Annals of Science and Inquiry
Vol. 2 No. 11 (2026)

Does Initial Adsorption Energy Improve Prediction of Relaxed Energy? A Held-Out Analysis of the OC20 Tutorial Data

Emily Bradley · Julie Hamilton · Kevin Wilson

Published: November 26, 2026 | Article 1094

Abstract

The Open Catalyst Project's instructional data store two calculated energies for each system. In the 100 training records, the relaxed energy was typically 5.81 eV below the initial value. Subtracting that amount looked reasonable as a quick correction, yet its mean absolute error was 4.10 eV on 20 held-out systems, compared with 2.41 eV for a constant training median. Freely fitted least-squares and robust regressions assigned the initial energy slopes near 0.11, far below the unit slope implied by an additive correction. Their holdout errors were slightly lower than the constant's, but paired bootstrap intervals included zero. In this sample, a calculated initial energy provided little dependable evidence of an improvement after its scale was estimated. The comparison describes an instructional archive in which the initial energy is already available.

Does Initial Adsorption Energy Improve Prediction of Relaxed Energy? A Held-Out Analysis of the OC20 Tutorial Data

Abstract

The Open Catalyst Project's instructional data store two calculated energies for each system. In the 100 training records, the relaxed energy was typically 5.81 eV below the initial value. Subtracting that amount looked reasonable as a quick correction, yet its mean absolute error was 4.10 eV on 20 held-out systems, compared with 2.41 eV for a constant training median. Freely fitted least-squares and robust regressions assigned the initial energy slopes near 0.11, far below the unit slope implied by an additive correction. Their holdout errors were slightly lower than the constant's, but paired bootstrap intervals included zero. In this sample, a calculated initial energy provided little dependable evidence of an improvement after its scale was estimated. The comparison describes an instructional archive in which the initial energy is already available.

Keywords — Open Catalyst 2020, adsorption energy, held-out evaluation, robust regression, baseline design

1. Introduction

The premise of predicting a relaxed energy from information on an initial structure is that expensive simulation results need to be estimated from inputs available earlier. The identity of those inputs matters. A model that begins with atomic information has a different task from a numerical rule that also receives an energy calculated for the particular initial state. The Open Catalyst 2020 dataset and community challenges include a structure-to-relaxed-energy task, using the initial atomic geometry to predict the relaxed adsorption energy (Chanussot et al., 2021). The project’s instructional download also stores an y_init scalar for its example records. That scalar permits a narrower and empirically tractable question about the value of information in a previously computed quantity.

An additive correction is a particularly intuitive use of an initial energy. If the relaxed value was usually lower by about the same amount, one could learn overall displacement from previous records, apply it to new initial values, and retain the numerical order and spacing of the initial estimates. The same approach would perform poorly if the size of the decrease varied systematically with the initial value or if only a few extreme observations accounted for a large share of the apparent relation between the two energy columns. A slope estimated from the training records allows their scales to differ. A constant training median offers an additional reference point because it does not receive any information about individual validation systems.

Li et al. (2026) present CatalyticMLLM, a graph-and-text approach to predictive tasks for catalytic materials including relaxed energy prediction. This motivates clarifying which inputs contribute evidence before comparing multimodal and other complex representations. Crystal graph convolution, universal graph networks, and continuous-filter neural networks illustrate how structural information can be learned rather than passed through as a fixed scalar (Xie & Grossman, 2018; Chen et al., 2019; Schütt et al., 2018). Line graphs and directional message passing further encode relationships among bonds and neighboring configurations (Choudhary & DeCost, 2021; Klicpera et al., 2020). A review of the value of an already available initial energy complements those representation questions. Each approach would require a shared data partition and matched inputs for a direct performance comparison.

Research question. On the archived OC20 instructional holdout, does an already computed initial energy reduce the error of predicting relaxed energy relative to the training median, and how does the answer depend on retaining or replacing a unit slope?

