Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Experimental Context and Source Shift in Machine-Learning Prediction of Polymer Nanocomposite Glass-Transition Temperature

Arash Salehpour

arash.salehpour@tabrizu.ac.ir

University of Tabriz

Research Article

Keywords: Polymer informatics, Polymer nanocomposites, Glass-transition temperature, Machine learning, Source-aware validation

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11660961/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Literature-derived polymer nanocomposite datasets combine substantial heterogeneity in chemistry, composition, processing, characterization, and publication source, raising concerns about whether machine-learning performance obtained under conventional validation reflects genuine cross-source generalization. This study investigates whether recorded experimental context improves prediction of polymer nanocomposite glass-transition temperature (Tg) beyond within-corpus interpolation. A frozen NanoMine cohort comprising 332 records from 52 publication-level source groups was evaluated using nested feature tiers incorporating matrix and filler identity, composition, processing, and measurement context. Ridge regression and CatBoost were assessed under row-random, grouped source-held-out, and leave-one-source-group-out (LOSO) validation, with source-cluster bootstrap uncertainty. For CatBoost, adding processing context reduced row-random source-macro MAE from 33.01 to 22.12°C, corresponding to a 10.89°C improvement (95% CI 4.19–18.30), whereas the same improvement under grouped source-held-out validation was only 1.41°C (95% CI − 5.93–9.17). LOSO produced an intermediate 7.66°C improvement (95% CI − 0.68–16.96). Source-held-out validation also introduced substantially greater chemistry novelty, including 70.7% unseen matrix–filler pairs. Importantly, even when the exact matrix–filler pair was represented in training, source-held-out error remained 17.31°C higher than row-random error (95% CI 8.04–27.30). These findings demonstrate that experimental context can be highly predictive within familiar corpus structure without being equally transportable across sources, underscoring the need for source-aware validation, chemistry-overlap analysis, and explicit applicability-domain reporting in polymer informatics.

1. Introduction

Polymer informatics seeks to connect chemical structure, composition, processing, and measured response using data-driven methods, while confronting sparse, heterogeneous, and incompletely reported literature data [1–4]. Polymer nanocomposites are especially challenging because the observed property reflects coupled contributions from matrix chemistry, filler identity and loading, processing history, dispersion-related structure, and characterization protocol. NanoMine was developed to make such information machine-readable by combining curated polymer-nanocomposite records with a schema spanning materials, processing, characterization, and response [5–7].

Glass-transition temperature (Tg) is widely used as a polymer-informatics benchmark because it is both technologically important and sufficiently represented in polymer databases [8, 9]. Yet Tg is not determined by repeat-unit identity alone. Molecular weight can alter Tg systematically [10], thermal history and sample state matter, and DSC and DMA can yield method-dependent transition estimates [11, 12]. Recent molecular-descriptor and language-model approaches show that richer chemical representations can improve polymer-property learning [13, 14], but they also emphasize that representation and training-domain coverage shape extrapolation.

A directly relevant NanoMine analysis by Ma et al. [15] showed that composition and filler descriptors carry useful information for polymer-nanocomposite Tg and highlighted processing methodology as an important direction. The present study asks a different question. It quantifies the incremental predictive information carried by composition, processing, and measurement context and tests whether that information survives when entire source groups are excluded from training.

The working hypothesis is that experimental-context variables can be highly predictive within familiar chemistry and source-group structure yet lose transportability when their physical meaning is encountered in different material systems. A processing temperature, residence time, or mixing route is not chemistry-independent: its relevance depends on polymer thermodynamics, rheology, interfacial interactions, and the operating window of the matrix–filler system. Accordingly, a model can learn useful within-corpus context relationships without learning a transferable physical law.

Validation design therefore changes the scientific question. Row-random splitting can place records from the same source group, chemistry family, or experimental campaign in both training and test sets, whereas source-held-out validation simultaneously introduces source, chemistry, composition-range, processing, measurement, and target-distribution shift. Materials-informatics studies increasingly emphasize split design, aggregation heterogeneity, and applicability domains for precisely this reason [3, 16, 17].

We organize the evidence hierarchically. The primary confirmatory claim is that the apparent gain from recorded experimental context depends strongly on validation regime. A secondary diagnostic tests how chemistry novelty accompanies source-held-out evaluation and whether a residual source-held-out penalty remains when chemistry overlaps. An exploratory diagnostic examines source-level target-distribution shift. The analysis concerns an identity-based tabular representation of the present NanoMine cohort; it is not a claim about the limits of all possible polymer representations.

2. Experimental

2.1 Data source, retrieval, and cohort freeze

The analysis used NanoMine/MaterialsMine XML records obtained through an authenticated export workflow. The retrieval cutoff was 24 September 2026; the principal export was generated at 2026-09-24T07:34:16Z, when the XML gallery indexed 1,335 records. The principal export and two targeted retry passes yielded 1,329 successfully retrieved unique record titles (99.55%). Retrieval batches were reconciled by record title, raw XML SHA-256 hashes were retained for provenance, and no exact raw-XML duplicate occurred in the frozen 332-record cohort. Six records remained persistently unavailable after repeated retrieval attempts; no property content was imputed for unavailable records.

Figure 1 summarizes cohort construction. A valid target required exactly one GlassTransitionTemperature XML node containing a direct scalar numeric value with unit Celsius or Kelvin. Kelvin values were converted to Celsius as $Tg(°C) = Tg(K) − 273.15$; the original value and unit

were retained, and no target clipping, winsorization, or range-based exclusion was applied. Of the 1,329 retrieved XMLs, 484 contained Tg information. Forty records with multiple Tg nodes were excluded because individual values could not be assigned reliably to specific measurement conditions. Of the remaining 444 single-Tg records, eligibility additionally required a source identifier, matrix identity, filler identity, a FillerComposition or FillerComponentComposition context, and processing context (PROCESSING, MatrixProcessing, or FillerProcessing). A total of 112 records failed these context requirements, leaving 332 records.

Hereafter, a source group denotes all records sharing the same normalized publication-level source identifier. The identifier was taken from the XML DOI/source field, lowercased and trimmed, with common DOI URL prefixes (http://doi.org/, https://doi.org/, http://dx.doi.org/, https://dx.doi.org/) and a leading “doi:” removed; terminal periods were stripped. Non-DOI database source identifiers were retained as distinct groups rather than merged heuristically. The frozen cohort contained 52 source groups (mean 6.38 records, median 4.5, range 1–38). Tg ranged from − 157.15 to 364.38°C (median 100.7°C). Measurement context was DMA for 151 records, DSC for 74, DSC + DMA for 80, and not explicitly encoded for 27.

A cohort-selection audit was performed for the 112 excluded single-Tg records. They spanned 50 source groups and had a Tg median of 102.0°C versus 100.7°C in the included cohort, although their mean Tg was lower (76.60 versus 95.07°C). All had matrix identity, but only 41.1% had filler identity. The dominant exclusion patterns were missing filler identity/composition (60 records), missing filler composition alone (26), and missing processing context (20). Measurement-method proportions were broadly comparable, although DSC and DSC + DMA proportions differed appreciably (Table 1).

