

Supplementary Information

Publication-date handling policies reshape temporal benchmarks for natural-product target retrieval

Chenyu Hu, Xiaoling Zhou, Yongkun Guo, Xinyou Zhang*

Additional file 1 · Supplementary Information

This supplement provides extended methods, validation, sensitivity analyses, and supplementary figures for the source-row-first reconstruction. Additional file 2 contains the corresponding machine-readable aggregate tables.

Scope and reading guide

This supplement describes a source-row-first framework for reconstructing temporal natural-product-target retrieval benchmarks. Publication-date handling policies are applied to one date-free strict source-row universe before relation membership is determined, and every downstream history, representation, mask, score, and evaluation object is regenerated. A three-condition analysis distinguishes exact-day processing, interval-certain processing within the day-only relation set, and full source-row-first interval-certain reconstruction. Here, universe means source rows passing the defined entity, mapping, sequence, verification, numeric-measurement, unit-conversion, and Tier A/B gates; it does not mean all NPASS records.

The primary analysis used interval-certain pre-cutoff history, Tier A and Tier B weights of 1.0, the symmetric endpoint, and exhaustive pair-neighbour ranking. The inclusive cutoff was 31 August 2022, after ChEMBL 31 [27] and before the recorded NPASS v2 update on 4 October 2022. Secondary analyses included a fixed 358-relation endpoint, the fixed-relation-key ablation, alternative evidence weights, and top-100 approximation. The study evaluates retrieval of later-recorded qualifying database relations; it does not identify first biological discoveries, establish direct binding, treat ChEMBL no-hits as negatives, or infer an assay hierarchy unsupported by the source fields.

The Supplementary Tables workbook is the machine-readable companion to this narrative. Its result and validation sheets cover the factorial analysis, date inputs and conflicts, molecular identity and duplicates, cold-scope transitions, fixed-relation-key attribution, paired robustness, top-100 tie stability, joint-scope positive-score support, and reproducibility checks. Restricted source rows, compound and target identifiers, relation keys, structures, sequences, endpoint membership lists, per-query values, and complete ranks are not reproduced.

Supplementary Methods

S1. Source-row-first design and estimands

The reconstruction proceeded through protocol, data processing, identity validation, scoring, evaluation, sensitivity analysis, and packaging stages. The analysis plan defined the cutoff, date-field priority, closed-interval semantics, strict evidence gates, primary configuration, target catalogue, baseline definitions, query-macro estimands, tie handling, uncertainty procedures, and software checks. The primary estimand was query-macro Recall@50 under exact uniform marginalization of score ties; Recall@10, NDCG@10, NDCG@50, and MRR were secondary. AUROC, ordinary AUPR, and calibration were excluded because unrecorded pairs were unlabeled rather than verified negatives.

Scoring and relevance evaluation were separated. The score stage used policy-specific v2 histories, role-specific molecular representations, and the fixed target catalogue but could not read endpoint relations or v3 activity fields. Its state receipts confirmed that endpoint data were inaccessible to the scoring stage. The evaluation stage loaded endpoint membership only after the score arrays and complete masked rank permutations had passed integrity checks. The analysis did not use reconstructed performance to select a baseline, weight, endpoint, threshold, scope, or representation.

The factorial design crossed day-only, interval-certain pre-cutoff, and interval-earliest-bound date policies with Tier B weights of 0.5, 0.7, and 1.0 while retaining Tier A weight at 1.0. These combinations produced nine declared policy cells. Both the symmetric endpoint and the fixed-358 comparator endpoint were evaluated, but the symmetric endpoint was primary. Interval-earliest-bound and interval-certain produced identical reconstructed histories in these data, so the nine declared policy cells corresponded to six distinct scoring states. The two interval policies remain logically different even when their observed states are numerically identical.

