90th percentile of Tc is 1.83 at top 10 and 2.19 at top 25, then falls back toward 1 as k grows. For Tc $\geq$ 600 K, the top-10 enrichment is 1.46. Rare-earth–Co (n = 72) is the least hostile subset (ρ = 0.65; top-10 enrichment 4.50 for the Tc 90th percentile), so the global head-of-list failure is not uniform across every chemistry slice. That favorable Co slice still sits inside an extract with only seven Sm–Co compositions and no $SmCo_5$-like cell under a 1-5 fraction window. Oxides (n = 42) are essentially uncorrelated (ρ = 0.11).

3.3. JARVIS Cross-Library Validation Changes the Head of the Ranking

The official JARVIS-DFT dft_3d snapshot dated 24 September 2025 contains 93,902 records. Filtering to Ce–Lu gives 16,921 records, of which 16,623 have a numeric OUTCAR moment and collapse to 13,712 composition keys. Exact alignment to the same NEMAD table yields 744 Tc pairs. On this independently assembled overlap, JARVIS |M| has Spearman ρ = 0.53 (bootstrap 95% interval 0.47–0.59; two-sided permutation $p < 0.001$) and recovers 5, 38 and 55 upper-decile Tc compositions at k = 10, 50 and 100, corresponding to enrichment 4.96, 7.54 and 5.46 (Table 4). The top-10 count exceeds a random-ranking null (nominal hypergeometric p = 0.0015). The result is stable to the older 12 December 2022 snapshot: 673 pairs, ρ = 0.51, and enrichment 5.94, 7.72 and 5.34 at the same depths. Thus, the weak AFLOW head-of-list result does not generalize unchanged to every DFT catalogue.

Table 4. JARVIS cross-library and three-way common-set validation.

Moment Score	Evaluation Set	n	Spearman ρ	P@10	E@10	P@50	E@50	P@100	E@100
JARVIS 2025	JARVIS–NEMAD	744	0.53	0.50	4.96	0.76	7.54	0.55	5.46
JARVIS 2022	JARVIS–NEMAD	673	0.51	0.60	5.94	0.78	7.72	0.54	5.34
JARVIS 2025	Three-way common	484	0.34	0.80	7.90	0.56	5.53	0.36	3.56
AFLOW	Three-way common	484	0.36	0.40	3.95	0.44	4.35	0.42	4.15

Coverage can still confound a library comparison, so we also restrict both scores to the 484 AFLOW–JARVIS–NEMAD compositions with the same experimental Tc. On this common set, JARVIS retrieves 8 of 10 upper-decile compositions, versus 4 for AFLOW (enrichment 7.90 vs. 3.95), and 28 of 50, vs. 22 (5.53 vs. 4.35). The ordering reverses by k = 100: JARVIS retrieves 36 positives and AFLOW 42 (3.56 vs. 4.15). Their global correlations with Tc are likewise similar and slightly favor AFLOW ($\rho = 0.34$ vs. 0.36), while the two moment scores correlate only moderately with each other ($\rho = 0.49$). The JARVIS-minus-AFLOW enrichment differences have paired-bootstrap 95% intervals of 0.00–8.92 at k = 10 and −1.03–0.19 at k = 100; two-sided paired-rank randomization gives $p = 0.135$ and 0.070. The correlation difference is also unresolved ($\Delta\rho = -0.026$, 95% interval −0.105–0.054; $p = 0.510$). JARVIS therefore has higher early-retrieval point estimates on shared chemistry, not demonstrated cross-library superiority.

Neither JARVIS snapshot contains a composition inside the stated 2-14-1 window. Its standalone top ranks instead contain Fe-rich phases such as $SmFe_{12}$, $Sm_2Fe_{17}N_3$ and $SmFe_5$, which are both high-moment and experimentally hot. This is useful coverage, but it remains a Tc-ranking result rather than a validation of anisotropy, coercivity or energy product.

3.4. CALR Converts Catalogue Disagreement into Empirical Residual Gating

The audit above shows that a DFT moment can help in one catalogue and fail in another, while a composition model is more stable. We therefore use CALR as an empirical residual-gating protocol (Figure 4a). The algorithm starts from a source-trained elemental-fraction ridge, represents $|M|$ by its percentile among all available moments in that library, and adds at most 50% of the moment-augmented residual. Five nested bridge folds separate residual/penalty selection from gate evaluation: every paired composition is scored with settings selected on the other four folds. The residual is enabled only when the conditional bootstrap 95% lower bound of the resulting cross-fitted $\Delta\rho$ is above zero. A failed gate sets the residual and penalties to zero, making CALR exactly equal to the composition ridge. Ridge regression, bootstrap resampling and transfer under dataset shift are established components [24,25,27]; the contribution assessed here is the cross-library protocol formed by a nested paired-bridge gate, bounded residual and exact fallback. It is not claimed as formal risk control or as an abstaining predictor.

For AFLOW-to-JARVIS transfer, all 663 AFLOW Tc pairs are the source, and the 484 paired keys supply target representations in source-model OOF folds. Final evaluation uses the 260 JARVIS Tc keys absent from AFLOW (Table 5). The nested bridge $\Delta\rho$ is +0.0079 (95% interval −0.0027–0.0187), so the gate closes before external labels are read. CALR therefore equals the composition ridge on the external set: $\rho = 0.779$ and E@10 = 4.81. The ungated composition-plus-moment model reaches $\rho = 0.843$ ($\Delta\rho$ vs. ridge 0.065, 95% interval 0.022–0.114) and E@10 = 5.78. This is a conservative false-negative transfer decision on this direction, but it is the prespecified consequence of requiring bridge evidence that survives nested selection.

The reverse direction shows why the same gate is needed. Its nested bridge $\Delta\rho$ is +0.0060 (−0.0125–0.0238), so CALR again falls back before the 179 AFLOW-only labels are evaluated and preserves the ridge at $\rho = 0.524$, E@10 = 4.97, E@25 = 5.17 and E@50 = 3.58. In contrast, the ungated composition-plus-moment model falls to $\rho = 0.280$ ($\Delta\rho = -0.244$, 95% interval −0.335 to −0.161), although E@10 is unchanged. Across directions, ungated mean/worst-direction ρ is 0.562/0.280, vs. 0.651/0.524 for ridge and CALR. The empirical cross-library evidence therefore supports degradation avoidance, not successful real-data activation. The composition ridge is also an explicit always-anchor policy. CALR equals that policy in all three real directions; relative to the ex post better of anchor and ungated

residual, its Δρ opportunity costs are 0.065, 0 and 0.027 (mean 0.030). Thus, the current real data show no performance gain over always using the anchor. The added value is an auditable accept-or-return decision whose sensitivity cost is visible.

The external JARVIS top 10 provides a retrospective ranking check: the fallback recovers five upper-decile phases whose labels were unavailable to gate selection. We then apply the identical frozen model to JARVIS compositions with no exact NEMAD Tc key. The stated chemistry filters leave 340 rows; Table 6 reports the ten highest dimensionless composition-ridge scores solely to make deployment behavior auditable. Because the forward gate closed, the list contains no moment-supported transfer hits and is excluded from every performance claim.

