Studies in Knowledge and Innovation
Vol. 2 No. 11 (2026)

A Held-Out Material Can Be in Training for Another Reference: An Audit of Two JARVIS Band-Gap Tasks

Rebecca Berry · Ellen Perkins · Simon L Mitchell

Published: November 26, 2026 | Article 1113

Abstract

The JARVIS-Leaderboard releases the same property under two electronic-structure methods with different evaluation partitions. Matching exact material identifiers showed that 1,463 of the 1,815 TBmBJ test materials were assigned to training for the OptB88vdW task. Only 187 materials were held out under both published assignments. On those 187, the mean absolute difference between the two reference labels was 0.578 eV. The published coGN predictions for OptB88vdW scored 0.105 eV against that reference and 0.608 eV against TBmBJ, while the coGN predictions for TBmBJ scored 0.654 and 0.277 eV, respectively. ALIGNN showed the same reversal. Combining references across benchmark tasks therefore changes both which materials are genuinely held out and what a reported error measures.

A Held-Out Material Can Be in Training for Another Reference: An Audit of Two JARVIS Band-Gap Tasks

Abstract

The JARVIS-Leaderboard releases the same property under two electronic-structure methods with different evaluation partitions. Matching exact material identifiers showed that 1,463 of the 1,815 TBmBJ test materials were assigned to training for the OptB88vdW task. Only 187 materials were held out under both published assignments. On those 187, the mean absolute difference between the two reference labels was 0.578 eV. The published coGN predictions for OptB88vdW scored 0.105 eV against that reference and 0.608 eV against TBmBJ, while the coGN predictions for TBmBJ scored 0.654 and 0.277 eV, respectively. ALIGNN showed the same reversal. Combining references across benchmark tasks therefore changes both which materials are genuinely held out and what a reported error measures.

Keywords — JARVIS-DFT, reference substitution, material identity, comparison of benchmark splits, multi-fidelity evaluation

1. Introduction

A benchmark can store two labels for a material, each obtained from a specified computational method. The distinction is often preserved in the dataset field name. It can disappear when several tasks are assembled for evaluation, either because training and test assignments are copied separately from each task or because the same nominal property is scored against a different method's reference. These are independent decisions. A material can qualify as held out in one task while its alternative label is used in training for the other. Even when both tasks hold the material out, their reference values can imply different errors for a prediction that has not changed.

The JARVIS electronic-structure data provide a concrete way to study the distinction. The two band-gap references, OptB88vdW and the Tran–Blaha modified Becke–Johnson description, have published differences in how they approximate or supply the relevant quantities (Choudhary et al., 2018). The broader computational context includes the density-functional formulation (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), work on derivatives and gaps in Kohn–Sham descriptions (Perdew & Levy, 1983), and comparisons of semilocal exchange and correlation approximations (Perdew et al., 1996). OptB88vdW and the modified Becke–Johnson potential belong to distinct methodological choices, with specific empirical behavior and limitations demonstrated in the literature (Klimeš et al., 2010; Tran & Blaha, 2009; Koller et al., 2011). Labels created under two such choices give related observations about a material, with an identifiable provenance for each stored value.

Li et al. (2026) combine graph and text inputs for predicting properties of catalytic materials. Their setting motivates making identity and label provenance explicit when models draw information from multiple representations or benchmark tasks. An input modality or graph architecture can improve the quality of a predictor while leaving unresolved the question of which reference it approximates and whether the evaluated material was exposed under another training task. This paper analyzes independent public JARVIS records and posted graph-model predictions. It examines that protocol question for two computed band-gap references, without reusing the target paper's model or task as the evaluation source.

Research questions. First, how much do nominal training and test assignments overlap when the two public band-gap tasks are joined on exact material identity? Second, on materials jointly held out by the two task manifests, how does the same set of published predictions score when the reference method changes?

The empirical expectation is that independent task manifests need not impose a common identity holdout. A separate expectation is that swapping the reference favors outputs associated with the matching computed target. The score comparison uses the same material identifiers and prediction files under both references so that switching the response is isolated from changing the set of evaluated cases.

2. Source Records and Identity Audit

The analysis obtains two zipped JSON benchmark disclosures from the NIST JARVIS-Leaderboard's OptB88vdW regression task and TBmBJ regression task. We use OPT and MBJ as abbreviations. Each JSON has separately named training, validation and test dictionaries keyed by exact JVASP material identifiers, with band gaps in eV. The OPT manifest contains 55,713 role entries and the MBJ manifest contains 18,167. Exposures are recorded at the exact material ID level. This prevents counting two differently labeled records of one material as two independent physical cases, while retaining distinct material IDs even if they could have similar or duplicated structures.