Table 1 summarizes the cohort-selection audit and the nested modeling tiers.

Table 1
Frozen cohort construction, exclusion audit, and nested feature tiers.

Item	N	Definition / note
Retrieved XML records	1,329	Successfully recovered from 1,335 indexed records
Tg-bearing records	484	At least one Tg node
Exactly one Tg node	444	Single scalar candidate before context filtering
Included confirmatory cohort	332 records / 52 source groups	Required source, matrix, filler, composition, and processing context
Excluded single-Tg cohort	112 records / 50 source groups	Mainly missing filler identity/composition or processing context
M0	—	Normalized matrix + filler identity
M1	—	M0 + mass fraction, volume fraction, basis, missingness indicators
M2	—	M1 + route; temperature count/min/max/mean; time count/min/max/sum; missingness indicators
M3-core	—	M2 + DSC/DMA record-level context + equipment-metadata presence

2.2 Nested feature tiers and representation scope

Four nested feature tiers were frozen before confirmatory evaluation. M0 contained matrix and filler identities. Identity normalization was deliberately conservative: strings were lowercased, leading/trailing whitespace was removed, internal whitespace was collapsed, and multi-component identities were retained as pipe-delimited combinations; no synonym or ontology-based merging was imposed. M1 added numeric mass-fraction and volume-fraction fields, composition basis, and explicit missingness indicators. Numeric composition values already encoded by NanoMine as dimensionless fractions were used on their native 0–1 scale; no mass-to-volume or volume-to-mass conversion was performed. Legacy text entries containing a percent sign were not silently rescaled and therefore remained missing in the numeric fraction field, with the corresponding missingness indicator preserving that state. Cohort eligibility required composition context to be present in the XML, not both numeric fraction values to be available.

M2 added processing-route category plus numerical summaries of scalar temperature and time fields encoded in matrix, filler, surface-treatment, and main processing contexts. Recognized temperature units were Celsius and Kelvin; Celsius values were retained and Kelvin values were converted as °C = K − 273.15. Recognized time units were seconds, minutes, hours, and days, converted to hours by division by 3600, division by 60, identity, and multiplication by 24, respectively. The numerical summaries were

temperature count, minimum, maximum, and mean in °C; time count, minimum, maximum, and sum in hours; and explicit missingness indicators for the temperature/time summaries. M3-core additionally used a record-level measurement-context category and a binary equipment-metadata-presence indicator. Measurement context was assigned from explicit Differential_Scanning_Calorimetry and Dynamic_Mechanical_Analysis XML tags: DSC, DMA, DSC + DMA when both tags were present, or not explicitly encoded when neither tag was present. Because the confirmatory cohort contains a single Tg target per record, the DSC + DMA label represents record-level measurement metadata rather than an automatic one-to-one assignment of Tg to one method. Instrument model/name was excluded from M3-core because of poor cross-source support. Source identifiers were used only for grouping and diagnostics and were never predictors. The chemistry representation is therefore categorical and identity-based rather than a molecular-distance representation [13, 14].

2.3 Models

Three model classes were evaluated. A global-median predictor served as a source-blind reference; for every outer split, its target median was calculated from the training partition only and then applied to the held-out partition. Ridge regression used one-hot encoding for categorical variables, median imputation and standardization for numerical variables, and $\alpha = 10$. All Ridge preprocessing operators were fitted independently within each training partition. OneHotEncoder used handle_unknown="ignore", so categories absent from training produced zeroed indicator blocks rather than test-set fitting or leakage. CatBoost was the primary nonlinear tabular model because it handles mixed categorical/numerical inputs and missing numerical values [18].

Model settings were prespecified and frozen before confirmatory evaluation and were not optimized against any outer-fold test performance. CatBoost used 200 iterations, depth 5, learning rate 0.05, L2 leaf regularization 5, and MAE loss. Five fixed random seeds (111111, 222222, 333333, 444444, 555555) were run and out-of-fold predictions were averaged. The executed environment used Python 3.13.15, CatBoost 1.2.8, scikit-learn 1.7.2, NumPy 2.1.3, and pandas 2.2.3.

2.4 Validation design and claim hierarchy

The primary evaluation used a frozen five-fold source-grouped split. Source groups were sorted by decreasing record count, with source identifier used as the deterministic tie-breaker, and then assigned sequentially to the currently smallest fold; ties in current fold load were resolved by the lowest fold number. Tg values were not used to construct the folds. Every record from a source group was assigned to one outer fold, giving fold sizes 66, 67, 67, 66, and 66. The archived frozen source-to-fold map is the authoritative assignment. The primary metric was source-macro MAE: MAE was computed separately within each held-out source group and then averaged with equal source weight. Equal weighting was chosen because the intended transfer unit is the source group and because source sizes range from 1 to 38 records; this prevents the largest sources from dominating the headline result. Pooled row-level MAE, RMSE, and R^2 were reported as complementary metrics. Equal source weighting also makes the macro estimate more sensitive to very small source groups, which is retained as a limitation.

A deterministic row-random five-fold split over the identical 332 records served only as a within-corpus interpolation diagnostic. The original assignment was generated with scikit-learn KFold(n_splits = 5, shuffle=True, random_state = 111111), then frozen and checksum-verified for all subsequent reproduction runs. True leave-one-source-group-out (LOSO) evaluation was additionally run for every CatBoost feature tier M0–M3-core.

For source-held-out predictions, matrix identity, filler identity, and the exact matrix–filler pair were classified as seen or unseen according to whether the corresponding normalized identity occurred in that fold's training set. Seen/unseen identity is an operational overlap definition, not a continuous measure of chemical distance.

2.5 Bootstrap uncertainty

For the primary validation-regime and feature-tier comparisons, uncertainty was estimated with 10,000 source-cluster bootstrap replicates (seed 20260924). Models were not refitted inside bootstrap replicates; instead, frozen out-of-fold errors were aggregated by source group and resampled at the source-group level. Tier contrasts were paired within source. For the source-held-out versus row-random penalty, the same resampled source indices were applied to both validation regimes before differencing. All reported 95% bootstrap intervals are two-sided percentile intervals defined by the 2.5th and 97.5th percentiles.

Chemistry-controlled and chemistry-stratified sensitivity analyses used 20,000 source-cluster bootstrap replicates. For chemistry-stratum point estimates, source-specific MAEs were computed within each stratum. For unseen-minus-seen contrasts, source groups were resampled jointly from the complete 52-source set and both stratum means were recomputed inside each replicate, preserving covariance when one source group contributed records to both strata. A source group can therefore contribute to both seen and unseen strata, and stratum source counts need not sum to 52. LOSO feature-tier uncertainty was estimated analogously by paired resampling of the 52 source-level LOSO MAEs across tiers; 95% intervals again used the percentile definition.

2.6 Additional sensitivity analyses

Four additional sensitivity analyses were performed. First, global-median and Ridge performance were completed under row-random validation so that validation-regime dependence could be assessed beyond CatBoost. Second, a chemistry-controlled transfer analysis restricted evaluation to held-out records whose exact matrix–filler pair was represented in the corresponding training data, and compared source-held-out with row-random error on those same records.