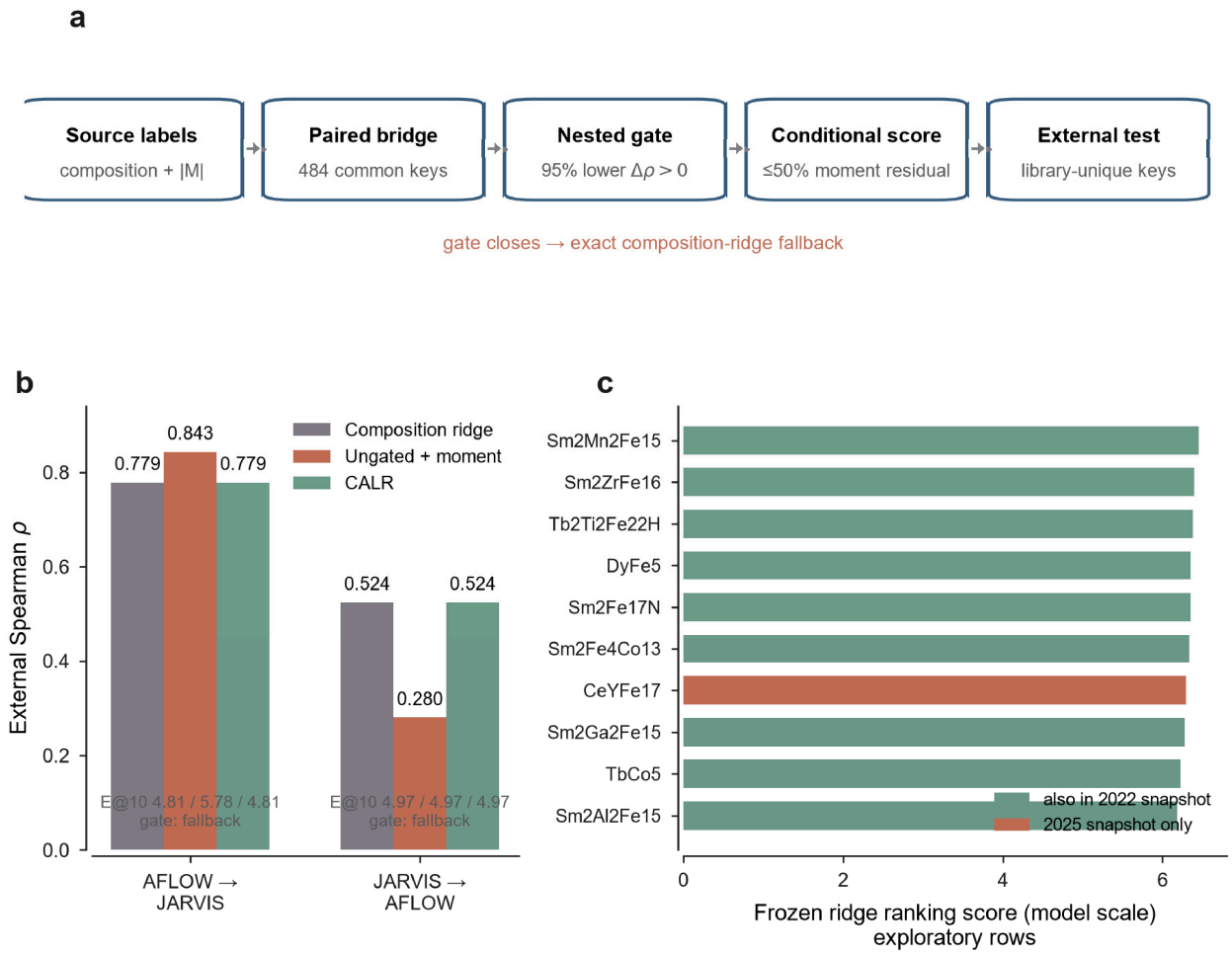

Figure 4. Nested residual-transfer gate, external ablation and exploratory deployment audit. (**a**), CALR uses nested, composition-disjoint paired-library OOF predictions to gate a bounded moment residual; a failed gate gives an exact composition-ridge fallback. (**b**), External Spearman ρ and E@10 for composition ridge, ungated composition plus moment and CALR on library-unique compositions. Both gates close. Arrows in (**b**) denote transfer from the source library to the target library (AFLOW → JARVIS and JARVIS → AFLOW). (**c**), Top 10 exploratory rows from a frozen-model deployment on unlabeled JARVIS compositions. Green bars mark nine rows also present in the 2022 snapshot. Bar lengths are dimensionless model-scale composition-ridge scores, not calibrated Tc predictions, measurements or materials-validation claims. Source: results/calr_cross_library.json and results/calr_jarvis_candidates.csv.

Table 5. Library-unique external validation and CALR ablation.

Source → External Target	Score	Gate	n	Spearman ρ	E@10	E@25	E@50
AFLOW → JARVIS	Raw DFT moment	—	260	0.753	5.78	5.39	4.24
AFLOW → JARVIS	Source-trained composition ridge	—	260	0.779	4.81	5.78	4.62
AFLOW → JARVIS	Ungated composition + moment	None	260	0.843	5.78	6.16	4.62
AFLOW → JARVIS	Nested gate, no risk penalty	Fallback	260	0.779	4.81	5.78	4.62
AFLOW → JARVIS	CALR	Fallback	260	0.779	4.81	5.78	4.62
JARVIS → AFLOW	Raw DFT moment	—	179	0.481	0.00	0.40	0.60
JARVIS → AFLOW	Source-trained composition ridge	—	179	0.524	4.97	5.17	3.58
JARVIS → AFLOW	Ungated composition + moment	None	179	0.280	4.97	5.17	3.58
JARVIS → AFLOW	Nested gate, no risk penalty	Fallback	179	0.524	4.97	5.17	3.58
JARVIS → AFLOW	CALR	Fallback	179	0.524	4.97	5.17	3.58

Note. Every bridge row is scored with hyperparameters selected on the other four nested bridge folds; final rows are library-unique compositions. The source-trained composition ridge is the explicit always-anchor policy. Nested bridge Δρ intervals are −0.0027–0.0187 forward and −0.0125–0.0238 reverse, so both gates close, and CALR equals that policy. On external labels, ungated Δρ vs. ridge is +0.065 (0.022–0.114) forward and −0.244 (−0.335 to −0.161) reverse. These are exploratory post-development tests and do not demonstrate a CALR gain over the anchor.

Table 6. Exploratory audit of frozen fallback deployment on unlabeled JARVIS rows.

Rank	Formula	JARVIS id	Frozen Ridge Score (Model Scale)	M (μB atom^{-1})	Formation Energy (eV atom^{-1})	Present in 2022
1	$Sm_2Mn_2Fe_{15}$	JVASP-122039	6.453	0.968	−0.00992	Yes
2	Sm_2ZrFe_{16}	JVASP-146044	6.398	1.611	−0.03894	Yes
3	$Tb_2Ti_2Fe_{22}H$	JVASP-119138	6.382	1.431	−0.11807	Yes
4	$DyFe_5$	JVASP-56971	6.355	1.559	−0.08505	Yes
5	$Sm_2Fe_{17}N$	JVASP-133415	6.351	1.271	−0.04789	Yes
6	$Sm_2Fe_4Co_{13}$	JVASP-140841	6.335	1.472	−0.09670	Yes
7	$CeYFe_{17}$	JVASP-169102	6.294	1.603	−0.02813	No
8	$Sm_2Ga_2Fe_{15}$	JVASP-133249	6.277	1.481	−0.04358	Yes
9	$TbCo_5$	JVASP-19620	6.229	1.157	−0.12787	Yes
10	$Sm_2Al_2Fe_{15}$	JVASP-146007	6.185	1.323	−0.09990	Yes

3.5. A Prespecified Activation Control Opens the Gate

The fixed-hash control separates the question of whether CALR can activate from whether the two tested catalogues justify activation (Figure 5). With the prespecified moment residual, all five nested gate folds selected the maximum allowed residual weight of 0.50. The concatenated gate score improved over the composition ridge by Δρ = 0.0816 (1500-bootstrap 95% interval 0.0416–0.1327), so the gate opened before the 145-composition test endpoint was read. On that held-out test, CALR increased ρ from 0.7689 to 0.7904; the

paired interval for the test difference was −0.0094–0.0531 and therefore does not establish a resolved test-set gain.