S2. Date-free v2 source-row universe

The v2 evidence input comprised 115,073 physical source rows. Each row was assigned one stable internal source-row identifier and one primary disposition before relation aggregation. A row entered the strict, date-free universe only if it had a complete InChIKey, a strict one-to-one reviewed-human UniProt mapping to the fixed human single-protein catalogue, an available canonical target sequence, the required automatic-verification status, an exact finite numeric measurement with units and finite molar conversion, and strict Tier A or Tier B evidence. Tier A was exact numeric molar Ki or Kd evidence; Tier B was exact numeric molar IC50 or EC50 evidence. AC50, generic Potency, censored or non-exact measurements, nonnumeric values, missing units or conversions, contextual-functional evidence, ambiguous mappings, and unavailable sequences did not satisfy the strict relation definition.

Applying these gates without reading a date retained 35,053 source rows, 13,573 relations, 4,685 compounds, and 1,559 targets. Day-only processing retained 13,885 rows, 4,990 relations, 1,726 compounds, and 1,131 targets. Interval-certain

processing restricted to the 4,990 day-only relation keys admitted 20,455 rows while retaining exactly the same relations, compounds, and targets. Unrestricted source-row-first interval reconstruction retained 34,803 rows, 13,544 relations, 4,680 compounds, and 1,557 targets. Earliest-bound and interval-certain histories were identical because no eligible v2 interval crossed the cutoff; this dataset therefore does not test their different treatment of a crossing interval.

Physical duplicates were retained rather than silently removed. Within the identity-validated day-only, fixed-relation-key, and interval-certain histories, 84, 122, and 212 exact full-row duplicate rows beyond first occurrence were available to the representative-selection rule. Repeating frequency-based structure selection after their removal changed the representative of zero compounds in all three states. Relation aggregation occurred only after row eligibility; a relation appeared once per configuration, Tier A took precedence over Tier B, and evidence weights were deterministic coefficients rather than calibrated activity probabilities.

S3. Publication-date reconstruction

PubMed ESummary metadata were read in the order epubdate, pubdate, and sortpubdate, taking the first parseable field. A day mapped to one calendar day, a month to its first and last valid days, and a year to 1 January and 31 December. No day was imputed for month- or year-precision records. Missing, invalid, seasonal, incompatible, non-PMID, and otherwise unresolved values retained null bounds. PubMed distinguishes publication dates from record-creation and entry dates [28]. The selected dates were treated as record-level valid-time proxies, not database transaction, ingestion, access, or discovery dates. The validation cross-tabulated selected field $\times$ original precision $\times$ publication era $\times$ inclusion status, compared parseable electronic and print intervals, and counted the minimum distance of selected intervals from the cutoff. The sortpubdate fallback was not selected in either PMID table.

For lower and upper date bounds L and U , and inclusive cutoff C , a row was historical-certain when $U \leq C$, definitely future when $L > C$, cutoff-crossing when $L \leq C < U$, and unresolved when a bound was unavailable. Day-only history required day precision and $L \leq C$. Interval-certain history required $U \leq C$. Interval-earliest-bound history required $L \leq C$. Relation-level aggregation was conservative: evidence for historical membership took precedence, followed by unresolved, cutoff-crossing, definitely future, and no eligible date. Thus, one retained historical source row prevented a relation from being classified as future.

Within the date-free v2 universe, exact-day history comprised 6,714 epubdate rows and 7,171 pubdate rows. Interval-certain processing additionally admitted 16,856 month-precision and 4,062 year-precision pubdate rows; 250 rows remained unresolved. These strata are row counts, while relation counts in the workbook are non-additive because a relation can have multiple supporting rows. Both electronic and print dates were parseable for 2,221 PMIDs supporting 6,714 strict rows and 2,189 non-additive relations. Electronic intervals were earlier and non-overlapping for 2,215 PMIDs, later for one, overlapping but unequal for four, and equal for one; a print-date-only alternative changed zero cutoff classifications. No selected strict v2 interval lay within 365 days of the cutoff, while 68 PMIDs/153 rows/96 non-additive relations lay 366–730 days away. The observed expansion is therefore visible by field and precision and was not driven by a near-cutoff electronic/print disagreement.