Third, categorical context variables were audited for cross-source support. Levels appearing in fewer than three source groups were collapsed to a training-fold-specific rare category and M2 was rerun under both validation regimes. Fourth, the source-level target-mean-shift diagnostic was subjected to an influence analysis excluding the source that had both the largest absolute standardized target-mean shift and the maximum M3-core prediction error.

2.7 Reproducibility

The executable pipeline verifies SHA-256 checksums for the three raw exports, the frozen row manifest, source-group folds, and the row-random fold map before fitting any model. Every frozen cohort row is linked to its raw XML by record title and record-level SHA-256. The executable reproducibility notebook exactly reproduced all eight frozen CatBoost source-macro MAEs. The associated run manifest records the execution environment, seeds, input checksums, fold sizes, and analysis settings; the reproducibility package also records the baseline, LOSO, chemistry-controlled, rare-category, cohort-selection, and influence-sensitivity outputs.

3. Results and Discussion

3.1 Cohort-selection audit does not indicate a narrow method-only subset

The cohort-selection audit clarified what the 112 single-Tg exclusions represent. Their Tg median was close to that of the included cohort (102.0 versus 100.7°C), but the excluded distribution had a lower mean and lower first quartile (46.11 versus 64.81°C). The principal loss of eligibility arose from missing filler identity/composition and processing context rather than from one characterization method.

Measurement-method proportions were broadly comparable: DMA represented 45.5% of excluded and 45.5% of included records; DSC represented 30.4% versus 22.3%; DSC + DMA 17.0% versus 24.1%; and not-explicitly-encoded 7.1% versus 8.1%. The cohort is therefore selected toward records with richer primary contextual metadata, and external representativeness should be interpreted accordingly.

3.2 Validation-regime dependence is not specific to CatBoost

Table 2 reports performance for the global-median, Ridge, and CatBoost models under both validation regimes. The global-median reference changed little between row-random and source-held-out evaluation (55.35 versus 57.53°C source-macro MAE), as expected for a minimally structured predictor. Ridge showed a strong validation-regime dependence. Under row-random evaluation, M1→M2 improved source-macro MAE from 45.29 to 40.02°C, a 5.27°C gain. Under source-held-out evaluation, the same transition worsened MAE from 61.73 to 71.28°C.

CatBoost showed the same qualitative dependence but a different absolute response: processing context produced a large within-corpus gain and only a small source-held-out gain. The central conclusion is therefore not peculiar to one nonlinear learner.

Table 2
Global-median, Ridge, and CatBoost performance under row-random and source-held-out validation. All reported error metrics are in °C.

Model	Tier	Row-random source-macro MAE (°C)	Source-held-out source-macro MAE (°C)	Row-random pooled MAE (°C)	Source-held-out pooled MAE (°C)
Global median	—	55.35	57.53	49.56	53.24
Ridge	M0	45.15	60.21	37.34	55.02
Ridge	M1	45.29	61.73	37.57	58.40
Ridge	M2	40.02	71.28	35.03	73.21
Ridge	M3-core	37.27	72.88	32.10	75.57
CatBoost	M0	32.28	55.11	23.40	52.00
CatBoost	M1	33.01	54.45	24.25	49.94
CatBoost	M2	22.12	53.04	14.56	49.78
CatBoost	M3-core	22.86	52.15	14.49	49.68

3.3 Recorded processing context is highly predictive under row-random interpolation

Figure 2 and Table 2 show that row-random CatBoost M0 and M1 yielded source-macro MAEs of 32.28 and 33.01°C. Adding processing context reduced M2 error to 22.12°C. The paired source-bootstrap M1→M2 improvement was 10.89°C (95% CI 4.19–18.30°C; Table 3). M3-core did not further reduce source-macro MAE (22.86°C), although its pooled MAE was essentially unchanged relative to M2 (14.49 versus 14.56°C).

These results support a specific claim: recorded processing/context variables contain substantial predictive information when the train/test split permits familiarity with the literature and chemistry structure represented in the corpus.

3.4 Context gain attenuates under source-held-out evaluation

Source-held-out CatBoost error decreased only from 55.11°C at M0 to 54.45°C at M1, 53.04°C at M2, and 52.15°C at M3-core. The M1→M2 improvement was 1.41°C (95% CI − 5.93 to 9.17°C), the M2→M3-core

improvement was 0.89°C (95% CI − 1.92 to 3.74°C), and the overall M0→M3-core improvement was 2.95°C (95% CI − 5.07 to 11.21°C).

The M3-core source-held-out minus row-random source-macro penalty was 29.29°C (95% CI 20.03–39.53°C). We use “source-held-out transfer penalty” for this quantity because it is a bundled distribution shift: source group, chemistry, composition range, processing, measurement, reporting, and target distribution change together. Table 3 shows the incremental changes in source-macro MAE across feature tiers under row-random, source-held-out, and LOSO validation, together with their 95% source-cluster bootstrap confidence intervals.

Table 3
Incremental source-macro MAE contrasts across feature tiers. Positive ΔMAE favors the richer tier. Row-random and grouped five-fold intervals use the frozen out-of-fold (OOF) source-cluster bootstrap; LOSO intervals use paired resampling of the 52 source-level LOSO errors.

Validation	Contrast	Δ source-macro MAE (°C)	95% CI
Source-held-out	M0→M1	0.65	[− 3.27, 4.47]
Source-held-out	M1→M2	1.41	[− 5.93, 9.17]
Source-held-out	M2→M3-core	0.89	[− 1.92, 3.74]
Source-held-out	M0→M3-core	2.95	[− 5.07, 11.21]
Row-random	M0→M1	−0.73	[− 4.28, 2.99]
Row-random	M1→M2	10.89	[4.19, 18.30]
Row-random	M2→M3-core	−0.74	[− 2.63, 1.11]
Row-random	M0→M3-core	9.42	[3.45, 15.65]
LOSO	M0→M1	3.47	[− 2.79, 10.59]
LOSO	M1→M2	7.66	[− 0.68, 16.96]
LOSO	M2→M3-core	0.67	[− 3.92, 4.87]
LOSO	M0→M3-core	11.81	[1.45, 23.95]

3.5 Source-held-out validation induces substantially greater chemistry novelty

Figure 3 quantifies the shift in chemistry support. Under row-random validation, only 1.2% of held-out rows contained an unseen matrix, 1.8% an unseen filler, and 3.9% an unseen exact matrix–filler pair. Under source-held-out validation, the corresponding rates were 30.6%, 24.7%, and 70.7%. Thus the two validation regimes test materially different degrees of chemical extrapolation.

3.6 Source-aware chemistry strata reveal strong matrix-overlap stratification

Figure 4 and Table 4 report source-macro as well as pooled error and numerical 95% source-cluster bootstrap intervals. For matrix-seen records, source-held-out M3-core source-macro MAE was 34.04°C (95% CI 20.81–51.98; 230 rows, 36 source groups). For matrix-unseen records it was 86.32°C (95% CI 62.08–114.05; 102 rows, 19 source groups). Using a paired full-source cluster bootstrap that preserves covariance for source groups contributing to both strata, the unseen-minus-seen difference was 52.28°C (95% CI 22.19–83.45).