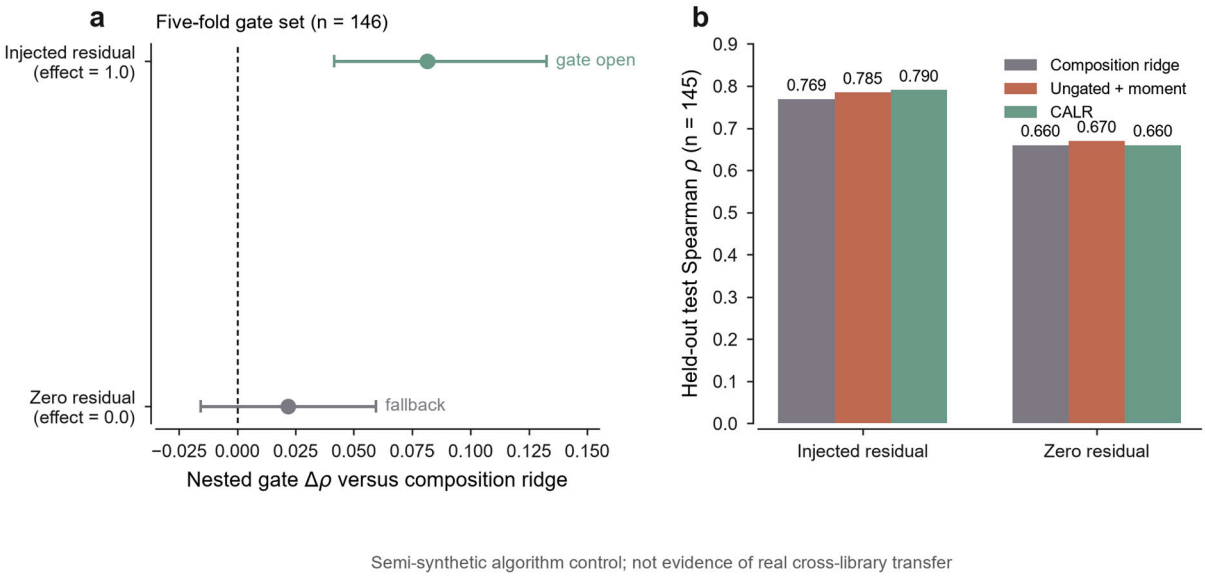

Figure 5. Fixed-hash semi-synthetic activation and zero-residual controls. (**a**), Nested gate Δρ and 1500-bootstrap 95% intervals relative to the composition ridge. The prespecified residual control opens; the matched zero-residual control falls back. (**b**), Spearman ρ on the 145-composition test partition, whose endpoint was not accessed before the gate decision. Both controls use 453 training and 146 gate compositions with real JARVIS composition and moment inputs. The endpoint is semi-synthetic; this figure validates gate operation, not real cross-library transfer or a material candidate. Source: results/calr_positive_control.json.

The matched zero-residual control used the same 453/146/145 composition split and deterministic noise but set the moment coefficient to zero. Its nested gate Δρ was 0.0219 (−0.0156–0.0595), so the gate closed, and CALR exactly reproduced the test-set composition ridge at ρ = 0.6600. The positive and zero-residual controls therefore demonstrate conditional opening and exact fallback under a fixed leakage-controlled protocol. Because their endpoints are semi-synthetic, they are an implementation and sensitivity check rather than empirical evidence that DFT moment transfer succeeds between AFLOW and JARVIS.

3.6. A Third Computational Source with Reused NEMAD Endpoints Closes the Gate

The checksum-verified Materials Project snapshot collapses 83,989 records to 55,885 composition keys and yields 459 MP–NEMAD pairs. Of these, 371 are also present in AFLOW; all 459 occur in JARVIS. MP-to-JARVIS therefore has 459 paired bridge compositions and 285 composition-disjoint JARVIS-only targets, whereas JARVIS-to-MP has no MP-only target. The computational source is new, but the NEMAD endpoint population is not method-level independent of the earlier AFLOW/JARVIS workflow: 113 of the 285 rows occurred in its bridge population, and 172 in its AFLOW-to-JARVIS external population. We therefore call this a third-source reanalysis on reused NEMAD endpoints. Within the frozen MP direction, labels remain withheld until evaluation. The MP-trained composition ridge has ρ = 0.7423. Ungated moment transfer reaches ρ = 0.7691 (Δρ = 0.0268, 95% interval 0.0095–0.0460), but the nested bridge Δρ is only 0.0101 (−0.0036–0.0250). The gate closes, and CALR returns the ridge. This reproduces conservative forward rejection, not independent method validation or real-data gate activation.

Alternative adaptation methods are not interchangeable on this same split. CORAL reduces ρ to 0.6724 (Δρ = −0.0699, 95% interval −0.1025 to −0.0361), and density-ratio

weighting gives 0.7147 (−0.0277, −0.0488 to −0.0089). Pooling MP and bridge labels gives 0.7425 (0.0002, −0.0014–0.0018). Retaining the 80% of external rows with lowest ensemble uncertainty raises the selective-set ρ to 0.8035, but changes the evaluation population and its Tc-p90 threshold; it is therefore a coverage-performance trade-off, not a full-coverage improvement claim.

The 48-case stress grid clarifies gate behavior beyond the original two controls (Figure 6). Each regime contains 12 fixed cases on one hash partition. All 12 zero-residual, 12 training-only/non-transferable and 12 sign-reversed cases close and fall back exactly. For nonzero transferable residuals, two of nine cases close. Across all 12 positive-mode cases, the gate opens in seven; three of those openings have harmful held-out Δρ, with worst ridge-minus-CALR regret 0.0234. Mean held-out regret across the positive mode is −0.0048. These are finite-grid outcomes, not calibrated future error rates. The rule is conservative on this grid but does not guarantee non-harm after opening. In the Spearman-only Stage 1 ablation, the forward gate still closes but the reverse gate opens (bridge lower bound 0.0017); external ρ then falls from 0.524 to 0.354 (Δρ = −0.171, 95% interval −0.243 to −0.108). Removing the early-retrieval component therefore does not turn the empirical gate into a safety guarantee and is materially worse on the observed reverse shift.

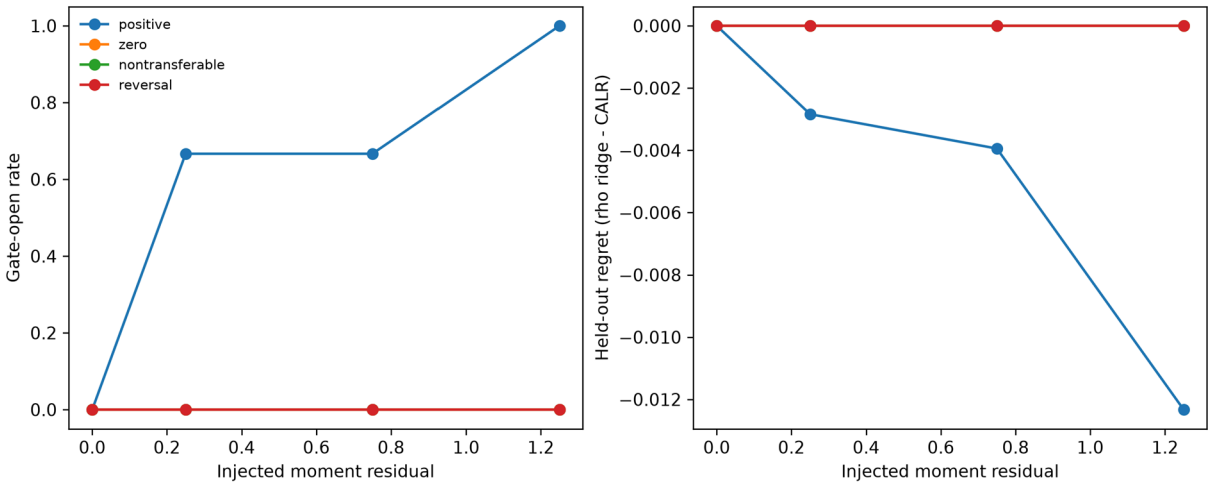

Figure 6. Fixed-hash CALR stress grid. (**Left**), gate-open rate across three deterministic noise levels for injected residual strengths 0–1.25. (**Right**), mean held-out regret relative to the composition ridge; negative values favor CALR. Transferable, zero, training-only/non-transferable and sign-reversed regimes use the same 453/146/145 composition partition. On the regret panel, the zero, non-transferable and sign-reversed curves coincide at zero because those gates close; all four legend entries are retained. This is an algorithmic control, not real materials evidence. Source: results/calr_stress_test.json.

