It's Not Leakage: The Crystal-Structure Advantage in Materials Property Prediction is Real and Property-Specific

Ali Asaria
Transformer Lab

Tony Salomone
Transformer Lab

Deep Gandhi*
Transformer Lab

Abstract

Structure-aware graph neural networks (GNNs) beat composition-only models on the Matbench materials-property benchmark, often by large margins. A common suspicion is that this advantage is inflated by benchmark leakage, near-identical crystals shared between train and test, so that the GNN is partly memorizing rather than generalizing. We test this hypothesis directly and find it false. Across the Matbench regression tasks we study, structural near-duplicate rates are negligible (0% on perovskites, $\leq$2.8% on elastic moduli even under permissive matching), and the large same-composition collision rates (up to 98%) reflect genuine polymorphism, not near-identical structure. Instead we identify a different, real benchmark artifact: because many crystals share a formula but differ in target, composition-only models face an irreducible error floor, 0.379 eV/atom on perovskites, 67% of the target mean absolute deviation, that no per-formula composition model can beat. Decomposing the composition-vs-structure gap with a compact from-scratch CGCNN, we find the structure advantage genuine on all five tasks but strongly property-specific (spanning $\sim$40$\times$): enormous for perovskite formation energy, modest for bulk formation energy and elastic moduli, and small for band gap, where appended structural descriptors actually *hurt* and only end-to-end learning recovers a small gain. A distance-resolved diagnostic corroborates genuine generalization: for bulk modulus the advantage *grows* with distance from the training set, and structural tasks retain a positive gap under an out-of-distribution cluster-split, the opposite of a memorization signature. The structure-GNN advantage is not a leakage artifact; where composition looks weakest, a composition-degeneracy floor is largely responsible.

1 Introduction

Machine-learning models for inorganic crystal properties fall into two families: *composition-only* models that see only the chemical formula [1, 2], and *structure-aware* graph neural networks that also see the 3D atomic arrangement [3, 4]. On the standardized Matbench benchmark [5], structure GNNs consistently win, and a recurring concern in the community is that these wins are inflated by *data leakage*: if near-identical crystals appear in both the training and test folds, a structure model can memorize rather than generalize, exactly as documented in other scientific benchmarks [6, 7]. Whether the Matbench structure advantage survives leakage control is, to our knowledge, untested.

We set out to measure how much of the advantage is leakage, and found that the premise does not hold. Our contributions are:

*Corresponding author: deep@lab.cloud

1. **The structure advantage is not near-duplicate leakage** (§5). Structural near-duplicate rates in Matbench are negligible and robust to matching tolerance; the alarming same-composition collisions are genuine polymorphism.

2. **A composition-degeneracy floor** (§5). Polymorphism imposes an irreducible error floor on per-formula composition models, 67% of the target mean absolute deviation on perovskites, which explains where composition "looks bad" as a benchmark-construction effect, not model weakness.

3. **A per-property decomposition of the structure advantage** (§5), using a compact from-scratch CGCNN and controlled descriptor baselines. Structure genuinely helps on all five tasks but with property-specific magnitude spanning ∼40×; band gap is the one task where appended structural descriptors hurt and only end-to-end learning helps.

4. **Evidence the advantage reflects generalization, not memorization** (§5). A distance-resolved diagnostic and an out-of-distribution cluster-split show the structure gap holds, and for bulk modulus grows, away from the training set.

2 Related Work

Composition vs. structure. Composition-only models such as Roost [1] and CrabNet [2] predict properties from stoichiometry alone, deliberately trading away polymorph resolution for structure-free screening. Structure GNNs, CGCNN [3] and ALIGNN [4], consistently lead the Matbench leaderboard [5]. Tian et al. [8] report that composition alone rivals composition+structure for near-hull compounds on random splits; we test whether such parity persists once leakage and the degeneracy floor are accounted for, and find that structure does help. Multimodal fusions confirm structure's formation-energy dominance on random splits [9, 10].

