What does a materials prediction target cost?

Distinct labels and error certificates on a common test population

Lluis Eriksson

19 November 2026

Abstract

A benchmark can report accuracy for fixed training sizes while leaving two operational questions open. How many distinct materials must be measured along a sequence of designs? Does the fitted predictor reach its error target on materials that remain unmeasured? We reanalyse two acquisition strategies in an open materials-regression benchmark. At four budgets, 89 of 150 paired property–seed runs reached a specified accuracy on the full library under both strategies. Counting distinct labels changes all 21 nominal ties into lower acquisition counts for the active strategy. We also bound subset error using two rounded published moments. This permits a comparison on the same never-acquired materials across budgets and strategies, without access to individual prediction errors. The resulting counts describe a retrospective grid of designs; operational stopping and monetary savings require further information.

1 Design and acquisition cost

Gallagher, Hasko and Webb evaluated training-set selection methods in chemical and materials prediction [1]. We use their public size experiment at release v1.0.0 [2]. We compare space-filling cluster selection (`sf-cluster`) against active acquisition (`al-sv-topk`) with the same final random-forest predictor. This comparison, the whole-library error target and pairing rule were fixed before reading paired outcomes. The subset analyses were developed subsequently. Results concern these two configurations.

For a fixed property and seed, let I_{aj} be the archived list of training row identifiers for arm a at budget j in $(100, 500, 1000, 5000)$. Let S_{aj} contain all distinct identifiers seen through j under a. The cost of visiting those budgets in sequence is $|S_{aj}|$ measurements, assuming earlier labels can be reused at zero additional measurement cost. Each predictor still trains on its original list. The replay does not retrain on the union. Of 1,200 lists, 105 contain a repeated identifier. Every one of the 900 successive pairs drops at least one earlier identifier from the training list. Jumping directly to one budget would be a different purchasing policy.

2 Which materials are scored?

We compare 15 database–property tasks with ten seeds apiece: band gap and formation energy in JARVIS18 and MP21, formation energy in OQMD14, five PolyOmics targets (Cp, Rg, density, tg and thermal conductivity), and five QM9 targets (cv, gap, r2, u0 and zpve). A matched seed is a pairing key. Repeated seeds on one library do not give independent experimental populations.

The target is a root mean squared error strictly below 1/2 of the full-library target standard deviation. This scale stays fixed over budgets and populations. Both arms use the same library of N rows for the first endpoint. On that library, cluster selection crosses first at a smaller nominal budget in 68 pairs, both cross at the same budget in 21, only the active arm crosses in one, and neither crosses in 60. Inclusion of equality at the displayed threshold changes none of these counts. Among the 89 joint successes, counting unique measurements yields 68 lower costs for cluster selection and 21 for active acquisition. At JARVIS18 band gap, seed 1, the two methods first cross at 5,000 rows yet use 6,200 and 5,781 distinct labels, respectively.

The second endpoint scores the complement of S_{aj}. It answers what remains to be predicted by that strategy at that budget. Different arms can leave different targets. A third endpoint uses the complement T of the union of *all eight* lists from that pair. Every predictor is thereby evaluated on precisely the same material identifiers, including when the methods cross at different budgets. The construction uses future selection information retrospectively. Neither design produces a random external holdout. The authors' evaluation code standardizes features across the known candidate library and calculates errors on that library; the subset results below concern the archived predictors.

3 Certifying error from rounded summaries

Let e_i be the residual divided by the original target standard deviation. On all N rows let $A = N^{-1}\sum_i |e_i|$ and $M = N^{-1}\sum_i e_i^2$. Let the subset of interest have fraction q and write $p = 1 - q$. Its contribution $s = N^{-1}\sum_{i\in T} e_i^2$ satisfies

$$A \le \sqrt{p(M - s)} + \sqrt{qs}.$$

This follows by applying Cauchy–Schwarz to each part. The subset mean squared error equals s/q. If $m \ge M$ and $m < q/4$, it follows immediately that $s/q < 1/4$. For $m \ge q/4$, the right-hand side with M replaced by m has its maximum at $s = qm$ and decreases on either side as one moves away from that point. At the boundary $s = q/4$ its value is $B = \sqrt{p(m - q/4)} + q/2$. If a lower bound on A exceeds B, the side of the boundary away from qm is excluded. For $m > 1/4$ this certifies failure; for $m < 1/4$ it certifies success. Remaining cases are left open.