As a boundary check on that exploratory list, current Materials Project records span 0.0004–0.286 eV atom^{-1} above hull, and only four rows have exact OQMD matches (Supplementary Note S18). The database and snapshot dependence reinforce why the list is not interpreted as a materials result. These values are not used to train, tune or evaluate CALR.

3.7. High-Moment Compositions Are Rarely High-Tc

We mark a composition as a trap if its maximum |M| lies at or above the aligned-set 90th percentile while its Tc lies below the aligned-set median. The median is 33.1 K, so this cut is deliberately conservative: it flags only compositions that are both high-|M| and experimentally cold, not every composition below a magnet-relevant Tc. Of 67 high-|M| compositions, only nine (13%) also sit in the experimental Tc 90th percentile (Figure 7).

Seven high-|M| compositions fall below the median Tc. The coldest examples are local-moment salts and Mn Laves phases: Eu_4As_3 (|M| = 3.91 μB atom^{-1}, Tc = 18 K), EuS (3.50 μB atom^{-1}, 16 K), Eu_3S_4 (2.71 μB atom^{-1}, 3.8 K), $HoMn_2$ (2.77 μB atom^{-1}, 24 K), and $ErMn_2$ (2.77 μB atom^{-1}, 18 K). These records are not DFT failures. They are cases in which a large 4f or local moment is real and the ordering temperature is not.

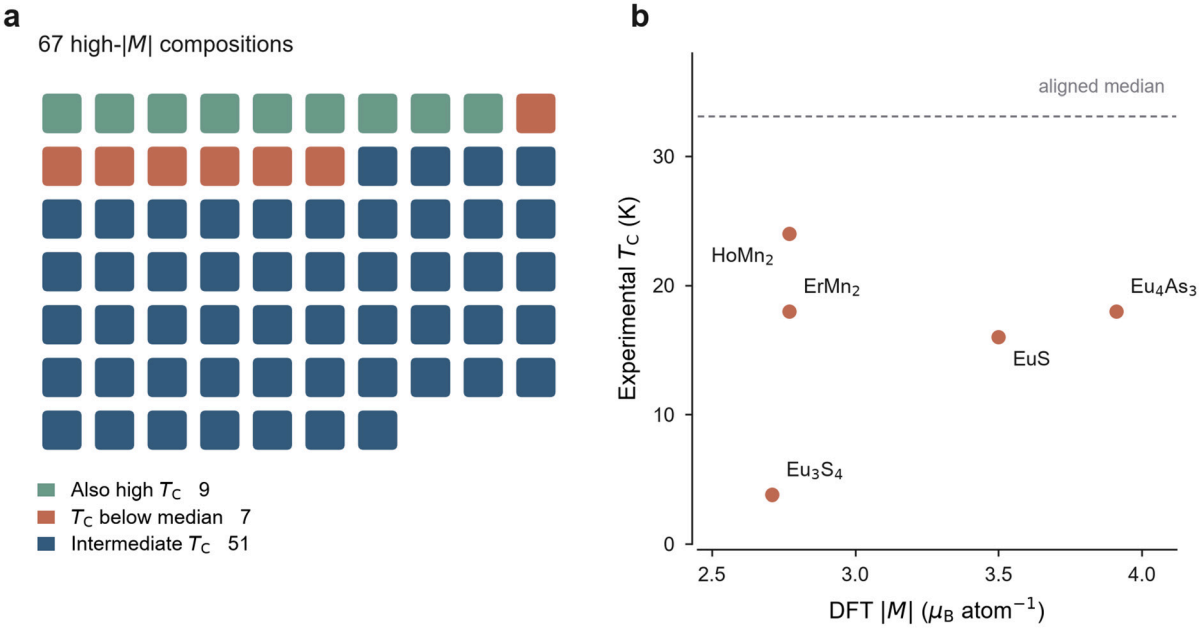

Figure 7. High-|M| compositions are rarely high-Tc. (**a**), Each tile is one of the 67 compositions at or above the |M| 90th percentile. Nine also reach the Tc 90th percentile; seven fall below the median Tc (33.1 K) and are counted as traps; the remaining 51 have intermediate Tc. (**b**), Labelled examples of these high-|M| compositions against experimental Tc; several lie below the aligned median (33.1 K), showing that a large local or 4f moment need not imply a high ordering temperature. Source: results/audit_summary.json.

The top of the |M| list is equally unhelpful for REPM screening even when Tc is not cryogenic. Elemental Gd and Tb appear with |M| = 7.94 and 6.29 μB atom^{-1} and Tc = 294 K and 220 K. Gd_3Ni, Gd_7Pd_3, and Gd_2Al occupy the same high-|M| band with Tc between 80 K and 334 K. A screener that trusts |M| promotes elemental gadolinium and europium pnictides over the 2-14-1 chemistry that is absent from the DFT extract.

Exact matching also exposes the deployment boundary: 4881 of 6746 AFLOW compositions (72.4%) have no official NEMAD partner, and 6083 (90.2%) have no exact Tc. These unmatched rows are excluded from performance claims rather than assigned pseudo-labels. Supplementary Notes S7 and S10 document legacy and privileged-signal diagnostics; neither contributes to the headline comparison.

3.8. The Anisotropy Dump Is Rich in Hcj and (BH)Max, but It Still Does Not Meet AFLOW

The magnetic table is not the only official NEMAD file. magnetic_anisotropy_materials.csv (8.23 MB; 33,916 rows; 19,672 DOIs; 16,235 unique parsed compositions) stores structured property cells with explicit units. Non-empty counts include coercivity 16,675, remanence 8572, K1 7272, anisotropy field 2876, and (BH)max 2884. The file is labelled by magnet class and sample form: 7453 Hard, 21,384 Soft, 24,938 Bulk, and 7014 Thin Film. Nd–Fe–B name strings number 1890, of which 1664 have coercivity, 1222 remanence, 997 (BH)max, and 297 a K or Ha field. A 2-14-1-like fraction window retains 585 rows (431 with Hc, 277 with (BH)max). Sm–Co name strings number 796 (629 with Hc, 314 with

(BH)max). Units remain heterogeneous (Hc in Oe, kOe, kA m^{-1}, T; (BH)max in MGOe and kJ m^{-3}; K1 in erg cm^{-3}, J m^{-3}, and MJ m^{-3}).

Exact-stoichiometry overlap with the AFLOW extract is 586 unique compositions. None of those overlapping rows is Nd–Fe–B or 2-14-1-like. The experimental REPM layer is therefore no longer the bottleneck. The computational extract is. A row-level audit of 50 Nd–Fe–B records (29 unique DOI) verifies matching Crossref metadata for all 50. Four released material names explicitly state Nd2Fe14B; the remaining 46 identify a family, modified sample or other nominal composition. None is promoted to an independently XRD/Rietveld-verified phase assignment, because DOI metadata and released names do not establish phase purity (Supplementary Note S4 and Table S1).