Leakage and benchmark integrity. Data-leakage taxonomies distinguish train/test overlap from clustered-data i.i.d. violations [7]; in drug discovery, near-duplicate leakage lets a zero-parameter memorization baseline match SOTA [6]. In materials, benchmark-integrity reviews warn that duplicates deflate effective test size and recommend reporting error against similarity-to-training [11], and Matbench-Discovery deduplicates train/test via protostructure matching [12]. Martirossyan et al. [13] show pymatgen's `StructureMatcher` tolerances are often too permissive, motivating our tolerance sweep.

Out-of-distribution splits. SOAP-descriptor leave-cluster-out (SOAP-LOCO) [14] and clustering-based validation splits [15, 16] construct harder, distribution-shifted test sets; Segal et al. [17] caution that domain leave-cluster-out can reduce to interpolation. Redundancy and coreset methods [18, 19] and GNN memorization studies [20] inform our distance-resolved diagnostic. Prior work does not decompose the composition-vs-structure gap under strict near-duplicate control, which is our focus.

3 Method

Structural near-duplicate detection. For each task we group crystals by reduced formula (only same-composition crystals can be duplicates), then within each block run pymatgen `StructureMatcher` (primitive-cell, scaled), forming near-duplicate connected components. Because default tolerances are

permissive [13], we sweep tight/default/loose settings. To scale to 10^5 crystals we add a space-group prefilter (matches require a shared space group).

Composition-degeneracy floor. Because many crystals share a reduced formula but carry different DFT targets, any predictor constrained to one value per formula has an irreducible in-sample error. We report the within-formula mean absolute deviation of the target as this floor; the exact MAE-minimizing per-formula predictor is the within-formula median, so our mean-based value is a close upper estimate of the true floor. The floor is a bound on *composition* models, not a model of structure's skill.

Decomposition. We compare composition and structure models by the gap $\Delta = \mathrm{MAE(comp)} - \mathrm{MAE(struct)}$ (positive means structure helps). We use two structure fidelities: (i) a *controlled descriptor* baseline, the same RandomForest as the composition model (Magpie features) with structural SOAP features appended, isolating the value of structural information under a fixed algorithm; and (ii) a compact *from-scratch CGCNN* (learnable atom embeddings, Gaussian-expanded bond distances, three gated convolution layers, mean pooling, MLP head), implemented in pure PyTorch without `dgl`. The CGCNN is a from-scratch reference, not a tuned leaderboard model.

Distance-resolved diagnostic. For each test crystal we compute the nearest-neighbor distance to its training set in SOAP-PCA space and bin test crystals into distance quartiles, reporting the structure gap per quartile. A memorization story predicts the advantage concentrates near the training set; genuine generalization predicts it holds or grows with distance. This diagnostic uses the descriptor (RandomForest) model, for which per-crystal predictions were available; we return to this caveat in Results and Limitations.

4 Experimental Setup

Data. Five Matbench regression tasks [5]: `mp_e_form` (132,752 crystals), `mp_gap` (106,113), `log_gvrh` and `log_kvrh` (10,987 each, shared structures), and `perovskites` (18,928). Targets are the DFT values shipped with each task, used as-is; all tasks have zero missing entries. The `matbench` package is unusable on modern toolchains (it pins `scikit-learn==1.0.1`); we load the identical data via `matminer` and reproduce the official five-fold nested CV with `KFold` seed 18012019.

Splits. (A) the stock random NCV, and (B) a SOAP-cluster leave-out out-of-distribution split built from a species-agnostic geometric SOAP (fixed 105-dim, tractable at 10^5) reduced by PCA and k-means clustered.

Models. Composition: Magpie element-property features with a RandomForest (an Automatminer-style reference; stronger composition models such as CrabNet/Roost would narrow the gaps we report). Structure: the descriptor+RF baseline above, and the from-scratch CGCNN trained with L1 loss, Adam, and target normalization fit on train. CGCNN training budgets were 100/80/80/40/100 epochs for perovskites/`mp_e_form`/`mp_gap`/`log_kvrh`/`log_gvrh`; the uneven budget is a limitation of the cross-task magnitude comparison. Experiments ran on single NVIDIA A10 GPUs via a cloud provider; the CGCNN uses a CUDA-matched torch build to avoid the `dgl` dependency entirely.