If the displacement has similar central tendency over the range of initial energies, the unit-slope correction should preserve informative differences between systems and outperform the constant. If high initial values exaggerate differences in the relaxed target, allowing the slope to contract may reduce error. These are empirical expectations about two rules under the same held-out score. The energy columns alone do not identify a physical mechanism behind a specific displacement.

2. Dataset and Research Design

The analysis uses the Open Catalyst Project's public tutorial_data.tar.gz download linked in the official instructional dataset guide. The unmodified archive also includes records for other tasks. This study selects the two IS2RE members explicitly labeled train_100 and val_20, giving 100 source records for estimating constants and coefficients and 20 validation records reserved for scoring. The downloader and analysis record the digest of the parent archive and the two selected files. All selected records had finite numeric values for both energy fields. Each archived system identifier occurred once within its partition, and no identifier appeared in both partitions. This check establishes separation by archived system ID. The training and validation files in the tutorial have no documented sampling design that could justify inferring representativeness of the full OC20 dataset from their names or their successful ID check.

For record (i), write (x_i) for the archived initial energy and (y_i) for the archived relaxed energy, in eV. The signed difference ($d_i=y_i-x_i$) describes the stored pair for each record. This difference has a clear sign and numerical meaning for these files. Its interpretation is limited to benchmark calculations, which rely on choices about computational models and records. Density functional theory establishes the foundations of the original calculation framework (Hohenberg & Kohn, 1964; Kohn & Sham, 1965), while work on the dependence of adsorption energetics on exchange-correlation formulations illustrates why computed reference values belong to a specific calculation setup (Hammer et al., 1999). These relationships motivate describing the two energies as archive values and separating them from physical measurements or universal conversion rules.

The analysis reads only system IDs and the two reference energies from the two designated files. It does not use the archived atom positions or chemistry fields. Atomic simulation formats can preserve richer structural data (Hjorth Larsen et al., 2017). Such information could support subsequent analysis with local environment descriptors or graph representations that preserve more structural distinctions (Himanen et al., 2020; Balasingham et al., 2024). Here each rule receives only the disclosed scalar or, for the constant, no information tied to the particular holdout system. That choice keeps observed differences attributable to the mathematical treatment of the given scalar under one consistent partition.