3.9. Direct Magnetic Endpoints Remain Sparse After Unit Control

To test whether the proxy gap also appears for an intrinsic magnetic endpoint, we parsed the anisotropy dump conservatively and aligned it to AFLOW by the same four-decimal composition key. K1 values with an energy-density unit were converted to |K1| in MJ m^{-3}; inequalities, ranges, unitless cells and incompatible units were excluded. This leaves 16 source rows and 14 exact composition pairs. The AFLOW moment has Spearman $\rho = 0.06$ with the median |K1|, and its top-10 recovery of the K1 upper decile is 1/10 (enrichment 0.70; two positive compositions). The corresponding controls are 25 pairs for μ_0Ha ($\rho = 0.24$; E@10 = 1.67) and 10 pairs for Js = μ_0Ms ($\rho = -0.12$; E@10 = 1.00). These endpoint counts are too small for a claim of universal ranking utility, but they establish a direct, unit-normalized magnet-property check rather than relying on Tc alone. No exact AFLOW pair is Nd–Fe–B or 2-14-1-like, so the endpoint audit does not become a processed-magnet validation.

3.10. Coercivity in the Magnetic Dump Remains a Weak Aligned Target

NEMAD's data dictionary lists coercivity in oersted and remanence in A m^{-1}. The GitHub/Zenodo machine-learning package does not contain those columns. The official website endpoint/backend/materials/download_magnetic returns magnetic_materials.csv with both fields as free text. The complete download used here has 33,668 records (4,525,021 bytes; 13 August 2026) and 4689 non-empty coercivity strings (13.9%) and 2009 remanence strings (6.0%) (Table 7). Those rates match the supplementary validation sample of 5015 records (coercivity true positives 648, 12.9%; remanence 295, 5.9%). The website headline on the same day was 67,584 magnetic entries. We report the downloadable table, not the headline.

Table 7. Official NEMAD dump (33,668 rows; 4.53 MB).

Field	Non-Empty	Rate	SI Rate on 5015
Curie (Tc)	21,343	63.4%	61.9%
Coercivity	4689	13.9%	12.9%
Remanence	2009	6.0%	5.9%
Magnetization	7064	21.0%	~17%
Nd–Fe–B names with Hc	132	—	—
Unit-parsed aligned Hc pairs	154	—	—

Name-string slices of that table contain 247 Nd–Fe–B records (132 with a coercivity field, 99 with remanence) and 5151 rare-earth–Fe records (1076 with coercivity). After strict unit parsing to kA m^{-1}, 154 exact AFLOW–NEMAD pairs remain. Spearman's ρ between |M| and parsed Hc is -0.17. Enrichment for the Hc 90th percentile is 1.54 at

top 50 and 1.25 at top 100. Remanence survives unit and range filters for seven pairs, which is too few to rank. The parser accepts single values and numeric ranges with explicit units; inequalities and multi-valued coated/uncoated strings are excluded. Supplementary field precision for coercivity is 87%. We therefore treat Hc as a unit-sensitive parser control, not as a gold-standard REPM endpoint, and we do not claim an Hcj ranking result for $Nd_2Fe_{14}B$.

3.11. Robustness of the Nested Residual Gate

Table 8 summarizes these robustness tests. The nsites-corrected Materials Project source, the frozen QQ percentile map, and four of five family hold-outs all close the gate and return the composition anchor. The exception is the AFLOW-to-JARVIS hold-out of the residual other chemistry class, which opens. Chemistry-group resampling never opens the forward gate and opens the reverse gate in two of six repeats, including one harmful opening. These results support the original closed-gate conclusions on the prespecified bridges, while showing that CALR is not a chemistry-agnostic safeguard once the bridge ceases to represent the deployment population.

Table 8. Robustness of the nested residual gate.

Test	Gate	Nested Δρ (95% Interval)	External Consequence
MP formula-atom (original)	Closed	0.0101 (−0.0036–0.0250)	CALR = ridge, ρ = 0.742
MP nsites-corrected	Closed	0.0093 (−0.0009–0.0191)	CALR = ridge, ρ = 0.732; decision unchanged
Frozen MP→JARVIS QQ percentile map	Closed	0.0094 (−0.0033–0.0212)	Held-out percentile ρ 0.677 vs. native 0.692
Family hold-out RE–Fe/RE–Co/oxide/pnictide	Closed (8/8)	All intervals include 0	CALR equals the retrained ridge
Family hold-out other (AFLOW→JARVIS)	Open	0.070 (0.041–0.103)	CALR ρ = 0.432 vs. ridge 0.389
Element-set resampling, AFLOW→JARVIS	0/6 open	Mean Δρ = −0.007	Stable closed fallback
Element-set resampling, JARVIS→AFLOW	2/6 open	Mean Δρ = 0.010	One helpful and one harmful opening
Same-budget target-domain ridge	—	—	Forward ρ = 0.793 vs. CALR 0.779; reverse 0.503 vs. 0.524
Same-budget validation selector	Chooses residual in reverse	—	Reverse external ρ = 0.456 vs. CALR 0.524
Spearman-only Stage 1 (single-objective ablation)	Reverse opens	0.011 (0.002–0.021)	External ρ falls from 0.524 to 0.354

Note. Nested intervals are paired-bootstrap 95% bounds for Δρ versus the composition ridge on the corresponding bridge. Resampling used a reduced ensemble (eight models; 400 CI replicates). Family hold-outs re-ran the full nested pipeline with the named family removed from source training. The Spearman-only ablation is the previously reported diagnostic in which Stage 1 no longer uses E@k; it is included here because it is the direct test of a single Spearman objective.

3.12. Sensitivity and Claim-Boundary Tests

Table 9 and Figure 8 collect these sensitivity and claim-boundary tests. Structure filters and median versus maximum |M| move the headline Spearman by at most 0.03. The composition ridge is the stronger ranking model under random folds, but family hold-out removes much of that apparent generalization. The MP run remains a closed gate on reused NEMAD labels. Nested re-seeding and a minimum-effect rule leave the three real gates closed. CALR therefore continues to equal the always-anchor policy on every evaluable real direction; the new material is the quantitative statement of when that policy should not be replaced.

Table 9. Structure, power, independence and endpoint tests.

Test	Setting	Result	Consequence for the Claim
Headline max |M| vs. Tc	n = 663 keys	ρ = 0.428; E@100 = 1.88	Catalogue-key ranking baseline
Median |M|/majority Pearson	n = 663	ρ = 0.417/0.428	Aggregation does not create the correlation
Unique Pearson/unique SG	n = 590/584	ρ = 0.418/0.417	Polymorph collisions are not the source of $\rho \approx 0.43$
Collision-free keys only	n = 277	ρ = 0.396; E@10 = 2.97	Headline top-10 is collision-sensitive; global ρ is not
Family-grouped ridge OOF	5 families; n = 663	ρ = 0.461 vs. random OOF 0.697	Random CV inflates composition-only generalization
Bridge power (LB > 0)	n = 484; SE = 0.0054/0.0094	80% power at $\Delta\rho \approx 0.015$–$0.020/0.030$	Observed 0.008 and 0.006 are below detection
Nested full-pipeline seeds	6 seeds × 2 directions (reduced ensemble)	1/6 forward open; 0/6 reverse	The one opening fails LB > 0.02
MP experimental labels	285 external keys	100% previously seen in AFLOW/JARVIS–NEMAD	Computational-source replication, not independent validation
2-14-1 in AFLOW extract	$Nd_2Fe_{14}B$ and 2-14-1-like window	n = 0 cells/0 keys	REPM screening in this family is a coverage failure
Direct K1/Ha/Js	n = 14/25/10	ρ = 0.06/0.24/−0.12; all CIs include 0	Intervals, not point estimates, are the supported statement
Wider residual weights	max 0.50 vs. max 1.00	Held-out $\Delta\rho$ CIs both include 0	The 0.50 cap is not what makes the control inconclusive
Stricter gate LB > 0.02	Stress 7 openings; 3 survive	3/7 harmful → 1/3 harmful	Reduces harm; does not create a real-data gain

Note. Spearman values are |M| versus experimental Tc except for the composition-ridge OOF rows and the direct-endpoint rows. Nested seeds used a reduced ensemble (eight models; 400 CI replicates). Power uses the observed nested bootstrap SE under a normal approximation to the LB > 0 test. Collision-free means one AFLOW cell on that four-decimal key. Unique Pearson/space group means all AFLOW cells on the key share that identifier.