For exact matrix–filler pairs, source-macro MAE was 36.65°C when the pair was seen and 57.62°C when unseen. The paired-bootstrap unseen-minus-seen difference was 20.96°C (95% CI − 17.60 to 50.94), so the pair-level contrast remains descriptive. For the broader “both identities seen” criterion, the difference was 28.86°C (95% CI − 0.42 to 57.30), which also includes zero under the covariance-preserving bootstrap.

Table 4
Source-held-out M3-core chemistry-overlap strata with source-macro MAE, 95% source-cluster bootstrap intervals, and pooled MAE.

Dimension	Stratum	Rows	Source groups	Source-macro MAE (°C)	95% source-cluster CI (°C)	Pooled MAE (°C)
matrix	seen	230	36	34.04	[20.81, 51.98]	35.87
matrix	unseen	102	19	86.32	[62.08, 114.05]	80.82
pair	seen	97	17	36.65	[15.76, 71.51]	34.77
pair	unseen	235	37	57.62	[40.77, 76.43]	55.83
both identity	seen	190	30	36.70	[21.21, 57.98]	38.62
both identity	unseen	142	27	65.57	[45.02, 89.03]	64.47

Note: a source group can contribute records to both seen and unseen strata, so source counts across strata are not expected to sum to 52.

3.7 Chemistry-controlled transfer remains difficult even when the exact pair is represented

To separate chemistry novelty more directly from source-group transfer, we restricted analysis to the 97 held-out records (17 source groups) whose exact matrix–filler pair was represented in that fold's training data. On those identical records (Table 5), M3-core source-macro MAE was 36.65°C under source-held-

out prediction and 19.34°C under row-random prediction. The residual source-held-out penalty was 17.31°C (95% source-cluster CI 8.04–27.30°C).

The result does not isolate a causal source-group effect, because composition range, processing, measurement, reporting, and target distribution can still differ across source groups. It does show, however, that categorical matrix–filler identity novelty alone is insufficient to account for the full validation-regime gap. Table 5 shows the chemistry-controlled comparison between source-held-out and row-random M3-core performance for records with overlapping matrix–filler chemistry

Table 5
Chemistry-controlled comparison of source-held-out and row-random M3-core error on the same chemistry-overlap subsets.

Overlap subset	Rows	Source groups	Source-held-out MAE (°C)	Row-random MAE (°C)	Penalty (°C)	95% CI (°C)
pair seen	97	17	36.65	19.34	17.31	[8.04, 27.30]
both identity seen	190	30	36.70	17.02	19.69	[11.27, 28.91]
matrix seen	230	36	34.04	16.71	17.32	[9.68, 25.63]

3.8 Rare/source-exclusive context categories do not explain the row-random processing gain

The source-proxy audit found four composition-basis levels, five processing-route levels, four measurement-context levels, and two equipment-metadata-presence levels. Two composition-basis levels occurred in only one source group; one processing-route level occurred in only one source group; and the rare equipment-metadata-absence level occurred in one source group. Only 15 rows belonged to processing-route categories supported by fewer than three sources, and measurement-context categories had no source-exclusive levels.

After collapsing all categorical levels supported by fewer than three training source groups (Table 6), M2 row-random source-macro MAE was 22.96°C versus 22.12°C without rare-category collapse. Source-held-out M2 MAE was 56.85°C versus 53.04°C without rare-category collapse. The large row-random processing gain therefore cannot be explained solely by rare/source-exclusive categorical labels, although context variables can still encode more subtle within-source-group signatures. Table 6 summarizes the cross-source support of the categorical context features and the results of the rare-category-collapse sensitivity analysis

Table 6
Cross-source support of categorical context features used in M1–M3-core.

Feature	Levels	Source-exclusive levels	Levels < 3 sources	Rows in source-exclusive	Rows in < 3-source levels
Composition basis	4	2	2	17	17
Equipment metadata present	2	1	1	4	4
Measurement context	4	0	0	0	0
Processing route	5	1	2	4	15

Rare-category-collapsed M2 sensitivity: row-random source-macro MAE 22.96°C; source-held-out source-macro MAE 56.85°C.

3.9 LOSO reveals protocol-dependent intermediate context benefit

The expanded LOSO analysis produced source-macro MAEs of 62.12°C (M0), 58.64°C (M1), 50.98°C (M2), and 50.31°C (M3-core) (Table 7). The LOSO M1→M2 gain was 7.66°C (95% source-cluster CI − 0.68 to 16.96°C), and the M0→M3-core gain was 11.81°C (95% CI 1.45–23.95°C).

LOSO therefore does not simply reproduce the strong attenuation observed in the frozen grouped five-fold split. Instead, the M1→M2 point-estimate gain was 10.89°C under row-random interpolation, 1.41°C under grouped five-fold source-held-out validation, and 7.66°C under LOSO. A likely contributor is training support: each LOSO fit trains on 51 source groups, whereas the five grouped folds train on only 41–44 source groups. The LOSO result is therefore consistent with processing context retaining greater cross-source predictive value when broader source diversity is available during training, while absolute cross-source error remains high. Table 7 summarizes LOSO CatBoost performance across the four feature tiers, including source-macro MAE, pooled MAE, RMSE, R^2, and bootstrap confidence intervals

Table 7
Leave-one-source-group-out CatBoost performance across M0–M3-core with paired source-cluster bootstrap uncertainty for source-macro MAE.

Tier	LOSO source-macro MAE (°C)	Pooled MAE (°C)	RMSE (°C)	R^2	95% source-cluster CI (°C)
M0	62.12	58.25	85.19	−0.426	[44.90, 81.27]
M1	58.64	52.94	79.86	−0.253	[41.45, 78.27]
M2	50.98	47.02	68.06	0.090	[37.01, 67.56]
M3-core	50.31	45.92	67.49	0.105	[35.65, 67.61]

3.10 Target-mean shift is an exploratory retrospective diagnostic

Figure 5 defines the standardized held-out-source mean-Tg shift as |mean Tg of the held-out source − mean Tg of the corresponding training data| divided by the training-sample Tg standard deviation. Source-level M3-core MAE increased with this quantity (Spearman $\rho = 0.625$; exploratory $p = 7.15 \times 10^{-7}$). Excluding the source with both the largest standardized mean-Tg shift and maximum M3-core error left $\rho = 0.603$.

Because this target-mean-shift diagnostic uses the true Tg values of the held-out source, it is retrospective and cannot be used as a deployable out-of-distribution detector. It is presented only as evidence that source-level target-mean shift accompanies source-held-out difficulty.

3.11 Interpretation and limitations

The analyses sharpen the interpretation. Recorded processing context is useful, but its apparent value varies with validation regime and is consistent with sensitivity to the amount of source diversity available during training. The CatBoost M1→M2 point-estimate gain was 10.89°C under row-random interpolation, 1.41°C under the frozen grouped five-fold split, and 7.66°C under LOSO (95% CI − 0.68 to 16.96°C). Ridge independently showed split-dependent feature utility, and the row-random processing gain survived rare-category collapse. Source-held-out evaluation also introduced substantially greater categorical chemistry novelty, and unseen matrix identity was strongly associated with higher error. Yet an exact-pair-controlled analysis still left a positive source-held-out penalty.