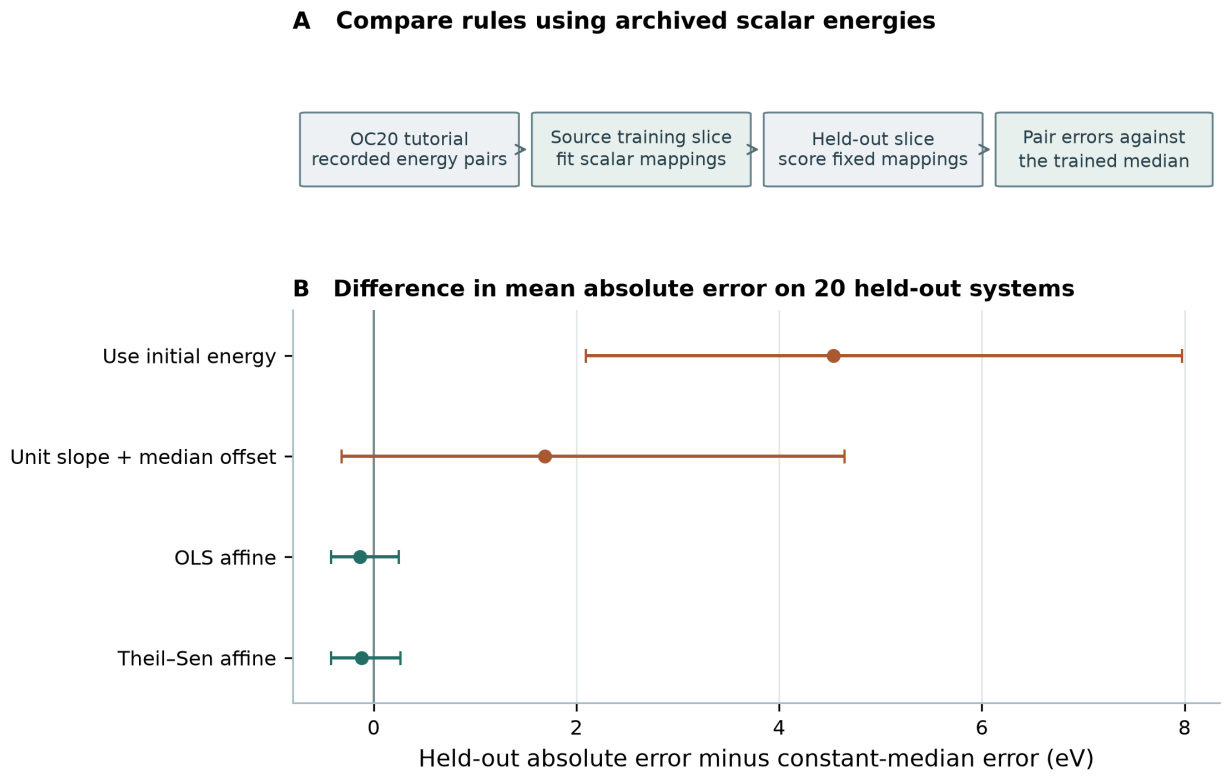

Figure 1. Research design. The training partition supplies location and regression coefficients. The fixed rules are scored on the same 20 other records; paired comparisons preserve common holdout systems.

3. Statistical Methods

The first rule predicts the median training relaxed energy for every record. It targets minimum absolute prediction loss for a constant evaluated on the training distribution. The second, an identity rule, copies the initial value. The third uses a unit slope and adds the median training difference between paired relaxed and initial values. Because subtracting the initial value gives a one-dimensional location problem, the third rule is the best fixed unit-slope adjustment for the training sample under absolute loss. It is an informative check of the additive story: a favorable loss function is used to fit its correction, while the required one-for-one relationship remains intact. The link between absolute loss and the median is a special case of the point forecasting and quantile regression framework (Gneiting, 2011; Koenker & Bassett, 1978).

For the fourth and fifth rules, both intercept and slope were free to change. Ordinary least squares fits the intercept and slope by minimizing squared training residuals (Draper & Smith, 1998). The Theil–Sen fit uses the median slope between pairs with unequal initial energies (Sen, 1968), with the intercept estimated as the median of the training relaxed values minus the fitted slope times the corresponding initial values. The two affine fits allow the observed output scale to vary relative to its input. Their contrast is also informative because fitting squared residuals and fitting slopes by pairwise medians can yield different responses to influential departures. Other classical robust methods modify the location loss or use the scale of squared errors (Huber, 1964; Rousseeuw, 1984). The current comparison keeps that question concrete by evaluating just one robust and one ordinary affine implementation. Validation values play no role in estimating any parameter; the rules remain frozen throughout scoring.

Mean absolute error is the primary descriptive measure. Root mean squared error is also reported because larger errors receive disproportionate weight under squaring (Chai & Draxler, 2014). Table 1 gives the point differences in MAE relative to the constant, with negative values favoring the alternative, and paired

95% percentile intervals from 5,000 resamples of the 20 validation records. Each resample uses the same held-out index draws for both rules before their absolute losses are subtracted and averaged. This follows the bootstrap logic of estimating variability by resampling the empirical data (Efron, 1979; Davison & Hinkley, 1997). The intervals are conditional on the fitted parameters and the sampled holdout units. System-level clustering beyond common IDs is not resolved; a cluster-level bootstrap would require a defensible cluster definition and sampling scheme (Field & Welsh, 2007).

4. Results

The median initial energy was 3.93 eV in training and 3.78 eV in validation. The corresponding median relaxed energies were −1.54 eV and −0.74 eV, respectively. Relaxed energy was smaller than initial energy in every paired observation in both partitions. Yet the displacement was far from constant. The training median displacement was −5.81 eV, with an interquartile range of −8.85 to −3.20 eV. The validation median was −4.95 eV, with an interquartile range of −7.64 to −2.99 eV. The initial energy extended to 24.84 eV in training and 28.79 eV in validation whereas the corresponding upper bounds of the relaxed energy were 3.39 and 3.65 eV. These marginal summaries show why a comparison of methods that preserve versus alter the input scale is useful, while the paired fits quantify how that decision matters for prediction.