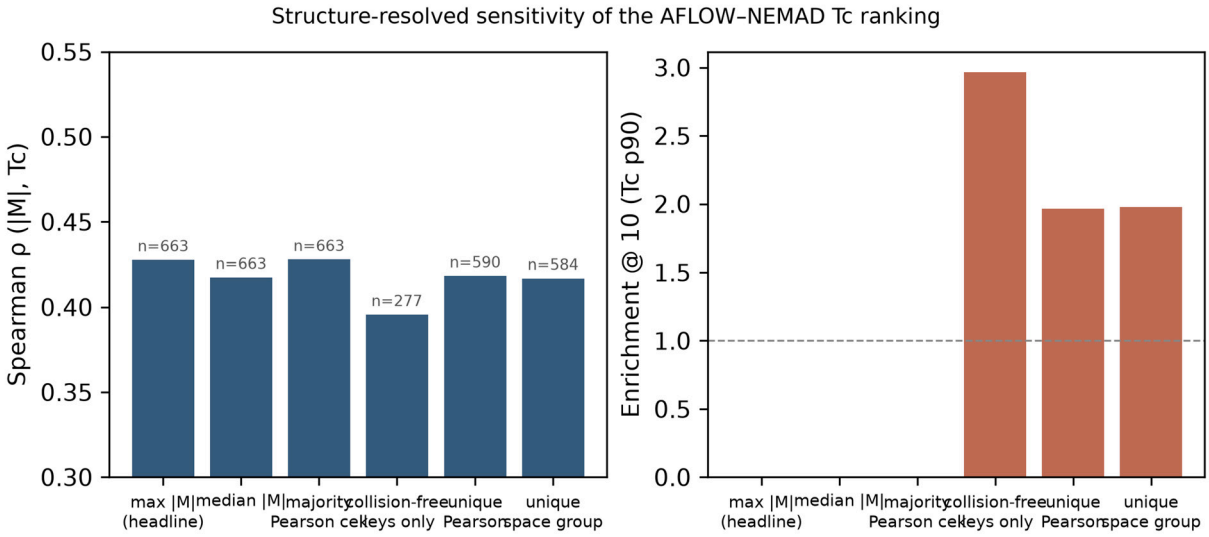

Figure 8. Structure-resolved sensitivity of AFLOW |M| versus NEMAD Tc. Left: Spearman ρ under alternative aggregations and crystallographic filters. Right: enrichment at k = 10 of the Tc 90th percentile. Sample sizes are printed above the ρ bars.

4. Discussion

The practical conclusion is stronger than a moment-versus-chance comparison but narrower than a universal ranking claim. In the AFLOW overlap, DFT |M| recovers no upper-decile Tc composition at top 10 and reaches enrichment 1.88 at top 100, below Fe + Co fraction and the leakage-controlled composition model. The independent JARVIS overlap is more favorable, with enrichment 4.96 and 5.46 at the same depths. The controlled three-way comparison then separates catalogue membership from scoring: JARVIS has higher point estimates than AFLOW at k = 10 and 50 on 484 identical compositions, but not at k = 100 or in global rank correlation, and none of the paired differences is resolved at the 0.05 level. Moment-to-Tc transfer is therefore a property of the database, overlap and decision depth together, not of a moment field in isolation. Because no experimental testing cost or minimum acceptable precision was prespecified, these point estimates are not a universal declaration that one score is useful or useless.

CALR turns that limitation into a defined residual-acceptance decision. The ungated moment residual improves AFLOW-to-JARVIS global ordering but sharply degrades the reverse direction. Nested bridge selection rejects the residual in both cases. The frozen MP source gives the same result: ungated MP-to-JARVIS transfer improves the external point estimate and paired interval, yet its bridge interval again crosses zero, and CALR falls back. CORAL and density-ratio adaptation cause resolved losses on the same external set, while pooled bridge training is neutral. These results expose a sensitivity–rejection trade-off. CALR's supported real-data behavior is identical to the always-anchor policy, with one residual rejected in the direction where ungated transfer was harmful and two favorable residuals also rejected under unresolved bridge evidence. It is residual rejection under unresolved bridge evidence, not prediction abstention, absolute safety, demonstrated gain over the anchor, or successful moment transfer.

The fixed-hash controls answer a narrower algorithmic question. CALR is not structurally locked in fallback: a known residual produces a positive gate lower bound, while removing that residual closes the gate under the same split and noise. The broader grid closes all 36 tested zero, non-transferable and sign-reversed cases and misses two of nine nonzero transferable cases. Three of seven positive-mode openings are harmful on the held-out partition. This supports conditional gate operation and label isolation, but not a

formal degradation guarantee. The real cross-library conclusions must therefore continue to be read from the three closed gates, not from semi-synthetic endpoints.

This is not an argument against DFT. Anisotropy energy, magnetization at finite temperature, exchange and processed-microstructure models remain the relevant computational route to intrinsic and extrinsic magnet performance [3,4,6,7,13,14]. It is an argument against treating an unvalidated |M| top-k list as a magnet list. The AFLOW failure against Fe/Co fractions shows why a random baseline is insufficient; the JARVIS result shows why a single computational catalogue is also insufficient. A scientifically meaningful proxy should add information beyond formula chemistry and retain that value under a stated deployment population.

Several alternative explanations should be stated. First, a composition key can merge polymorphs and miss off-stoichiometric 2-14-1 grains. Three-, four- and five-decimal AFLOW keys produce the same 663 Tc pairs and empty top 10; median rather than maximum |M| changes E@100 from 1.88 to 2.18. A separate AFLOW label census also finds zero tP68/$P4_2$/mnm (space group 136) rows. These checks strengthen the extract-level gap but do not constitute Rietveld phase identification. Second, AFLOW and JARVIS differ in structures, workflows and covered chemistry. Even on their common compositions their moment scores correlate only $\rho = 0.49$, so a database name is not a negligible nuisance variable. Third, a zero-Kelvin DFT spin moment and a literature-extracted ordering temperature are related but non-equivalent quantities. The 4f-in-core/default-GGA limit is visible in the Materials Project $SmCo_5$ cell (mp-1429; 0.106 μB/cell, unphysical under that setting). Fourth, the AFLOW 600 K tail has only 22 positives, so an inference of under-enrichment is unsupported. Finally, Materials Project already hosts $Nd_2Fe_{14}B$, whereas neither tested Ce–Lu catalogue contains a 2-14-1 match under our window. These are catalogue-specific coverage statements, not claims about all public DFT. Single-crystal anisotropy of $R_2Fe_{14}B$, not a per-atom moment, is the intrinsic quantity that later becomes Hcj after processing [7].

The algorithm has additional limits. It was developed on AFLOW and JARVIS rather than preregistered, and all three evaluable empirical gates close. The semi-synthetic controls prove that the implemented decision rule can activate under an injected transferable residual, but no real library demonstrates successful activation. The Spearman-only ablation further shows that a positive bridge lower bound can coexist with resolved external harm. The ridge features are composition-level and cannot distinguish polymorphs, site occupancy or magnetic order. Because the real-data gates set both penalties to zero, the present empirical data do not validate the lower-confidence penalty as a calibrated error control. The unlabeled-list exercise is only an audit of frozen fallback behavior and has no role in method validation.