5 Results

Leakage is negligible. Same-composition collision rates are high (0.98 perovskites, 0.41 `mp_e_form`, 0.37 `mp_gap`, 0.19 elastic), but the true *structural* near-duplicate rate is near zero on every task we measured to completion: 0.000 on perovskites, 0.016 on elastic, and $\sim$0.0002 on `mp_gap` (measured

Table 1: Per-property decomposition. Structure gap = MAE(comp) − MAE(struct); positive means structure helps. Composition is Magpie+RF (5-fold); CGCNN is our from-scratch model (epochs per task in text); descriptor gap is Magpie+SOAP+RF. Degeneracy floor is the irreducible per-formula composition MAE. Values are point estimates in task units (eV/atom for formation energies, eV for band gap, $\log_{10}$GPa for moduli).

Task	Comp. MAE	CGCNN MAE	GNN gap	Descriptor gap	Deg. floor
perovskites	0.581	0.071	+0.510	+0.362	0.379
`mp_e_form`	0.122	0.066	+0.056	+0.029	0.027
`mp_gap`	0.329	0.295	+0.034	−0.027	0.062
`log_kvrh`	0.083	0.068	+0.015	+0.014	0.012
`log_gvrh`	0.105	0.093	+0.012	+0.014	0.014

over the first ∼75% of candidate pairs before the sweep was stopped). We did not run the exact 10^5-scale sweep to completion on `mp_e_form`; the partial evidence and the strong tolerance-robustness elsewhere make a materially higher rate there unlikely, but we do not claim a completed count. The finding is robust to `StructureMatcher` tolerance: perovskites stays at 0.000 across tight, default, and loose settings, and elastic rises only to 0.028 under deliberately over-permissive matching. A de-duplicated split is therefore essentially identical to the random split, and near-duplicate leakage cannot explain the structure advantage.

The composition-degeneracy floor. Polymorphism imposes a large irreducible floor on composition-only error where it is severe: 0.379 eV/atom on perovskites (67% of the target mean absolute deviation), and small elsewhere (2.7–4.9%). On perovskites the measured composition MAE (0.581) is 0.202, 53%, above this floor, so composition leaves substantial error beyond the floor; but the floor alone already caps any per-formula composition predictor at 0.379, while the CGCNN (0.071) beats it outright, resolving polymorphs to which composition is blind.

Per-property decomposition. Table 1 reports the full decomposition. The CGCNN beats composition on all five tasks, the structure advantage is real, but the magnitude is strongly property-specific, from +0.510 (perovskites) to +0.012 (shear modulus), roughly a 40× span (we caution the small gaps are point estimates, not tested against fold variance, and the cross-task magnitudes are affected by the uneven CGCNN training budget above). Band gap is the task where the *descriptor* approach fails: appended structural features *hurt* it (−0.027), and only the end-to-end CGCNN recovers a small gain (+0.034, beating composition's 0.329). End-to-end learning thus extracts a band-gap signal from structure that appended features cannot, though the band-gap GNN gain is the paper's smallest clear structural effect and we do not over-read it.

Out-of-distribution split. Under the SOAP-cluster leave-out split (B), the descriptor structure gap stays positive for the structural tasks, perovskites +0.295, `mp_e_form` +0.063, `log_kvrh` +0.022, `log_gvrh` +0.023, and stays negative for band gap (−0.030). Structure's value survives, and for formation energy even exceeds, a deliberately harder distribution-shift test, consistent with genuine generalization rather than interpolation to nearby training crystals.

Distance-resolved diagnostic. Table 2 and Figure 2 bin test crystals by distance to the training set. For bulk modulus (`log_kvrh`) the descriptor gap grows monotonically from +0.003 (nearest quartile) to +0.033 (farthest), the advantage is largest exactly where memorization is impossible. Perovskites keeps a large gap in every quartile but, unlike `log_kvrh`, *declines* with