Before comparing predictions, we checked whether either manifest itself assigns the same ID to multiple roles. We then formed the three-by-three table of material-ID intersections between task roles. Its cell counts can include the same ID in several cells when that ID appears in more than one role in a source manifest. The overall number of linked materials is separately counted over unique IDs. Finally, the analysis selected the intersection of the two test dictionaries and removed any identifier appearing in either task's training or validation dictionary. This operational definition draws on the requirement to preserve meaningful grouping in validation (Roberts et al., 2017). It is stricter than taking the intersection of nominative test roles alone. More demanding checks would require a structure identity audit, a characterization of chemical similarity, or information about prior training exposures, as used for more comprehensive out-of-distribution benchmark design (Omee et al., 2024; Chen et al., 2026).

One reason for keeping the stages explicit is that data merging errors can occur at a higher level than column names. A learner that is fitted to the union of separately generated training dictionaries may consume a different label for a nominal test identity. Leakage work highlights this distinction between an allowed input in a particular application and information made available through an evaluation setup that was supposed to hold the unit out (Kaufman et al., 2012; Kapoor & Narayanan, 2023). Here the cross-role counts are a statement about the published manifests. They do not establish which files any contributor actually used for an individual fit.

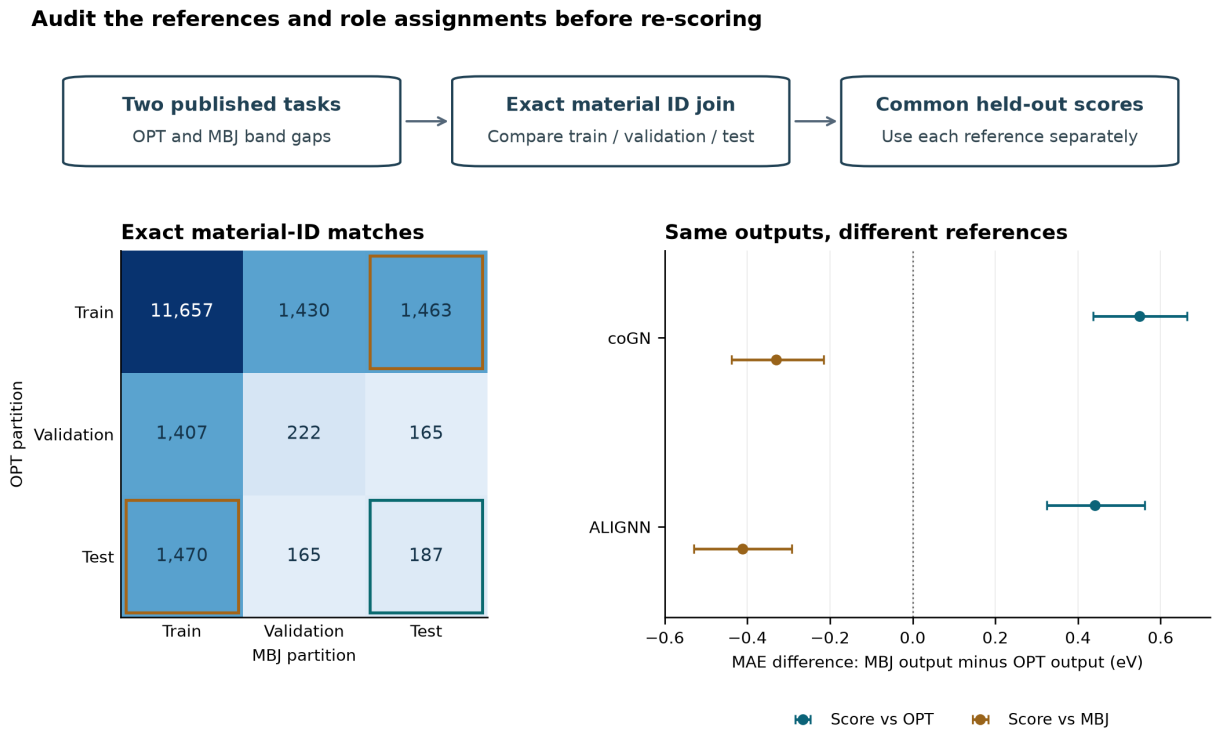


Figure 1. The overlap matrix counts exact ID matches between roles. The prediction contrast is mean absolute error of the MBJ-task output minus the OPT-task output for a common held-out group. Positive values favor the OPT-task output.