The nsites correction confirms that the original MP conversion used too few atoms whenever $Z > 1$, yet the catalogue-level ranking and the closed MP gate are insensitive to that scale error. The provenance ledger makes the aggregation rule inspectable: maximum |M| and median Tc do not describe a unique cell or a unique experimental record. Percentile harmonization does not make AFLOW, JARVIS and MP moments physically interchangeable; it only preserves within-library rank, and a frozen QQ map does not open a previously closed gate. The bridge represents the named rare-earth–Fe/Co, oxide and pnictide slices of this extract, but not the residual other class and not every chemistry-group resample of the reverse direction. CALR's value under a 484-label budget is not a gain over a target-trained composition ridge on the mild forward split; it is the refusal to follow a bridge-tuned selector into reverse-transfer harm. Defining Spearman ρ as the sole protected estimand makes that refusal explicit. It also shows, via the Spearman-only Stage-1 ablation, that dropping the conservative early-retrieval filter can open the harmful reverse gate. We

therefore keep Spearman as the activation criterion and retain Stage 1 only as a setting filter that cannot declare success.

Structure-resolved ranking and median-versus-max aggregation show that the headline $\rho \approx 0.43$ is a catalogue-key fact, not an artefact of rounding four decimals onto disparate polymorphs, although CeO_2-like collisions remain a physical mismatch between a computed cell and an experimental median. Family-grouped CV shows that the composition ridge's random-fold $\rho = 0.70$ should not be read as family-agnostic generalization. The 484-key bridge can reliably detect only larger residual effects; a closed gate therefore should not be interpreted as evidence that a small useful residual is absent. MP is a third DFT source on reused NEMAD labels. Direct K1/Ha/Js correlations are compatible with a wide range of true values. No 2-14-1 chemistry is present in the computational extract, so permanent-magnet screening claims stop at that coverage failure. CALR remains an auditable return to the composition ridge. A practical benefit would require an independent experimental library, a nested lower bound above the 80% power threshold on repeated partitions, and a held-out interval that excludes zero. Those conditions are not met here.

The paper also bounds what a follow-up can claim. The anisotropy dump already contains 1664 Nd–Fe–B coercivity strings and 997 (BH)max strings, so experimental sparsity is no longer the reason to withhold an Hcj study. The frozen MP snapshot supplies a same-protocol moment field, while current OPTIMADE supplies structure and stability metadata but no bulk moment. Materials Project contains $Nd_2Fe_{14}B$ (mp-5182; 31.304 μB/cell; $P4_2/mnm$; PAW Nd_3), related $R_2Fe_{14}B$ cells and $SmCo_5$ (mp-1429; 0.106 μB/cell; P6/mmm; PAW Sm_3) [9]. The downloaded JARVIS snapshots close the Tc cross-library test but contain no composition in the stated 2-14-1 window. A direct Hcj study still requires magnetic-order-resolved, spin-orbit and finite-temperature calculations and explicit separation of intrinsic phase properties from microstructure and processing.

The exploratory unlabeled-list audit is not a materials-discovery result. Its score is neither a calibrated Tc prediction nor evidence that moment transfer activated, and database hull metadata do not establish magnetic performance. We do not treat CALR as a substitute for magnetic-ground-state searches, spin-orbit MAE, finite-temperature magnetism, grain-boundary diffusion or sintering.

5. Conclusions

No real CALR gate activates or improves on the composition anchor. One harmful and two favorable residuals are rejected in evaluated directions. A Spearman-only selector opens a harmful reverse transfer. The method is an empirical audit protocol, not formal risk control. Audit rows remain exploratory and lack magnetic validation.

CALR converts a catalogue audit into an empirical cross-library residual-gating test. An ungated composition-plus-moment model improves AFLOW-to-JARVIS ρ from 0.779 to 0.843 but degrades JARVIS-to-AFLOW from 0.524 to 0.280. A third-source MP-to-JARVIS reanalysis improves the ungated external score from 0.742 to 0.769, yet its bridge interval also crosses zero. CALR therefore rejects the residual and returns the composition anchor in all three evaluable real directions. It has no observed performance advantage over always using that anchor, but makes one evaluated residual rejection and two rejected favorable transfers auditable. CORAL and density-ratio weighting lose performance on the common MP split. All 36 tested zero, non-transferable and reversed grid cases close; three of seven positive-mode openings are harmful on held-out global ranking. The supported advance is a reproducible materials-database audit and empirical residual-gating protocol, not formal risk control or top-k protection. The exploratory audit rows are not materials claims. After correcting MP moments by current cell atom counts, tracing matches to auid/jid/mpid/DOI records, freezing percentile calibration, holding out chemical families,

and resampling element-set groups; the three prespecified real gates remain closed. CALR is not a substitute for a target-domain model when many target labels already exist, and it is not stable if the bridge is stripped of the majority chemistry class. Its supported role is an auditable residual-acceptance rule with a single Spearman deployment objective. Structure-resolved and family-hold-out tests leave that role unchanged. The composition anchor is weaker across chemical families than random-fold CV suggests, the 484-key bridge is underpowered for the observed residual, and Materials Project does not supply independent experimental labels. We do not claim a ranking benefit over the composition ridge, and we do not claim 2-14-1 permanent-magnet screening from an extract that contains no 2-14-1 cells.

AFLOW records are available from aflow.org through AFLUX. The JARVIS-DFT snapshots are official Figshare files 64,391,379 and 38521619. The frozen Materials Project file is Matminer's mp_nostruct_20181018 Figshare file 13309253; current stability metadata are from Materials Project OPTIMADE, and OQMD stability records are from its public formation-energy API. The NEMAD Tc table is in Zenodo record 17042814. The official NEMAD magnetic dump and anisotropy dump are obtained from nemad.org as described in Section 2; we do not redistribute those full files. Novomag is Zenodo record 17399362. MAGNDATA is hosted at cryst.ehu.eus/magndata. Aligned pairs, CALR external predictions, activation/stress controls, transfer benchmarks, stability-query outputs, candidates, uncertainty tables, direct-endpoint pairs and the 50-row Nd–Fe–B DOI/nominal-identity audit are provided as Source Data/Supplementary Files. The authors did not generate experimental measurements.

Alignment, AFLOW/JARVIS/MP cross-library ranking, CALR training/bridge gating/exploratory deployment audit, activation/stress controls, method benchmarking, stability queries, composition baselines, direct-endpoint parsing, DOI audit, uncertainty and figure scripts accompany this article. The empirical entry points are run_coverage_aware_ranker.py and run_mp_cross_library.py; the benchmark and controls are run_transfer_benchmark.py, run_calr_positive_control.py, run_calr_stress_test.py and run_calr_spearman_only_ablation.py; stability reconciliation is query_candidate_stability.py. Generated files include results/calr_cross_library.json, results/calr_spearman_only_ablation.json, results/mp_cross_library.json, results/transfer_benchmark.json, results/calr_stress_test.json, results/candidate_independent_stability.csv and the audit/prediction tables. The legacy AFLOW moment-model archive is described only in the Supplementary Information. A persistent repository DOI will be minted before submission.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/magnetochemistry12090107/s1, Table S1: Nd–Fe–B anisotropy sampling frame and nominal-identity audit; Table S2: Four-decimal key collisions; Table S3: Spearman and hypergeometric checks; Table S4: Ranking sensitivity; Table S5: Library-unique external performance; Table S6: CALR gate controls; Supplementary Note S1: Alignment key and collisions; Supplementary Note S2: Coercivity and remanence parser; Supplementary Note S3: Uncertainty for already-reported point estimates; Supplementary Note S4: Nd–Fe–B anisotropy sampling frame; Supplementary Note S5: Materials Project page checks (13 August 2026); Supplementary Note S6: Files not redistributed; Supplementary Note S7: Prior ExtraTrees archive (not refit); Supplementary Note S8: Aggregation and key-rounding sensitivity; Supplementary Note S9: AFLOW structure-label census; Supplementary Note S10: Exploratory neighbourhood rule; Supplementary Note S11: Composition baselines; Supplementary Note S12: Materials Project status and JARVIS cross-library validation; Supplementary Note S13: Direct K1, Ha and Js endpoints; Supplementary Note S14: CALR algorithm, bridge gate and exploratory audit; Supplementary Note S15: Fixed-hash activation and zero-residual controls; Supplementary Note S16: Third computational source with reused NEMAD endpoints; Supplementary Note S17: Common-split transfer benchmark and stress grid; Supplementary Note S18:

Independent candidate stability reconciliation; Supplementary Note S19: Reviewer 1 robustness analyses; Supplementary Note S20: Reviewer 2 structure, power and claim-boundary analyses.