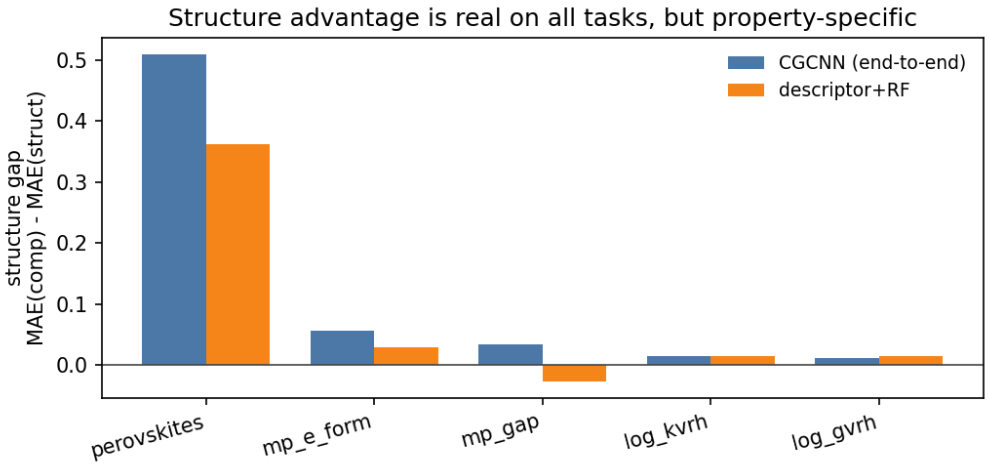

Figure 1: The structure advantage (gap = MAE(comp) − MAE(struct)) is positive for the end-to-end CGCNN on all five tasks but strongly property-specific. Appended structural descriptors (orange) hurt band gap, whereas the end-to-end GNN (blue) helps. Point estimates.

Table 2: Distance-resolved structure gap (descriptor model) by nearest-neighbor distance-to-train quartile, Q1 (near) to Q4 (far), in task units. For bulk modulus the gap grows with distance; for band gap the descriptor gap worsens; perovskites stays large but declines.

Task	Q1 (near)	Q2	Q3	Q4 (far)
`log_kvrh` (bulk modulus)	+0.003	+0.011	+0.018	+0.033
perovskites	+0.422	+0.429	+0.365	+0.238
`mp_gap` (band gap)	−0.001	−0.018	−0.036	−0.054

distance (+0.422 → +0.238); the advantage there is largest near the training set, though far-quartile crystals still show a substantial +0.238. For band gap the descriptor gap worsens with distance (−0.001 → −0.054), i.e. appended structural features fail to generalize while composition holds. The per-crystal-system breakdown shows these effects are systemic, not driven by one system (perovskites: cubic +0.392, orthorhombic +0.347, tetragonal +0.365). We stress this diagnostic uses the descriptor (RandomForest) model, not the CGCNN; a RandomForest on SOAP features does not memorize individual training crystals the way an end-to-end GNN might, so these trends corroborate the OOD-split result rather than directly testing GNN memorization; a per-crystal CGCNN version is future work.

6 Discussion

The results reframe a common assumption. The materials-ML community often discounts structure-GNN gains as possible leakage; we find near-duplicate leakage is not the mechanism on the tasks we study. Two effects that *are* real were being conflated with it. First, polymorphism creates a composition-degeneracy floor: when a benchmark pairs one formula with several differently-valued structures, per-formula composition models are mathematically capped, and their large error is partly a property of the benchmark. This bounds how good composition can look; it does not itself