For illustration, a hypothetical strict row with publication year 2021 maps to [1 January 2021, 31 December 2021]. It remains unresolved under the day-only rule but enters interval-certain history because its upper bound is before the cutoff. Only after that decision is its compound–target relation aggregated; the compound can then contribute a historical structure, fingerprint, and scaffold, the target can contribute evidence, the query-specific mask is rebuilt, and an endpoint query can leave a cold scope. A hypothetical interval spanning 31 August 2022 would be excluded by interval-certain but admitted by earliest-bound. No retained study row had that crossing pattern.

At equal evidence weights, the fixed-relation-key ablation admitted 6,570 additional interval-certain rows while holding the 4,990 day-only relations fixed. No selected structure, mask, score, rank, operational denominator, cold-scope member, or Recall@50 estimate changed. Full interval-certain reconstruction instead added 8,554 relations, changed masks for 97 endpoint queries and 787 query–target cells, and rebuilt all dependent objects. The comparison separates source-row accrual within a fixed relation set from relation-membership expansion.

All 1,131 day-only historical targets were retained and 426 new targets entered interval history. Among the retained targets, equal-weight relation counts increased for 891, were unchanged for 240, and decreased for none; interval-minus-day changes had quartiles 1, 2, and 6, a maximum of 261, and a mean of 6.814. Eligible source-row support increased for 967 common targets and was unchanged for 164. The resulting fixed-salt popularity top 50 changed for 225/225 endpoint queries: 9,008 of 11,250 slots were retained and 2,242 each exited and entered. Of 364 endpoint relations, 346 changed popularity rank and 30 crossed top-50 membership. This is a broad redistribution of target evidence, not an isolated outlier addition.

S4. Symmetric future-endpoint construction

The same date-free entity, mapping, sequence, and measurement gates were applied to the v3 source before temporal classification. Of 169,813 v3 source rows, 41,498 passed the strict date-free gates. Relation-level classification was restricted to consistently v3-only relations. Under interval-certain semantics, 457 relations were definitely future before the ChEMBL 31 leakage screen. The exact overlap with reconstructed v2 history was zero.

The ChEMBL 31 analysis excluded 28 candidates with a historical activity record and set aside 65 candidates whose entities could not be resolved under the mapping rules. The unresolved group comprised 51 compound-unresolved and 14 target-unresolved relations. Among the ChEMBL-excluded relations, 11 carried Tier A evidence and 17 Tier B; the corresponding counts were 7/44 for compound-unresolved, 12/2 for target-unresolved, and 76/288 among the 364 retained relations. All strict candidate-disposition rows had day-precision publication records spanning 2022–2024. Target-family information was unavailable, so no family-stratified inclusion/exclusion claim is made. A ChEMBL no-hit was not treated as evidence of inactivity. No cutoff-crossing or historical-overlap relation remained in the endpoint flow. The symmetric endpoint therefore contained 364 relations, 225 query compounds, and 158 targets, supported by 495 strict source rows from 127 PMIDs. Every endpoint target occurred in the fixed catalogue of 4,123 reviewed human single-protein targets.

The fixed-358 comparator endpoint contained 358 relations, 222 queries, and 156 targets. All 358 comparator relations occurred in the symmetric endpoint, which added six qualifying relations and removed none; their relation-level Jaccard similarity was 0.983516. This secondary comparison evaluates robustness to a fixed endpoint without substituting it for the primary symmetric construction. Supplementary Figure S1 compares the endpoint scope counts.

S5. Scenario-local structures, masks, and cold scopes

Historical and query molecular representations were generated separately and validated at the complete-InChIKey level before scoring. Across 8,357 state–role representative records (1,726 day-only historical, 1,726 fixed-relation-key historical, 4,680 interval-certain historical, and 225 endpoint queries), 7,986 selected SMILES regenerated the declared full key directly. Exact eligible source InChI resolved 286 records, and a narrowly scoped declared-key-directed protonation normalization resolved 82 when the connectivity and stereochemical key blocks already matched and uncharging restored the complete key. Three state occurrences of one molecule used a documented RDKit /b-layer round-trip exception: the raw InChI regenerated the declared key and retained/direct forms passed canonical graph, default Morgan, primary/achiral scaffold, and tautomer-canonical graph/fingerprint invariance checks. No connectivity-block mismatch or unresolved representative was retained. Counts are state–role records and therefore repeat a molecule used in more than one historical state.