Author Contributions: Conceptualization, J.-F.L. and Q.C.; methodology, J.-F.L.; software, J.-F.L.; validation, L.Z.; formal analysis, J.-F.L. and L.Z.; investigation, J.-F.L. and Y.D.; resources, Y.D.; data curation, Y.D. and Z.L.; writing—original draft preparation, J.-F.L.; writing—review and editing, Q.C., L.Z. and Z.L.; visualization, Z.L.; supervision, Q.C.; project administration, Q.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (grant number 2023YFB3506401).

Institutional Review Board Statement: Not applicable. This study used public computational and literature-extracted tables and involved no human participants or animals.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding author.

Acknowledgments: We thank the maintainers of AFLOW, JARVIS, NEMAD, the Materials Project, Novomag and MAGNDATA for public access to the source tables.

Conflicts of Interest: Jun-Feng Li, Qian Chen, Lei Zhou, Yang Du, and Zhen Liang were employed by Advanced Technology & Materials Co., Ltd. All authors participated in the design, experiments, data analysis and manuscript preparation of this study. The company provided experimental conditions but had no role in study design, data analysis, writing of the manuscript or the decision to submit for publication. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Gutfleisch, O.; Willard, M.A.; Brück, E.; Chen, C.H.; Sankar, S.G.; Liu, J.P. Magnetic materials and devices for the 21st century: Stronger, lighter, and more energy efficient. *Adv. Mater.* **2011**, *23*, 821–842. [CrossRef] [PubMed]
2. Skokov, K.P.; Gutfleisch, O. Heavy rare earth free, free rare earth and rare earth free magnets—Vision and reality. *Scr. Mater.* **2018**, *154*, 289–294. [CrossRef]
3. Coey, J.M.D. Perspective and prospects for rare earth permanent magnets. *Engineering* **2020**, *6*, 119–131. [CrossRef]
4. Coey, J.M.D. Permanent magnetism. *Solid State Commun.* **1997**, *102*, 101–105. [CrossRef]
5. Sagawa, M.; Fujimura, S.; Togawa, N.; Yamamoto, H.; Matsuura, Y. New material for permanent magnets on a base of Nd and Fe. *J. Appl. Phys.* **1984**, *55*, 2083–2087. [CrossRef]
6. Herbst, J.F. R2Fe14B materials: Intrinsic properties and technological aspects. *Rev. Mod. Phys.* **1991**, *63*, 819–898. [CrossRef]
7. Hirosawa, S.; Matsuura, Y.; Yamamoto, H.; Fujimura, S.; Sagawa, M.; Yamauchi, H. Magnetization and magnetic anisotropy of R2Fe14B measured on single crystals. *J. Appl. Phys.* **1986**, *59*, 873–879. [CrossRef]
8. Curtarolo, S.; Setyawan, W.; Hart, G.L.; Jahnatek, M.; Chepulskii, R.V.; Taylor, R.H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218–226. [CrossRef]
9. Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002. [CrossRef]
10. Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R.H.; Nelson, L.J.; Hart, G.L.; Sanvito, S.; Buongiorno-Nardelli, M.; et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **2012**, *58*, 227–239. [CrossRef]
11. Choudhary, K.; Garrity, K.F.; Reid, A.C.E.; DeCost, B.; Biacchi, A.J.; Walker, A.R.H.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A.G.; Centrone, A.; et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **2020**, *6*, 173. [CrossRef]
12. Choudhary, K. The JARVIS infrastructure is all you need for materials design. *Comput. Mater. Sci.* **2025**, *259*, 114063. [CrossRef]
13. Sakurai, M.; Wang, R.; Liao, T.; Zhang, C.; Sun, H.; Sun, Y.; Wang, H.; Zhao, X.; Wang, S.; Balasubramanian, B.; et al. Discovering rare-earth-free magnetic materials through the development of a database. *Phys. Rev. Mater.* **2020**, *4*, 114408. [CrossRef]
14. Vishina, A.; Vekilova, O.Y.; Björkman, T.; Bergman, A.; Herper, H.C.; Eriksson, O. High-throughput and data-mining approach to predict new rare-earth-free permanent magnets. *Phys. Rev. B* **2020**, *101*, 094407. [CrossRef]

15. Nelson, J.; Sanvito, S. Predicting the Curie temperature of ferromagnets using machine learning. *Phys. Rev. Mater.* **2019**, *3*, 104405. [CrossRef]
16. Boora, N.; Ahmad, R.; Rahman, S.; Dung, N.Q.; Ahmad, A.; Alshammari, M.B.; Lee, B.-I. Recent Advances of Colossal Magnetoresistance in Versatile La-Ca-Mn-O Material-Based Films. *Magnetochemistry* **2025**, *11*, 5. [CrossRef]
17. Gou, X.; Sun, X.; Cheng, P.; Shi, W. Molecular Nanomagnets with Photomagnetic Properties: Design Strategies and Recent Advances. *Magnetochemistry* **2025**, *11*, 77. [CrossRef]
18. Rouabah, S.; Didouche, F.-Y.; Khebli, A.; Aguib, S.; Chikh, N. The Advancing Understanding of Magnetorheological Fluids and Elastomers: A Comparative Review Analyzing Mechanical and Viscoelastic Properties. *Magnetochemistry* **2025**, *11*, 62. [CrossRef]
19. Itani, S.; Zhang, Y.; Zang, J. The northeast materials database for magnetic materials. *Nat. Commun.* **2025**, *16*, 9415. [CrossRef] [PubMed]
20. Zhang, W.; Deng, L.; Zhang, L.; Wu, D. A survey on negative transfer. *IEEE CAA J. Autom. Sin.* **2023**, *10*, 305–329. [CrossRef]
21. Wang, Z.; Dai, Z.; Póczos, B.; Carbonell, J. Characterizing and avoiding negative transfer. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 July 2019; pp. 11285–11294. [CrossRef]
22. Geifman, Y.; El-Yaniv, R. SelectiveNet: A deep neural network with an integrated reject option. *Proc. Mach. Learn. Res.* **2019**, *97*, 2151–2159.
23. Angelopoulos, A.N.; Bates, S.; Candès, E.J.; Jordan, M.I.; Lei, L. Learn then test: Calibrating predictive algorithms to achieve risk control. *Ann. Appl. Stat.* **2025**, *19*, 1641–1662. [CrossRef]
24. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
25. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Stat.* **1979**, *7*, 1–26. [CrossRef]
26. Gallego, S.V.; Perez-Mato, J.M.; Elcoro, L.; Tasci, E.S.; Hanson, R.M.; Momma, K.; Aroyo, M.I.; Madariaga, G. MAGNDATA: Towards a database of magnetic structures. I. The commensurate case. *J. Appl. Crystallogr.* **2016**, *49*, 1750–1776. [CrossRef]
27. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [CrossRef]