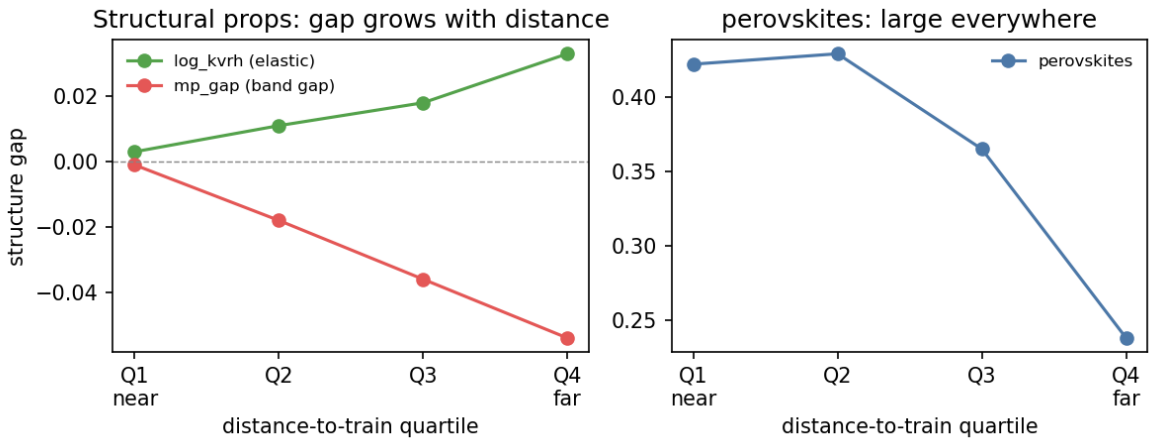

Figure 2: Distance-resolved structure gap (descriptor model, $\log_{10}$GPa for moduli / eV for band gap) by distance-to-train quartile. Bulk modulus (left, green) grows with distance; band gap (left, red) worsens; perovskites (right) stays large but declines. Descriptor-model diagnostic; see text.

explain structure's skill: the CGCNN's perovskite error (0.071) is far below the floor (0.379), so most of structure's advantage there is resolving polymorphs the floor grants composition no credit for. Second, the genuine structural signal is strongly property-specific and, for structural properties, survives distribution shift and (for bulk modulus) grows with distance from training data. Band gap is the honest exception: it is largely composition/electronic-determined, appended structural descriptors hurt, and even an end-to-end GNN gains only a little.

7 Limitations

Statistical. All reported metrics are point estimates; we do not report fold-level variance or significance, so the smallest gaps (+0.012 to +0.034) should be read as directional rather than as established fine-grained rankings, and the many task×model×quartile×subgroup comparisons are exploratory (no multiple-comparison correction). *Construct.* The CGCNN is a compact from-scratch model, not tuned to leaderboard SOTA; on both moduli the descriptor RF matches or slightly exceeds it, so the CGCNN is not an upper bound on structural value. Symmetrically, the composition baseline is Magpie+RF, weaker than CrabNet/Roost; a stronger composition model would shrink every reported gap. The CGCNN training budget is uneven across tasks (40–100 epochs), confounding the cross-task magnitude comparison. *Internal.* The distance-resolved and species-aware analyses use the descriptor (RandomForest) structure model, not the CGCNN, because per-crystal predictions were available there; the diagnostic corroborates but does not directly test GNN memorization, and it depends on a single distance metric (SOAP-PCA) whose sensitivity we did not probe. Structural near-duplicate rates are threshold-dependent (we show robustness across a sweep), and the `mp_e_form` exact sweep was not completed. *External.* We study five Matbench DFT regression tasks; DFT (PBE) labels encode systematic band-gap underestimation, and findings may not transfer to experimental targets or other benchmarks.

8 Availability

Artifacts, code, the leakage-checked splits, and analysis, are available from the authors on request. All splits derive deterministically from the public Matbench data and the fixed seed reported above.

9 Conclusion and Future Work

Across five Matbench tasks, the structure-GNN advantage is not a near-duplicate leakage artifact: leakage is negligible, the advantage is genuine on every task, and for structural properties it survives distribution shift. Where composition appears to fail badly, a composition-degeneracy floor from polymorphism is largely responsible. Future work: a fully-tuned and budget-equalized model matrix with fold-level variance, a stronger composition baseline, per-crystal CGCNN distance-resolved curves, and extension to experimental (non-DFT) targets.

References

[1] Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. 2019. URL https://arxiv.org/abs/1910.00617.

[2] Anthony Yu-Tung Wang, Steven K. Kauwe, Ryan J. Murdock, and Taylor D. Sparks. Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 2021.

[3] Tian Xie and Jeffrey C. Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. 2017. URL https://arxiv.org/abs/1710.10324.

[4] Kamal Choudhary and Brian DeCost. Atomistic Line Graph Neural Network for Improved Materials Property Predictions. 2021. URL https://arxiv.org/abs/2106.01829.

[5] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6:138, 2020.