The representative rule remained most-frequent nonempty source SMILES with lexicographic tie resolution. Raw-InChI repair and the declared-key-directed protonation rule were the only structure-changing resolutions; any remaining full-key mismatch failed closed. Historical and query pools were never merged. No salt stripping, parent extraction, tautomer canonicalization for scoring, or general charge normalization was introduced; canonical tautomer comparison was used only as an exception-validation invariant. The exact-duplicate sensitivity described in S2 changed no selected representative.

UniProt mappings and canonical sequences obtained from resources available after the cutoff were used only for retrospective entity standardization and consistent feature construction. They must not be interpreted as information that was demonstrably available to a retrieval system at the cutoff date.

Same-query target masks, exact Bemis–Murcko scaffold sets, and sequence-homology masks were rebuilt for each policy state. A query was exact-scaffold-cold only when its Bemis–Murcko scaffold was absent from the scenario-specific historical compounds; this is an exact structure criterion, not a chemical-similarity threshold. Target homology was evaluated with the pinned MMseqs2 workflow over the fixed catalogue, using thresholds of 0.30, 0.50, and 0.70 and minimum bidirectional coverage of 0.80. These masks defined diagnostic exact-scaffold-cold and joint exact-scaffold/homology scopes. They were not treated as conventional prospective unseen-drug or unseen-target splits.

Under the primary interval-certain/equal-weight state and symmetric endpoint, the temporal scope contained 364 relations, 225 queries, and 158 targets. The primary stereochemistry-aware exact-scaffold-cold scope contained 61 relations, 48 queries, and 32 targets; all 48 queries/61 relations were retained from the day-only scope, 39 queries/63 relations exited, and none entered. The fixed-relation-key ablation retained the day-only scope of 124/87/71. Removing scaffold stereochemistry produced the same direction: 104/77/64 under day-only/fixed-relation-key histories versus 54/43/25 under interval-certain reconstruction. Joint exact-scaffold/homology scopes contained 9/7/5, 9/7/5, and 10/8/6 relations/queries/targets at identity thresholds 0.30, 0.50, and 0.70, compared with day-only 24/19/17, 29/22/21, and 29/22/21.

For the 48 primary scaffold-cold queries, maximum default radius-2/2,048-bit Morgan Tanimoto similarity to any identity-validated interval-history compound ranged from 0.198113 to 1.000000, with median 0.476177. Four queries had similarity exactly 1.0. Manual case audit showed distinct complete stereochemical identities with common connectivity and achiral scaffolds; because the RDKit Morgan calculation used its default chirality-disabled setting, their fingerprints were identical. These are representation-boundary cases rather than full-identity leakage. Exact primary scaffold absence therefore does not imply uniformly low fingerprint similarity or prospective chemical out-of-distribution status.

For the fixed-358 comparator under the primary state, the temporal scope contained 358 relations, 222 queries, and 156 targets, and the scaffold-cold scope contained 58 relations, 47 queries, and 29 targets. Its three joint-scope denominators matched the symmetric primary endpoint. These counts are supplied as a secondary comparator only.

S6. Retrieval baselines and complete-ranking audit

Four predefined retrieval baselines were evaluated. Target popularity scored candidates by their policy-specific historical evidence mass. Sequence transfer propagated target evidence through the precomputed target-sequence similarity relation. Morgan transfer used molecular fingerprint similarity between a query and historical compounds. Pair-neighbour transfer combined chemical proximity with target-side historical evidence and was evaluated exhaustively over all eligible historical targets in the primary analysis. A top-100 historical-neighbour restriction was retained only as an approximation sensitivity.