3. Frozen Prediction Files and Measures

The two architectures are connectivity-optimized graph networks (coGN) and atomistic line graph neural networks (ALIGNN), each with one reference-specific public submission per task (Ruff et al., 2024; Choudhary & DeCost, 2021). Graph-based materials prediction systems can build representations with different choices about neighborhood and connectivity, as seen in crystal graph convolution, continuous-filter models and models shared across molecules and crystals (Xie & Grossman, 2018; Schütt et al., 2018; Chen et al., 2019). Selecting two families tests whether the reference sensitivity appears in more than one file pair. Architectures, hyperparameters, and the amount of data used by each original contribution are outside the estimand. The posted vectors are treated as fixed outputs, with no fitting performed in this study.

Every submission ID was required to match the complete official test role in its own task; where the submission provided a target column, that column was checked against the official value to one millionth of an eV. The coGN submissions give ID and prediction without the target, so their references come directly from the benchmark JSON. In each paired comparison, all four submitted predictions are present for every one of the same 187 IDs. Predictions are retained as submitted, including small negative outputs. The analysis computes the mean absolute error relative to each reference. Mean absolute error weights materials equally and corresponds to a prediction loss whose optimal point target is a conditional median (Gneiting, 2011). This interpretation is useful only after the target distribution is identified.

We compute three quantities from the paired rows. First, absolute and signed differences between the references characterize the discrepancy in measurement convention. Second, for each fixed submission, the loss difference after switching from OPT scoring to MBJ scoring shows its sensitivity to the scoring reference. Third, the MAE difference of the MBJ-task output minus the OPT-task output is computed within each architecture and held-out reference. A simple numerical check follows from the triangle inequality: the absolute change in one row’s prediction loss after switching references cannot exceed the

absolute difference between the two reference values themselves. The analysis checked this property on every scored row.

Uncertainty in mean contrasts is summarized by 4,000 paired percentile bootstrap resamples of the 187 shared test identities (Efron, 1979; Davison & Hinkley, 1997). The prediction and the two labels of each sampled ID remain together. Intervals show dispersion from the observed holdout cases with all four predictions held fixed. The archive manifests do not provide similarity clusters for these 187 entries, so the intervals do not capture dependence between closely related materials or variation caused by retraining. Calibrating uncertainty beyond this matched set would require more information about the specific target distribution (Pernot, 2022).

4. Results

In the released role dictionaries, two OPT IDs and one MBJ ID appear in both their own training and test roles. The respective task labels are identical for those repeated IDs. No fit/validation or validation/test duplicate IDs occur inside either manifest. The cross-task join produced 18,165 unique IDs. Table 1 shows the role intersections. Among the 1,815 MBJ test entries, 1,463 (80.61%) appear in the OPT training dictionary and 165 appear in the OPT validation dictionary. In the other direction, 1,470 OPT test IDs appear in MBJ training out of 5,572 OPT test entries. All 187 shared test records remain after also excluding any role from either fitting or validation dictionary. The clean exact-ID intersection therefore represents 10.30% of the original MBJ test role and 3.36% of the original OPT test role.

Table 1. Published split-role intersections between the two benchmark archives.

OPT role	MBJ train	MBJ validation	MBJ test
Train	11,657	1,430	1,463
Validation	1,407	222	165
Test	1,470	165	187

Cells count exact ID intersections. The small number of within-manifest repeated IDs means the sum can exceed the number of shared unique identifiers.

Within the 187 evaluation pairs, the mean signed MBJ-minus-OPT difference was +0.577 eV, while the mean absolute difference was 0.578 eV (95% interval [0.461, 0.693]). The median absolute discrepancy was zero and the ninetieth percentile was 1.892 eV. Six records changed from an archived zero in OPT to a positive number in MBJ. Therefore agreement at zero and substantial positive-label discrepancies can coexist in a single aggregate. Switching the reference can have little effect on many individual entries yet markedly change MAE for the evaluated group.

Table 2. Public graph-model outputs scored on the same 187 jointly held-out IDs, in eV.

Submission	Full native test	Matched vs OPT	Matched vs MBJ	MBJ − OPT (95% interval)
coGN for OPT	0.122	0.105	0.608	+0.503 [0.395, 0.613]
coGN for MBJ	0.264	0.654	0.277	−0.377 [−0.481, −0.269]
ALIGNN for OPT	0.142	0.173	0.674	+0.501 [0.389, 0.619]
ALIGNN for MBJ	0.310	0.614	0.262	−0.351 [−0.453, −0.250]

The first numerical column reproduces the separate leaderboard evaluations. Remaining columns use the identical matched IDs. Intervals use paired resampling of these IDs.