[6] Amber Huang, Ian Scott Knight, and Slava Naprienko. Data Leakage and Redundancy in the LIT-PCBA Benchmark. 2025. URL https://arxiv.org/abs/2507.21404.

[7] Christina Sauer, Anne-Laure Boulesteix, Luzia Hanßum, Farina Hodiamont, Claudia Bausewein, and Theresa Ullmann. Beyond algorithm hyperparameters: on preprocessing hyperparameters and associated pitfalls in machine learning applications. 2024. URL https://arxiv.org/abs/2412.03491.

[8] Siyu Isaac Parker Tian, Aron Walsh, Zekun Ren, Qianxiao Li, and Tonio Buonassisi. What Information is Necessary and Sufficient to Predict Materials Properties using Machine Learning? 2022. URL https://arxiv.org/abs/2206.04968.

[9] Janghoon Ock, Joseph Montoya, Daniel Schweigert, Linda Hung, Santosh K. Suram, and Weike Ye. UniMat: Unifying Materials Embeddings through Multi-modal Learning. 2024. URL https://arxiv.org/abs/2411.08664.

[10] Jithendaraa Subramanian, Linda Hung, Daniel Schweigert, Santosh Suram, and Weike Ye. XxaCT-NN: Structure Agnostic Multimodal Learning for Materials Science. 2025. URL https://arxiv.org/abs/2507.01054.

[11] Nawaf Alampara, Mara Schilling-Wilhelmi, and Kevin Maik Jablonka. Lessons from the trenches on evaluating machine-learning systems in materials science. 2025. URL https://arxiv.org/abs/2503.10837.

[12] Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. Matbench Discovery – A framework to evaluate machine learning crystal stability predictions. 2023. URL https://arxiv.org/abs/2308.14920.

[13] Maya M. Martirossyan, Thomas Egg, Philipp Hoellmer, George Karypis, Mark Transtrum, Adrian Roitberg, Mingjie Liu, Richard G. Hennig, Ellad B. Tadmor, and Stefano Martiniani. All that structure matches does not glitter. 2025. URL https://arxiv.org/abs/2509.12178.

[14] Liqin Tan, Pin Chen, Menghan Liu, Xiean Wang, Jianhuan Cen, and Qingsong Zou. Benchmarking GNNs for OOD Materials Property Prediction with Uncertainty Quantification. 2025. URL https://arxiv.org/abs/2511.11697.

[15] Andrea Napoli and Paul White. Clustering-Based Validation Splits for Model Selection under Domain Shift. 2024. URL https://arxiv.org/abs/2405.19461.

[16] Michal Sanocki and Julija Zavadlav. Generalization of Long-Range Machine Learning Potentials in Complex Chemical Spaces. 2025. URL https://arxiv.org/abs/2512.10989.

[17] Nofit Segal, Aviv Netanyahu, Kevin P. Greenman, Pulkit Agrawal, and Rafael Gomez-Bombarelli. Known Unknowns: Out-of-Distribution Property Prediction in Materials and Molecules. 2025. URL https://arxiv.org/abs/2502.05970.

[18] Xuqiang Shao, Yuqi Zhang, Di Zhang, Zhaoyan Dong, Tianxiang Gao, Mingzhe Li, Xinyuan Liu, Zhiran Gan, Fanshun Meng, Lingcai Kong, Zhengyang Gao, Hao Lic, and Weijie Yangd. Data-Efficient Machine Learning Potentials via Difference Vectors Based on Local Atomic Environments. 2025. URL https://arxiv.org/abs/2501.16398.

[19] Sazzad Hossain, Ponkrshnan Thiagarajan, Shashank Pathrudkar, Stephanie Taylor, Abhijeet S. Gangan, Amartya S. Banerjee, and Susanta Ghosh. Surprisingly High Redundancy in Electronic Structure Data Across Materials Explained by Low Intrinsic Dimensionality. 2025. URL https://arxiv.org/abs/2507.09001.

[20] Adarsh Jamadandi, Jing Xu, Adam Dziedzic, and Franziska Boenisch. Memorization in Graph Neural Networks. 2025. URL https://arxiv.org/abs/2508.19352.