For each state and Tier B weight, the scorer wrote complete masked target permutations rather than endpoint-only ranks. Six unique state-weight score sets were required because interval-certain and earliest-bound shared a verified state. Every score set stored five complete candidate arrays: target popularity, sequence 3-mer, Morgan transfer, exhaustive pair-neighbour, and the separately labelled top-100 pair-neighbour sensitivity. Thus, a receipt that totals rank cells counts the top-100 array as another stored ranking, not as additional unique query-target candidates. Each array covered 225 queries against the 4,123-target catalogue, and every allowed portion was checked as a duplicate-free complete permutation after masking. Endpoint-hidden receipts passed before relevance evaluation.

Operationality was baseline-specific. Popularity, Morgan transfer, and exhaustive pair-neighbour transfer were operational for all 225 primary temporal queries. Sequence transfer was operational only when at least one allowed target received a nonzero sequence-transfer score; 128 primary temporal queries met that condition. All-zero sequence vectors remained in total-denominator accounting but did not enter the signal-bearing sequence macro mean. When a scope contained no operational sequence query, the value was stored as not estimable because no operational queries were available, not as zero, NaN, or infinity.

S7. Exact ties and descriptive uncertainty

Retrieval measures were computed per query and then equally averaged over operational queries. Uniform within-block averaging followed tie-aware retrieval evaluation [33]; exact score ties were not resolved solely by a fixed pseudo-random ordering. For a tie block spanning ranks a through b, the evaluator calculated the exact expected contribution under a uniform within-block permutation and exact lower and upper values at each cutoff. Fixed-salt values, uniform-tie expectations, and exact bounds were retained as different estimands.

The primary temporal uniform-tie Recall@50 values were 0.177598 for target popularity, 0.058464 for sequence transfer over 128 operational queries, 0.294806 for Morgan transfer, and 0.296963 for exhaustive pair-neighbour transfer. Their exact tie intervals were 0.176857–0.178339, 0.058464–0.058464, 0.280667–0.321302, and 0.295852–0.298074. The wider Morgan interval shows why an expected value and a particular fixed ordering must not be conflated.

Two descriptive resampling procedures used 10,000 PCG64 replicates with deterministic labelled substreams. The query bootstrap resampled operational queries and used percentile intervals [34]. Following the cluster-bootstrap principle [35], the PMID-component bootstrap resampled connected components in the query-PMID graph and aggregated only operational members. For primary temporal Recall@50, query-bootstrap intervals were 0.132634–0.226784 for popularity, 0.027214–0.095443 for sequence, 0.239662–0.352920 for Morgan, and 0.241407–0.354444 for pair-neighbour. Component intervals were 0.091943–0.310187, 0.023374–0.123459, 0.192242–0.431623, and 0.197291–0.430225. The all-query baselines used 97 effective components covering 225 queries; sequence used 64 covering 128 operational queries. The graph defines resampling groups but does not establish their independence.

The tie and uncertainty output contained 1,800 aggregate rows. Of these, 360 rows had zero operational queries and carried the explicit non-estimability status with blank numerical fields. No serialized NaN or infinite sentinel was permitted in the aggregate package.

S8. Factorial policy contrasts

All nine date-policy $\times$ Tier-B-weight configurations were evaluated using identity-validated structures. Interval-certain and earliest-bound cells remained separately labelled despite identical inputs. The aggregate factorial table contains every metric, tie estimand, scope, endpoint family, and operational denominator; effect decomposition is restricted to matched endpoint, scope, baseline, metric, and estimand combinations. The fixed-relation-key ablation is evaluated at weight 1.0 and is not substituted for any full factorial cell.

For temporal Recall@50 under the symmetric endpoint, interval-certain minus day-only at Tier B weight 1.0 was -0.047566 for popularity, $+0.012015$ for sequence, $+0.015657$ for Morgan, and $+0.023312$ for exhaustive pair-neighbour. Within interval history, weight 1.0 minus 0.5 was $+0.010190$, -0.008594 , $+0.041265$, and $+0.051778$, respectively. The pair-neighbour date-by-weight interaction against 0.5 was -0.021979 ; against 0.7, its weight contrast was $+0.011778$ and interaction -0.011270 . In the unmatched scaffold-cold scope, the pair-neighbour total date contrast was $+0.034124$ and the weight 1.0-minus-0.5 contrast $+0.170139$. These full-population quantities mix membership and derived-state changes and must be interpreted with the matched analyses below.