The tabulated MAE a, RMSE r and relative RMSE z have five decimal places. With $h = 0.000005$ we take

$$m = (z + h)^2, \qquad A_{\text{low}} = \frac{\max(0, a - h)\max(0, z - h)}{r + h}.$$

The denominator is positive for all entries. Comparisons use exact rational arithmetic and square only positive sides. There is no independence or exchangeability assumption on errors. The certificate is sufficient. Its failure does not establish that two residual vectors producing opposite orderings exist.

4 First crossings and fixed test results

At each of four budgets the certificate returns success, failure or undetermined. For each arm a possible first crossing must survive all earlier certified outcomes. We keep each non-failing step up to and including the first certified success. If no success is certified, we also allow no crossing. We compare the cost of every remaining pair of candidates. A direction is certified only if both arms cross in all retained cases and every cost comparison gives that direction. This enumeration is conservative, even if some joint assignments are infeasible.

Scored population	Below target	Misses target	Undetermined
Whole library	382	818	0
Remaining under the current arm	364	369	467
Fixed materials unused by both arms	304	58	838

On the arm-specific remaining population, cluster selection has a certified lower first-crossing cost in 16 of 150 pairs. On the fixed common population this conclusion survives in one pair. The remaining 149 comparisons are unresolved by the certificates; loss of a certificate gives no evidence of a reversal. Across pairs, the common set retains between 73.0% and 95.9% of the library.

The surviving case is QM9 target `r2`, seed 8. The common test set contains 122,398 of 133,885 molecules. Cluster selection meets the target at its first budget, with 100 distinct labels. Active acquisition misses at 100 and meets it at the next budget, whose 500-row training list brings its cumulative cost to 561 distinct labels. Thus the verified difference on this grid is 461 measurements on identical evaluation rows. It is a single benchmark case. All 150 pairs were evaluated and retained.

A difference in labels can be translated into a monetary threshold only given a measurement price and selection, fitting and validation costs for both paths. If c is the common price per label and H_a denotes other costs accumulated by crossing, cluster selection is cheaper exactly when $H_c - H_a < c(|S_a| - |S_c|)$, with the two sets taken at their respective crossings. The archive supplies neither prices nor a live stopping signal. Kottke and colleagues explicitly optimize annotation and error costs in transductive classification [3]; here the endpoint is a fixed regression accuracy and the evidence consists of aggregate errors.

5 Reproduction, attribution and scope

The companion program `reconstruct.py` reads the published results and four index archives at the pinned commit, checks their Git object identifiers, reads the 15 database headers, checks index bounds, and reconstructs the three populations. It requires Python 3, NumPy and read access to the public archive. Run `python reconstruct.py results.zip` in a clean directory. The two output tables report aggregate counts and all matched curves, including cases that never demonstrably cross. The accompanying expected summary makes it possible to compare a rerun without relying on a figure or a model score. The materials data and predictors originate with the cited benchmark; this paper contributes a secondary analysis.

The fixed subset excludes materials selected by either arm over the whole grid. It removes changes in test-set composition *within* a paired comparison. Library coverage, outcome-dependent selection, access to candidate features, and the choice of two configurations still limit generalization. Neither the success count nor the unresolved count is an estimate of a population probability. A common target drawn prospectively and individual model predictions would enable direct comparisons for more pairs. The counts do not establish optimality of an acquisition strategy or evidence about other prediction tasks.

Model assessments and review context

OpenAI Codex participated in deriving and checking the certificates, implementing the reconstruction and preparing this manuscript. The frozen version and any subsequent assessment must be identified by the PDF digest. Assessment by a model involved in this work is an internal evaluation. This manuscript carries no claim of independent peer review.

References

[1] Q. M. Gallagher, S. Hasko and M. A. Webb. Minimizing data requirements for chemical and material property prediction via cluster-based training set selection. *Nature Computational Science* (2026). doi:10.1038/s43588-026-01076-9.

[2] Q. M. Gallagher, S. Hasko and M. A. Webb. Data and code for data-efficient regression, release v1.0.0. Zenodo 22133401. Pinned size experiment and scoring implementation.

[3] D. Kottke, C. Sandrock, G. Krempl and B. Sick. A stopping criterion for transductive active learning. *Machine Learning and Knowledge Discovery in Databases*, LNCS 13716. doi:10.1007/978-3-031-26412-2_29.