The analysis does not establish a laboratory effect or identify the causal contribution of missing metadata. Source groups bundle chemistry, composition range, processing, measurement protocol, target distribution, and reporting behavior. Nor do these results establish an intrinsic limit of polymer ML: M0–M3 use categorical chemistry identities rather than molecular descriptors, repeat-unit fingerprints, physicochemical descriptors, or learned polymer embeddings [13, 14]. Unseen categorical identities are necessarily difficult for such a representation to extrapolate from.

Additional limitations are the exclusion of ambiguous multi-Tg records; sparse molecular-weight, particle-size, crystallinity, and microstructure fields; one frozen grouped five-fold partition for the confirmatory feature-tier contrasts; selection toward records with richer contextual metadata; and the greater variance that equal source weighting can assign to very small source groups. LOSO, cohort-selection, rare-category, chemistry-controlled, and influence analyses were used as complementary robustness checks. Bootstrap intervals condition on the frozen out-of-fold predictions and therefore quantify source-level sampling variability rather than full model-refitting or split-construction uncertainty.

4. Conclusions

In a frozen 332-record, 52-source-group NanoMine Tg cohort, the apparent predictive value of recorded experimental context depended on the transfer protocol. The CatBoost M1→M2 point-estimate gain was 10.89°C under row-random validation, 1.41°C under grouped five-fold source-held-out validation, and 7.66°C under LOSO (95% CI − 0.68 to 16.96°C). The intermediate LOSO result shows that cross-source context benefit is not a single fixed quantity and is consistent with greater context benefit when broader source diversity is available during training.

Source-held-out evaluation also introduced much greater categorical chemistry-identity novelty. Unseen matrix identity was associated with substantially higher source-aware error, but categorical matrix–filler identity novelty did not fully explain the gap: among records whose exact matrix–filler pair was represented in training, source-held-out error still exceeded row-random error by 17.31°C (95% CI 8.04–27.30). Rare-category and chemistry-controlled sensitivities support a cross-source transfer difficulty that is coupled to, but not reducible to, categorical chemistry novelty.

The practical message is methodological rather than deployment-oriented. For identity-based, literature-derived polymer-nanocomposite Tg modeling, experimental context can be strongly predictive within familiar corpus structure without being equally transportable across held-out source groups. Source-aware validation, explicit chemistry-overlap reporting, and representation-aware limitations should therefore accompany conventional row-random performance.

Declarations

Ethics approval and consent to participate

Not applicable.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Not applicable.

Author Contribution

Arash Salehpour: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Data curation; Visualization; Writing – original draft; Writing – review & editing; Project administration.

The author approved the final version of the manuscript and accepts responsibility for the integrity of the work.

Acknowledgments

The authors thank anonymous reviewers and respected editors for taking the time to review the manuscript. All authors listed have made a substantial, direct and intellectual contribution to the work and have approved it for publication. We also acknowledge the use of Grammarly for grammar and language checks, which improved the clarity and readability of the manuscript. The responsibility for the technical and scholarly content remains entirely with the authors.

Data Availability

The data analyzed in this study were obtained from the NanoMine/MaterialsMine database. The reproducibility materials include the analysis code, frozen cohort manifest, source-group assignments, validation fold maps, model seeds, checksum-verification procedures, sensitivity-analysis scripts, and derived result tables. Where redistribution of the original XML records is restricted by database terms, the accompanying repository will provide record identifiers, retrieval instructions, file-level SHA-256 checksums, and the frozen derived manifest required to reconstruct the study cohort. The code and derived reproducibility materials will be deposited in a permanent public repository before publication.

References

1. Audus DJ, de Pablo JJ. Polymer Informatics: Opportunities and Challenges. ACS Macro Lett. 2017;6(10):1078–82. https://doi.org/10.1021/acsmacrolett.7b00228.
2. Andraju N, Curtzwiler GW, Ji Y, Kozliak E, Ranganathan P. Machine-Learning-Based Predictions of Polymer and Postconsumer Recycled Polymer Properties: A Comprehensive Review. ACS Appl Mater Interfaces. 2022;14(38):42771–90. https://doi.org/10.1021/acsami.2c08301.
3. Ge W, De Silva R, Fan Y, Sisson SA, Stenzel MH. Machine Learning in Polymer Research. Adv Mater. 2025;37(11):2413695. https://doi.org/10.1002/adma.202413695.
4. Xu X-Y, Hu X, Wang L-Q, Jiang Y. Data-efficient Machine Learning for Polymer Informatics. Chin J Polym Sci. 2025;43(10):1707–17. https://doi.org/10.1007/s10118-025-3401-z.
5. Zhao H, Li X, Zhang Y, Schadler LS, Chen W, Brinson LC. Perspective: NanoMine: A Material Genome Approach for Polymer Nanocomposites Analysis and Design. APL Mater. 2016;4:053204. https://doi.org/10.1063/1.4943679.
6. Zhao H, Wang Y, Lin A, Hu B, Yan R, McCusker J, Chen W, McGuinness DL, Schadler LS, Brinson LC. APL Mater. 2018;6(11):111108. https://doi.org/10.1063/1.5046839. NanoMine Schema: An Extensible Data Representation for Polymer Nanocomposites.

7. Brinson LC, Deagen M, Chen W, McCusker J, McGuinness DL, Schadler LS, Palmeri M, Ghumman U, Lin A, Hu B. Polymer Nanocomposite Data: Curation, Frameworks, Access, and Potential for Discovery and Design. ACS Macro Lett. 2020;9(8):1086–94. https://doi.org/10.1021/acsmacrolett.0c00264.
8. Tao L, Varshney V, Li Y. Benchmarking Machine Learning Models for Polymer Informatics: An Example of Glass Transition Temperature. J Chem Inf Model. 2021;61(11):5395–413. https://doi.org/10.1021/acs.jcim.1c01031.
9. Miccio LA, Schwartz GA. Mapping Chemical Structure–Glass Transition Temperature Relationship through Artificial Intelligence. Macromolecules. 2021;54(4):1811–7. https://doi.org/10.1021/acs.macromol.0c02594.
10. Fox TG, Flory PJ. Second-Order Transition Temperatures and Related Properties of Polystyrene. I. Influence of Molecular Weight. J Appl Phys. 1950;21(6):581–91. https://doi.org/10.1063/1.1699711.
11. Rieger J. The Glass Transition Temperature Tg of Polymers—Comparison of the Values from Differential Thermal Analysis (DTA, DSC) and Dynamic Mechanical Measurements (Torsion Pendulum). Polym Test. 2001;20(2):199–204. https://doi.org/10.1016/S0142-9418(00)00023-4.
12. Gracia-Fernández CA, Gómez-Barreiro S, López-Beceiro J, Tarrío Saavedra J, Naya S, Artiaga R. Comparative Study of the Dynamic Glass Transition Temperature by DMA and TMDSC. Polym Test. 2010;29(8):1002–6. https://doi.org/10.1016/j.polymertesting.2010.09.005.
13. Kuenneth C, Ramprasad R. polyBERT: a chemical language model to enable fully machine-driven ultrafast polymer informatics. Nat Commun. 2023;14:4099. https://doi.org/10.1038/s41467-023-39868-6.
14. Zhou Y, Zhao Y, Li J, Miao C, Zhang H, Qi J, Wang Y, Zhou H, Hu Y. Molecular descriptor-driven machine learning for predicting polymer glass transition temperature. Polymer. 2026;346:129612. https://doi.org/10.1016/j.polymer.2026.129612.
15. Ma B, Finan NJ, Jany D, Deagen ME, Schadler LS, Brinson LC. Machine-Learning-Assisted Understanding of Polymer Nanocomposites Composition–Property Relationship: A Case Study of NanoMine Database. Macromolecules. 2023;56(11):3945–53. https://doi.org/10.1021/acs.macromol.2c02249.
16. Witman MD, Schindler P. MatFold: Systematic Insights into Materials Discovery Models' Performance through Standardized Cross-Validation Protocols. Digit Discovery. 2025;4:625–35. https://doi.org/10.1039/D4DD00250D.
17. Ottomano F, De Felice G, Gusev VV, Sparks TD. Not as Simple as We Thought: A Rigorous Examination of Data Aggregation in Materials Informatics. Digit Discovery. 2024;3:337–46. https://doi.org/10.1039/D3DD00207A.
18. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: Unbiased Boosting with Categorical Features. Adv Neural Inf Process Syst. 2018;31:6638–48.