Each full date-policy contrast changed historical membership and the structures, masks, evidence masses, operability, scopes, and scores derived from it. The fixed-relation-key ablation added 6,570 interval-certain rows within the day-only relation set but reproduced every day-only mask and primary metric exactly. For the fully operational temporal baselines over all 225 queries, interval-minus-day Recall@50 was -0.047566 for popularity, $+0.015657$ for Morgan transfer, and $+0.023312$ for exhaustive pair-neighbour transfer. The unmatched sequence contrast was $+0.012015$ while operability changed from 61 to 128 queries, so it is not a paired query effect. A representation-only date effect remains not separately identified, and the factorial results are descriptive policy sensitivities rather than multiplicity-adjusted tests.

Matched analyses used the 48 queries cold in both histories and the 61 queries sequence-operational in both. For pair-neighbour Recall@50, the common-cold interval-minus-day estimate was -0.006944 ; paired query-bootstrap and PMID-component 95% intervals were -0.072917 to 0.062500 and -0.085714 to 0.056826 . The 26 PMID components had a largest component of 12 queries (25%); deleting one component shifted the estimate by at most 0.024063 . The unmatched interval-cold minus day-cold difference was $+0.034124$ and decomposed exactly into matched ranking state -0.006944 plus day-state population composition $+0.041068$. This arithmetic identity is descriptive and non-causal.

For sequence Recall@50 on the 61 common operational queries, the paired estimate was $+0.017486$, with query-bootstrap interval 0 to 0.051366 and PMID-component interval 0 to 0.069767 . The 39 components had a largest component of 20 queries (32.8%); the largest leave-one-component-out shift was 0.016375 . Over all 225 temporal queries, pair-neighbour paired change was $+0.023312$, with query and component intervals of -0.016985 to 0.065409 and -0.016452 to 0.082067 . These dependence checks cross zero where noted and do not establish method superiority.

S9. Analytic null quantities and zero-hit scopes

The null audit used an analytic finite-catalogue random-ordering reference. For each query, it recorded the number of allowed targets, the number of relevant endpoint targets, the cutoff, expected Recall under a uniform random permutation, and the hypergeometric probability of at least one top-K hit. It generated 46,032 query-level null rows. Primary observed-to-null enrichment divided the per-query uniform-tie Recall expectation by its nonzero query-specific random-ordering expectation, matching the primary retrieval estimand; the fixed-salt numerator was retained as a separately labelled sensitivity. This is a catalogue-conditioned reference rather than an empirical negative control.

At $K = 50$ in the primary temporal scope, mean uniform-tie-aligned observed-to-null enrichment was 14.631872 for popularity over 225 queries, 4.753870 for sequence transfer over 128 operational queries, 24.261898 for Morgan transfer over 225 queries, and 24.439468 for exhaustive pair-neighbour over 225 queries. Corresponding fixed-salt sensitivities were 14.692954 , 4.753870 , 24.927883 , and 24.347890 . These are catalogue-conditioned random-ordering references rather than empirical negative controls, calibrated activity probabilities, or biological mechanism validation.

In each primary joint scaffold/homology scope, all estimable baselines had Recall@50 equal to zero, but their score support differed. Relevant positive-score coverage for both popularity and Morgan transfer was 0/9, 0/9, and 0/10 at identity thresholds 0.30, 0.50, and 0.70; popularity assigned positive scores to 1,557 eligible candidates per query and Morgan to 1,542–1,557, so their zero relevant support was a structural consequence of requiring direct historical target evidence. Pair-neighbour transfer assigned a positive score to every relevant relation (9/9, 9/9, and 10/10) and to all 4,123 catalogue candidates, but ranked no relevant relation in the top 50. Sequence transfer had all-zero candidate vectors and was non-estimable, not a zero-performing operational method. One-sided 95% exact binomial upper limits [36] for a query-level

any-hit probability were 0.348164 for the two seven-query scopes and 0.312344 for the eight-query scope. These small-sample limits do not adjust for documentary dependence or establish irretrievability.