The within-architecture comparison illustrates the same phenomenon from the perspective of choosing which output to use. On OPT, the MBJ-task coGN output has a mean error higher by 0.549 eV [0.437, 0.664] than the OPT-task output. On MBJ that difference becomes −0.331 eV [−0.439, −0.215]. ALIGNN shows the same reversal, with differences of +0.441 eV [0.324, 0.562] on OPT and −0.411 eV [−0.529, −0.293] on MBJ. These differences arise while holding submissions and tested instances fixed. They quantify the consequences of changing the reference for these outputs, and do not estimate the effect of changing the target at training time with a fixed training dataset.

5. Discussion

Kennedy and O'Hagan (2000), Perdikaris et al. (2017), and Peherstorfer et al. (2018) describe how modeling can combine observations from different levels of approximation while keeping their roles explicit. Graph networks also provide an example of how different calculation fidelities can be modeled jointly (Chen et al., 2021). The manifest audit addresses an accompanying evaluation decision: whether the same physical identity can enter fitting through one task and evaluation through another. If a study aims to evaluate completely unseen material IDs across tasks, enforcing one grouped partition before selecting labels satisfies that requirement. If it aims to estimate a higher-level value for materials with a lower-level result already known, then the lower-level number defines an explicitly conditional prediction task. Under either design, naming the source of the prediction and reference makes an error value interpretable.

The two test roles also demonstrate the difference between checking a source score and comparing scores. Reproducing a posted MAE verifies that the analysis joined each submission with the intended default reference. It does not make the original test sets identical. A protocol such as Matbench prescribes task splits to make within-task comparisons reproducible (Dunn et al., 2020). Cross-task comparison needs an additional decision about how the role assignments and label conventions interact. Comparing a limited common subset controls material identity and score definition at the cost of coverage. The full-size evaluations remain the appropriate records of each submission's performance on its own task.

The matrix adds useful detail about coverage. Counting how many OPT test entries appear in MBJ training answers one question for the complete OPT archive. Conditioning on the OPT test entries that are present in any MBJ role answers a different question for the shared subset. The corresponding denominator is smaller because MBJ has fewer released materials. This asymmetry is visible even before fitting an estimator or examining a prediction vector. An evaluation report that supplies only one overlap percentage can obscure it. The separate counts in Table 1 make it possible to state which population and which task arrangement drive a potential exposure.

The zero median in the reference comparison also shows why a single summary can obscure the sensitivity of predicted error. A substantial fraction of materials have the same recorded value, while the upper tail contains discrepancies in the range that makes a visible contribution to absolute loss. Those differences can matter even if a new model improves most on labels where the two references agree. A paired error table measures this without attempting to choose which simulation should be considered more accurate in nature. In particular, an observed difference between simulations cannot serve as an experimental uncertainty estimate.

The direction and magnitude of the reference effect apply to the 187 dual-held-out examples and the four chosen submissions. They do not establish the ranking of methods relative to experimental measurements. A future head-to-head comparison would require a third reference or a separately collected evaluation, with all selection decisions isolated from that evaluation (Cawley & Talbot, 2010; Varma & Simon,

2006). The existing records show why preserving the source reference and explicitly joining partitions by material identity can materially change the interpretation of multi-source results.

6. Conclusion

In the JARVIS band-gap releases, many nominal test materials have a different training role under the alternative reference. The jointly held-out materials also show sizable changes in model error when the software outputs stay fixed and the scored reference changes. Reporting both exact identity exposure and the calculation method attached to each label makes clear whether a benchmark assesses a new material, a new reference for an already known material, or the accuracy of a particular output within its native task.

Data and Code Availability

The two OPT and MBJ task pages link the published reference and submission archives used here, accessed 23 November 2026. The project terms describe use of the repository software and documentation. The accompanying result record gives SHA-256 for all six inputs, and the analysis and figure scripts reproduce the reported aggregates.

References

Cawley, G. C.; Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11, 2079–2107. https://jmlr.org/papers/v11/cawley10a.html

Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. (2019). Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chemistry of Materials*, 31(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294

Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Ong, S. P. (2021). Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science*, 1(1), 46–53. https://doi.org/10.1038/s43588-020-00002-x

Chen, H.; Ye, T. T. (2026). A controlled framework for chemistry-shift evaluation in descriptor-based materials screening. *Computational Materials Science*, 274, 115027. https://doi.org/10.1016/j.commatsci.2026.115027

Choudhary, K.; Zhang, Q.; Reid, A. C. E.; Chowdhury, S.; Van Nguyen, N.; Trautt, Z.; Newrock, M. W.; Congo, F. Y.; Tavazza, F. (2018). Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Scientific Data*, 5(1). https://doi.org/10.1038/sdata.2018.82