Figures

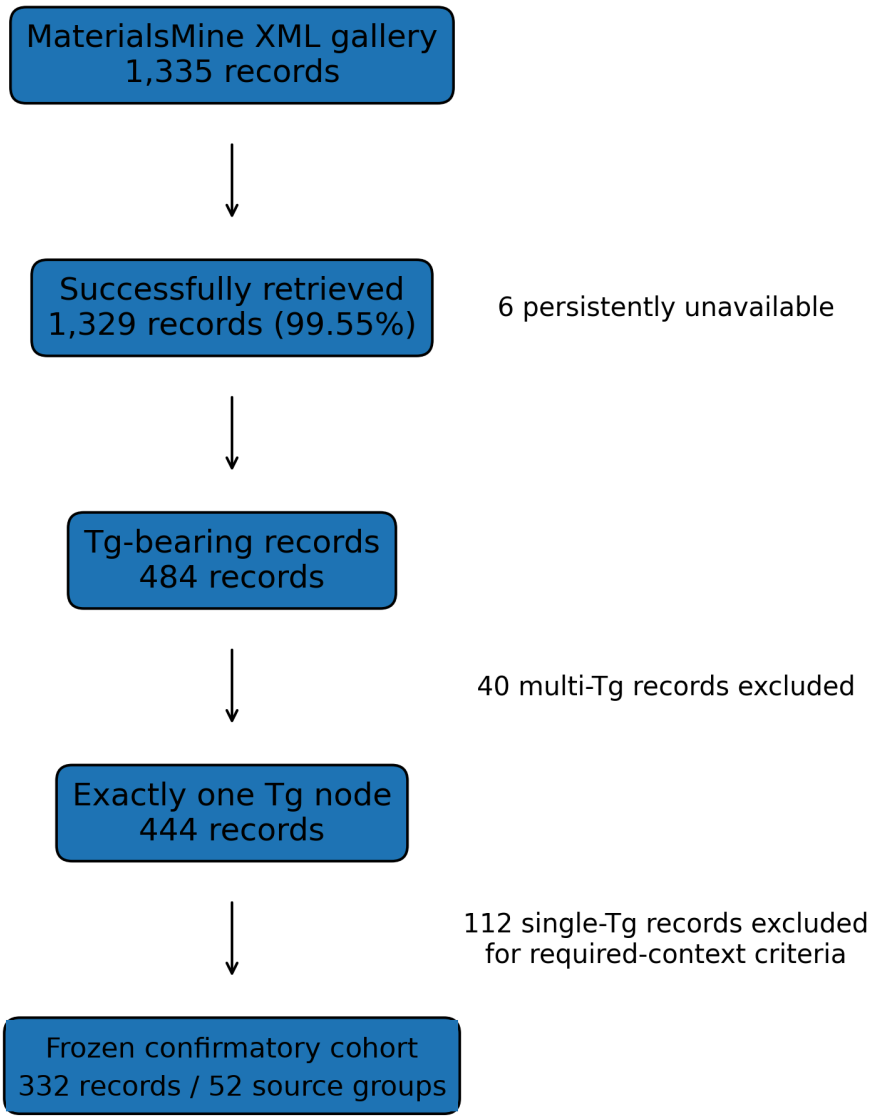

Figure 1

Cohort construction and confirmatory freeze. The workflow shows retrieval, Tg curation, conservative exclusion of multi-Tg/context-incomplete records, and the frozen 332-record/52-source-group cohort.