S10. Potency and measurement-type sensitivity

Endpoint-side strict evidence was summarized without a potency threshold, at $p\text{Activity} \geq 5$, and at $p\text{Activity} \geq 6$. The unthresholded symmetric endpoint had 495 source rows, 364 relations, 225 queries, and 158 targets. The $p\text{Activity} \geq 5$ subset retained 293 rows, 233 relations, 147 queries, and 118 targets; the $p\text{Activity} \geq 6$ subset retained 139 rows, 120 relations, 81 queries, and 70 targets. A relation entered a threshold subset if at least one strict endpoint evidence row met the rule. The precomputed scenario rankings were evaluated on each endpoint subset; histories were not retrained for a potency threshold.

In the fixed-rank sensitivity for the primary scenario, Morgan $\text{Recall}@50$ was 0.302889 without a threshold, 0.297538 at $p\text{Activity} \geq 5$, and 0.323045 at $p\text{Activity} \geq 6$; pair-neighbour values were 0.295852, 0.291302, and 0.319444. Measurement-type strata overlap at relation level and must not be added: the endpoint contained 63 Tier-A-only relations, 288 Tier-B-only relations, and 13 relations with both evidence classes. Inclusive K_i/K_d and $\text{IC}_{50}/\text{EC}_{50}$ groups contained 76 and 301 relations across 52 and 188 queries. Their fixed-rank $\text{Recall}@50$ values were 0.273077 versus 0.319086 for Morgan and 0.253846 versus 0.315983 for pair-neighbour. Popularity, Morgan, and pair-neighbour used all queries in each subset; sequence used only its reported operational denominator. These remain database-measurement descriptions, not mechanistic assay comparisons.

These differences remain descriptions of qualifying database measurement relations. The available NPASS fields did not support a defensible binding, functional, cellular, or indirect assay hierarchy. Accordingly, the assay level was recorded as unknown because it could not be inferred from the available NPASS fields. No mechanistic comparison among assay types is inferred.

S11. Exhaustive versus top-100 pair-neighbour ranking

An approximation restricted to the 100 historical target neighbours with highest sequence similarity was compared with exhaustive pair-neighbour transfer. In the primary cell, both implementations had fixed-salt $\text{Recall}@50$ of 0.295852, uniform-tie $\text{Recall}@50$ of 0.296963, and exact bounds of 0.295852–0.298074. Fixed top-50 relation membership and every query-level uniform-tie $\text{Recall}@50$ contribution agreed, although 136 of 364 endpoint-relation ranks differed. The supported stability claim is therefore limited to primary-cell $\text{Recall}@50$, not complete score/rank equality, other cutoffs, endpoints, or policy cells. Exhaustive evaluation remains primary.

S12. Software and reproducibility

The workflow records source versions and retrieval routes, protocols, scripts, identity-validation fixtures, restricted-intermediate receipts, score-array hashes, aggregate outputs, reports, and figure-source tables with byte count, SHA-256 digest, origin, generator, schema, and rights class. Validation includes complete-InChIKey checking of every representative, exact source-InChI resolution, narrowly scoped protonation normalization, unresolved fail-closed behavior, the bounded /b-layer exception, duplicate sensitivity, date-field and cutoff checks, endpoint-hidden scoring, complete ranks, cross-table invariants, software tests, and identifier scans for public artifacts.

The archived v1.2.0 software release [37] contains the project-authored workflow, source-acquisition instructions, environment specification, identity-validation tests, checksums, and bounded workflow checks. The companion aggregate-data archive [38] contains identifier-free aggregate outputs and the source data underlying the figures and tables. Full reconstruction requires direct acquisition of the NPASS and ChEMBL source databases and the computational resources described in the software release.

NPASS and ChEMBL source records required for reconstruction should be obtained directly from their respective providers. The software and aggregate-data archives [37,38] record source versions, access routes, checksums, validation steps, and the identifier-free derived outputs that can be redistributed. Source rows, canonical relation keys, source structures, endpoint membership, per-query values, and complete ranks are not redistributed. A ChEMBL no-hit remains a leakage-screen outcome, not a negative activity label.