Choudhary, K.; DeCost, B. (2021). Atomistic Line Graph Neural Network for improved materials property predictions. *npj Computational Materials*, 7(1). https://doi.org/10.1038/s41524-021-00650-1

Davison, A. C.; Hinkley, D. V. (1997). Bootstrap Methods and their Application. *Cambridge University Press*. https://doi.org/10.1017/cbo9780511802843

Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. (2020). Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6, 138. https://doi.org/10.1038/s41524-020-00406-3

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Gneiting, T. (2011). Quantiles as optimal point forecasts. *International Journal of Forecasting*, 27(2), 197–207. https://doi.org/10.1016/j.ijforecast.2009.12.015

Hohenberg, P.; Kohn, W. (1964). Inhomogeneous Electron Gas. *Physical Review*, 136(3B), B864–B871. https://doi.org/10.1103/physrev.136.b864

Kapoor, S.; Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kaufman, S.; Rosset, S.; Perlich, C.; Stitelman, O. (2012). Leakage in data mining. *ACM Transactions on Knowledge Discovery from Data*, 6(4), 1–21. https://doi.org/10.1145/2382577.2382579

Kennedy, M. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1), 1–13. https://doi.org/10.1093/biomet/87.1.1

Klimeš, J.; Bowler, D. R.; Michaelides, A. (2010). Chemical accuracy for the van der Waals density functional. *Journal of Physics: Condensed Matter*, 22(2), 022201. https://doi.org/10.1088/0953-8984/22/2/022201

Kohn, W.; Sham, L. J. (1965). Self-Consistent Equations Including Exchange and Correlation Effects. *Physical Review*, 140(4A), A1133–A1138. https://doi.org/10.1103/physrev.140.a1133

Koller, D.; Tran, F.; Blaha, P. (2011). Merits and limits of the modified Becke-Johnson exchange potential. *Physical Review B*, 83(19). https://doi.org/10.1103/physrevb.83.195134

Omee, S. S.; Fu, N.; Dong, R.; Hu, M.; Hu, J. (2024). Structure-based out-of-distribution (OOD) materials property prediction: a benchmark study. *npj Computational Materials*, 10(1). https://doi.org/10.1038/s41524-024-01316-4

Peherstorfer, B.; Willcox, K.; Gunzburger, M. (2018). Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization. *SIAM Review*, 60(3), 550–591. https://doi.org/10.1137/16m1082469

Perdew, J. P.; Levy, M. (1983). Physical Content of the Exact Kohn-Sham Orbital Energies: Band Gaps and Derivative Discontinuities. *Physical Review Letters*, 51(20), 1884–1887. https://doi.org/10.1103/physrevlett.51.1884

Perdew, J. P.; Burke, K.; Ernzerhof, M. (1996). Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18), 3865–3868. https://doi.org/10.1103/physrevlett.77.3865

Li, Y.; Xu, J.; Zhang, X. Y.; Xiang, S.; Ran, N.; Li, W.; Liu, C. L. (2026). CatalyticMLLM: A Graph-Text Multimodal Large Language Model for Catalytic Materials. *arXiv*. https://arxiv.org/abs/2605.17254v5

Perdikaris, P.; Raissi, M.; Damianou, A.; Lawrence, N. D.; Karniadakis, G. E. (2017). Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198), 20160751. https://doi.org/10.1098/rspa.2016.0751

Pernot, P. (2022). The long road to calibrated prediction uncertainty in computational chemistry. *The Journal of Chemical Physics*, 156(11). https://doi.org/10.1063/5.0084302

Roberts, D. R.; Bahn, V.; Ciuti, S.; Boyce, M. S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J. J.; Schröder, B.; Thuiller, W.; et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8), 913–929. https://doi.org/10.1111/ecog.02881

Ruff, R.; Reiser, P.; Stühmer, J.; Friederich, P. (2024). Connectivity optimized nested line graph networks for crystal structures. *Digital Discovery*, 3(3), 594–601. https://doi.org/10.1039/d4dd00018h

Schütt, K. T.; Sauceda, H. E.; Kindermans, P. J.; Tkatchenko, A.; Müller, K. R. (2018). SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24). https://doi.org/10.1063/1.5019779

Tran, F.; Blaha, P. (2009). Accurate Band Gaps of Semiconductors and Insulators with a Semilocal Exchange-Correlation Potential. *Physical Review Letters*, 102(22). https://doi.org/10.1103/physrevlett.102.226401

Varma, S.; Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1). https://doi.org/10.1186/1471-2105-7-91

Xie, T.; Grossman, J. C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Physical Review Letters*, 120(14). https://doi.org/10.1103/physrevlett.120.145301