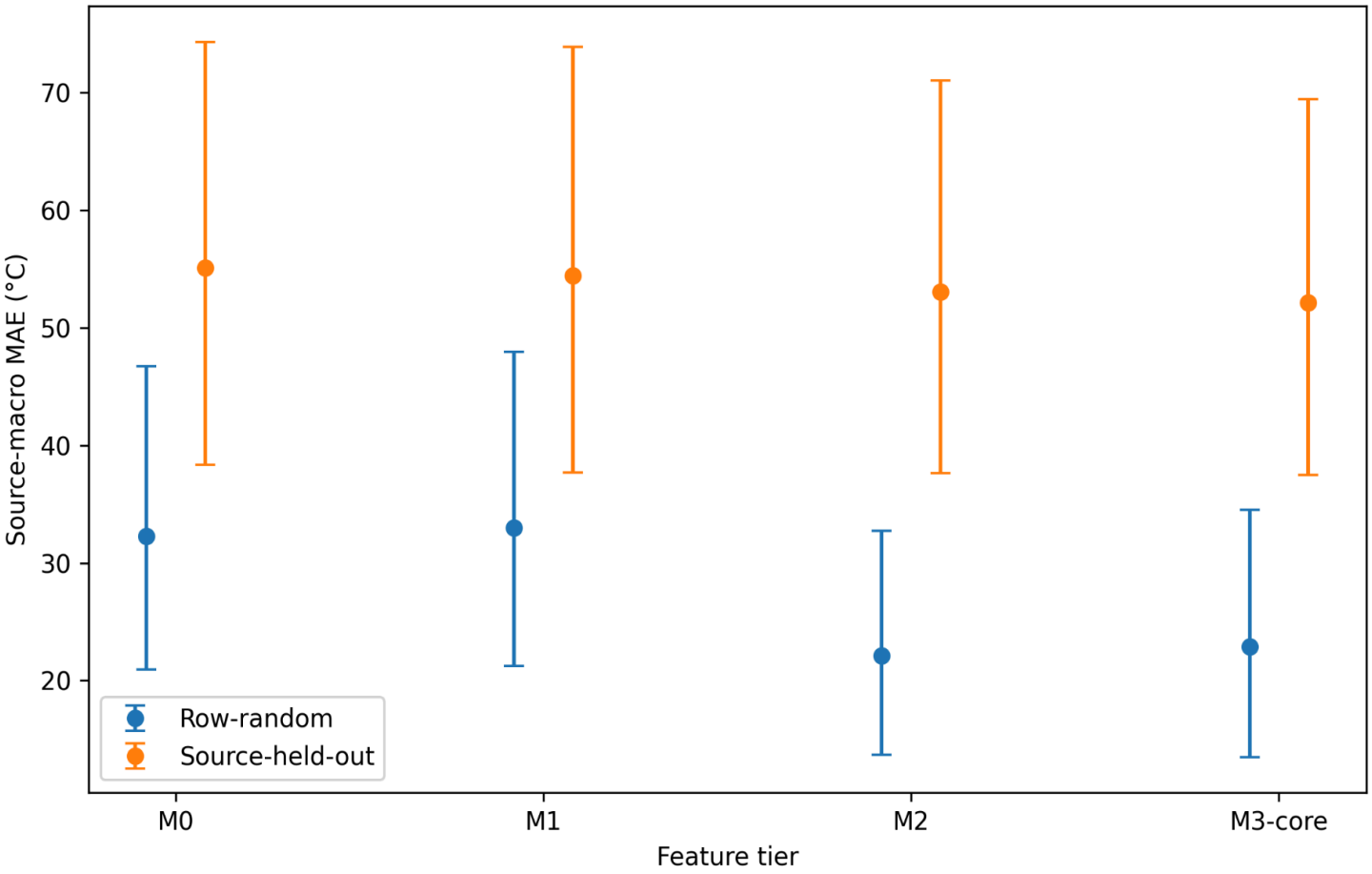

Figure 2

CatBoost source-macro MAE across nested feature tiers under row-random and source-held-out validation. Error bars are 95% source-cluster bootstrap intervals.

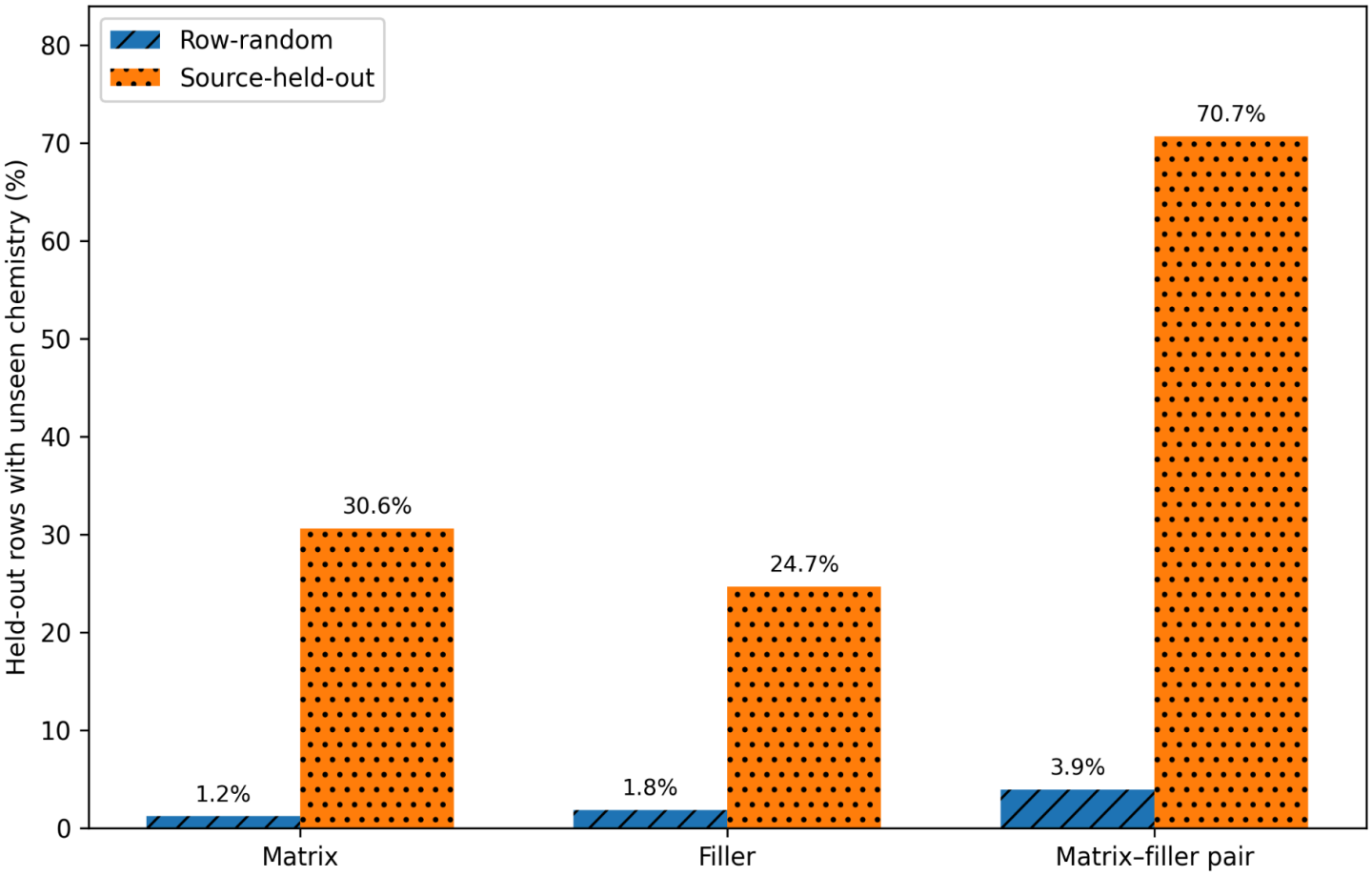

Figure 3

Operational chemistry novelty under the two validation regimes. Percentages are the mean held-out-row rates for unseen matrix identity, filler identity, and exact matrix–filler pair.