Table 1. Held-out errors for predicting relaxed energy from the archived initial energy.

Rule	Fitted slope	MAE (eV)	RMSE (eV)	MAE difference versus constant (95% interval), eV
Constant training median	0.000	2.415	2.770	Reference
Raw initial energy	1.000	6.949	9.974	+4.535 [2.093, 7.974]
Median displacement	1.000	4.100	7.245	+1.685 [−0.316, 4.644]
Ordinary least squares	0.111	2.279	2.786	−0.135 [−0.418, 0.245]
Theil–Sen	0.116	2.296	2.789	−0.119 [−0.420, 0.264]

One score per system, with 20 held-out systems. Intervals resample the validation systems with the source-trained rules held fixed. Displayed differences use unrounded scores.

The raw input had the largest error, with each observed relaxed value below its raw predictor. After applying the median training displacement, error fell considerably relative to copying the input. Nevertheless, the additive rule still performed worse on average than the constant. Its median absolute error was 2.70 eV, below its own 4.10 eV mean but slightly above the constant's 2.41 eV median absolute error. The fitted slope appears to be more consequential than adjusting the calibration intercept alone. The least-squares and robust slopes were 0.111 and 0.116 eV per eV. Put another way, a 10 eV difference between initial energies translates into roughly 1.1 eV between their affine predictions. The different fitting methods reached similar amounts of compression using the same training records.

Fitting the scale did not settle whether the initial scalar consistently improves prediction. The affine rules were only about 0.12–0.14 eV better than the constant in held-out MAE, and both paired intervals spanned zero. RMSE was slightly lower for the training median. To locate the large errors that could account for that contrast, a descriptive check separated the five highest initial energy records from the other fifteen. The MAE of the additive rule was 8.54 eV in the high-input portion and 2.62 eV in the remainder. The corresponding values were 3.18 and 2.16 eV for the constant, and 3.16 and 1.98 eV for

least squares. The high-input group was defined by the covariate, which makes it a diagnostic of the numerical rule under high (x), with only five cases and no independent inference about that subgroup.

5. Discussion

The results expose two distinct requirements for a useful correction: a shift of location and a suitable scale. Typical relaxed values were lower and a median adjustment reduced much of the raw error. Leaving the initial scale unchanged still made large positive initial energies yield high estimates of relaxed energy. The near agreement of the freely fitted slopes makes it harder to explain the fitted compression solely as a consequence of choosing a squared loss. Both affine models, however, learned from the same archive and showed only a small empirical improvement under absolute loss. The result primarily supports including and reporting an inexpensive constant benchmark when examining the extra information supplied by a calculated scalar. A more elaborate predictor may lower training error with uncertain benefit at evaluation, a general concern in discussions of model complexity and estimation variance (Geman et al., 1992).

There are three evaluation concerns for a larger follow-up. First, software and papers should state explicitly when the initial calculated energy itself would be available when predictions are made. Obtaining that value may consume computational resources and alters the information given to the model, so a fair measurement would account for both cost and the corresponding input boundary. Second, subgroup distributions may warrant broader evaluation. A split that breaks documented chemical or structural associations, if provided by the complete benchmark, would support a stronger evaluation of generalization. Work on extrapolation in materials discovery and structure-based out-of-distribution benchmarks shows how a change in the separation of training and evaluation examples can materially alter the question answered by a score (Meredig et al., 2018; Omee et al., 2024). The success of a simple scalar in one partition would be evidence about that partition and the fit available to it.