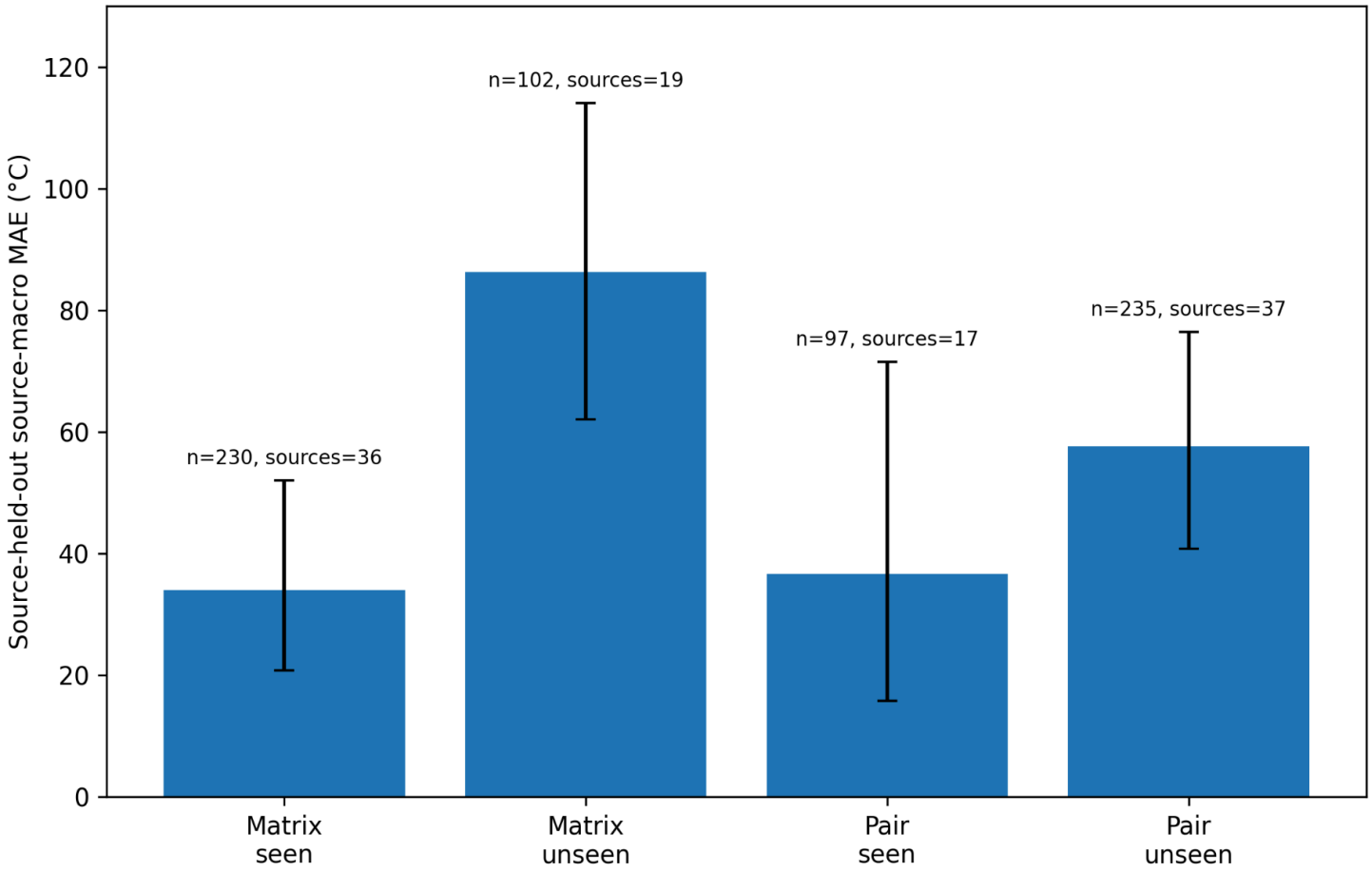

Figure 4

Source-held-out M3-core source-macro error stratified by matrix and exact matrix–filler-pair overlap. Error bars are 95% source-cluster bootstrap intervals; row and source-group counts are shown above each bar.

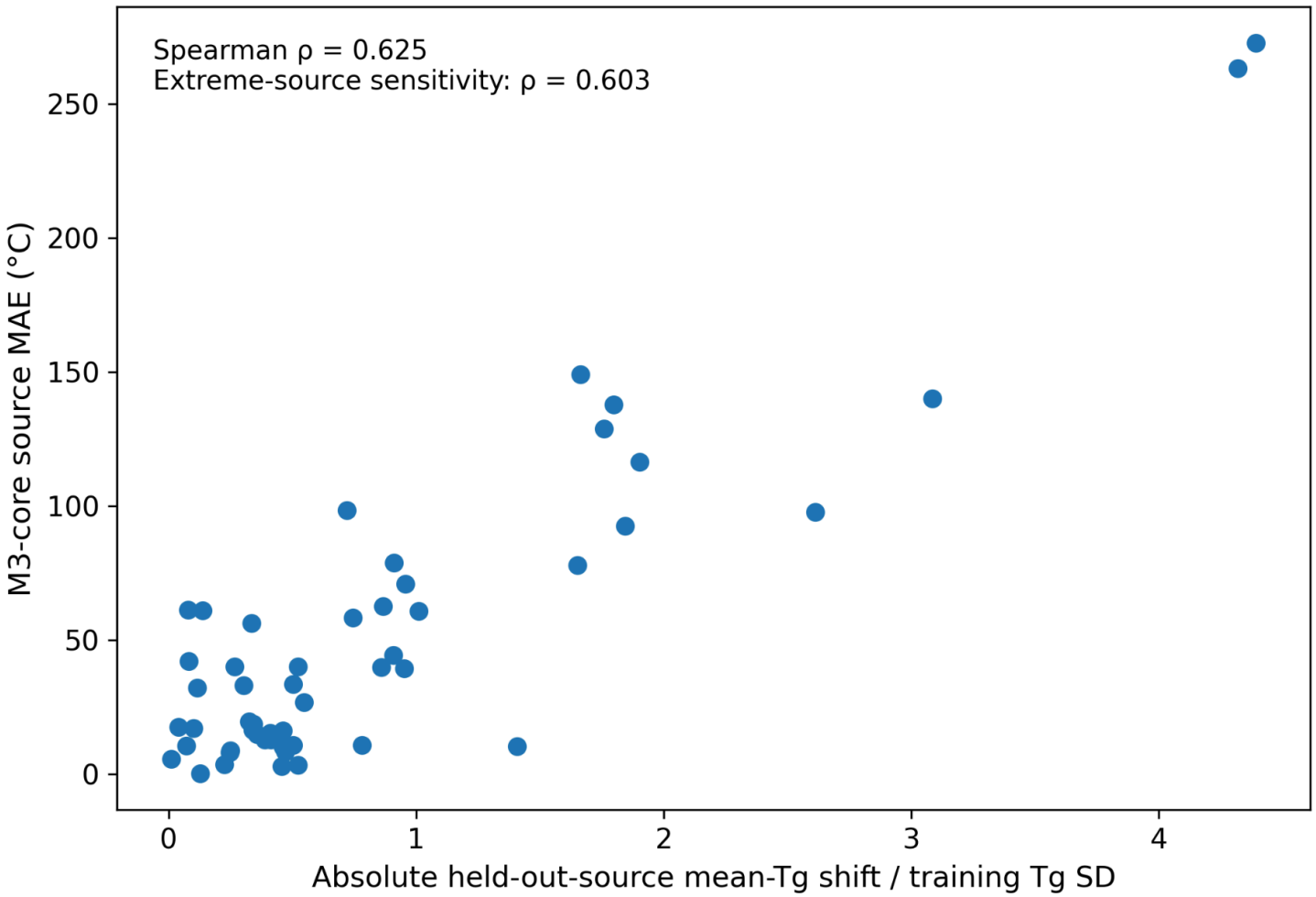

Figure 5

Exploratory retrospective target-mean-shift diagnostic. The x-axis is the absolute held-out-source mean-Tg shift divided by the training-sample Tg standard deviation. The relationship remains similar after excluding the extreme high-leverage source.