Third, repeated choices about how to handle extreme values, estimate slopes, or compare metrics can absorb information from benchmark evaluations. Established materials benchmarks define evaluation protocols for direct comparisons (Dunn et al., 2020). When the research involves choices based on previous performance, work on selection bias recommends a separate stage for evaluating the resulting choice (Cawley & Talbot, 2010; Varma & Simon, 2006). Here the quantitative intervals summarize the fixed-rule comparison on one small slice; dependable claims about whether more flexible numerical predictions improve in a specified domain would need more independent evidence. Calibration work in computational chemistry also emphasizes that uncertainty about errors requires comparison with the population or setting to which the resulting claims apply (Pernot, 2022).

6. Conclusion

The small numerical range of the affine predictions also matters for interpreting individual cases. Compression of the input scale reduces the effect of an unusually high initial value, but it has no way to distinguish two systems with similar initial values and different relaxed energies. In a setting where the extreme relaxed values are substantively important, this limitation could persist even if an aggregate fit slightly improved. A comparison using the complete benchmark could report how prediction spread and errors vary across prespecified input or structural groups. The tutorial provides a tightly bounded illustration of this issue because the observed predictions can be traced directly to the estimated slopes.

In the OC20 tutorial, an initial-energy value could be converted into a usable numerical prediction by substantially contracting its scale, but the contraction improved average absolute error by little relative to the constant. A fixed additive correction performed worse despite a uniformly negative measured

displacement. Any evaluation of the initial energy's incremental predictive value should specify whether that calculation is available in the intended setting and use a holdout with enough independently identified systems to measure the benefit.

Data and Code Availability

The public tutorial archive was retrieved on 23 November 2026 (SHA-256 81fe6511f87d435f463be34eccae9ac1dbe1e3e86789d4904966cb70e5a3077a). The upstream OC20 documentation specifies CC BY 4.0. The accompanying analysis script, figure script, and aggregate result files identify the two archive members and state the fixed procedures for scoring and resampling.

References

Balasingham, J.; Zamaraev, V.; Kurlin, V. (2024). Material Property Prediction Using Graphs Based on Generically Complete Isometry Invariants. *Integrating Materials and Manufacturing Innovation*, 13(2), 555–568. https://doi.org/10.1007/s40192-024-00351-9

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

Chai, T.; Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chanussot, L.; Das, A.; Goyal, S.; Lavril, T.; Shuaibi, M.; Riviere, M.; Tran, K.; Heras-Domingo, J.; Ho, C.; Hu, W.; et al. (2021). The Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis*, 11(10), 6059–6072. https://doi.org/10.1021/acscatal.0c04525

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

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

Draper, N. R.; Smith, H. (1998). Applied Regression Analysis. *Wiley*. https://doi.org/10.1002/9781118625590

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

Field, C. A.; Welsh, A. H. (2007). Bootstrapping Clustered Data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(3), 369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

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

Geman, S.; Bienenstock, E.; Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1), 1–58. https://doi.org/10.1162/neco.1992.4.1.1

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

Hammer, B.; Hansen, L. B.; Nørskov, J. K. (1999). Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals. *Physical Review B*, 59(11), 7413–7421. https://doi.org/10.1103/physrevb.59.7413

Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. (2020). DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247, 106949. https://doi.org/10.1016/j.cpc.2019.106949

Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. (2017). The atomic simulation environment—a Python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27), 273002. https://doi.org/10.1088/1361-648x/aa680e

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

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1), 73–101. https://doi.org/10.1214/aoms/1177703732

Klicpera, J.; Groß, J.; Günnemann, S. (2020). Directional Message Passing for Molecular Graphs. *International Conference on Learning Representations*. https://openreview.net/forum?id=B1eWbxStPH

Koenker, R.; Bassett, G. (1978). Regression Quantiles. *Econometrica*, 46(1), 33. https://doi.org/10.2307/1913643

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

Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; Mehta, A.; Ward, L. (2018). Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering*, 3(5), 819–825. https://doi.org/10.1039/c8me00012c

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

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

Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105

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

Sen, P. K. (1968). Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association*, 63(324), 1379–1389. https://doi.org/10.1080/01621459.1968.10480934

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