Supplementary Results synthesis

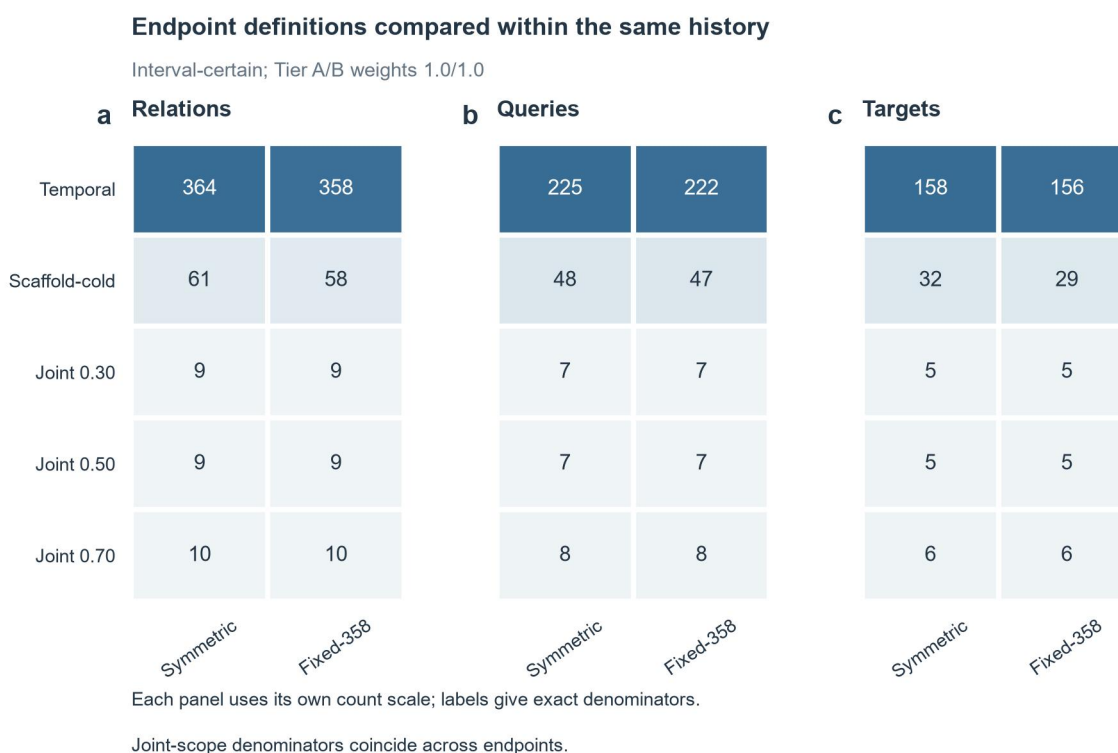
The fixed-relation-key ablation localized the source-row-first effect. Admitting 6,570 more interval-certain source rows within the 4,990 day-only relations changed no representation, mask, score, rank, scope, denominator, or retrieval metric. Full relation expansion to 13,544 historical relations changed 97 query masks and contracted the scaffold-cold population

from 87 to 48 queries. The observed benchmark changes therefore arose when the date policy expanded relation membership.

Under the primary configuration, pair-neighbour temporal Recall@50 was 0.296963. Its all-query interval-minus-day contrast was +0.023312, but the common-cold paired contrast was -0.006944, and the positive unmatched cold-scope difference decomposed mainly into population composition. Exact ties, paired query and PMID-component resampling, joint-scope score-support checks, and the top-100 tie-aware comparison bounded these findings. Complete-InChIKey and duplicate analyses established that unresolved molecular identities or physical-row multiplicity did not determine the estimates.

Together, these analyses support a bounded methodological conclusion: applying date policies before relation aggregation changes the historical evidence and, through it, the population and estimand being evaluated. The evidence does not support biological target discovery, verified negative labels, prospective validation, causal mediation by representation, or superiority inferred from unmatched cold subsets. The contribution is the traceable attribution chain and the conditions under which it can be reproduced.

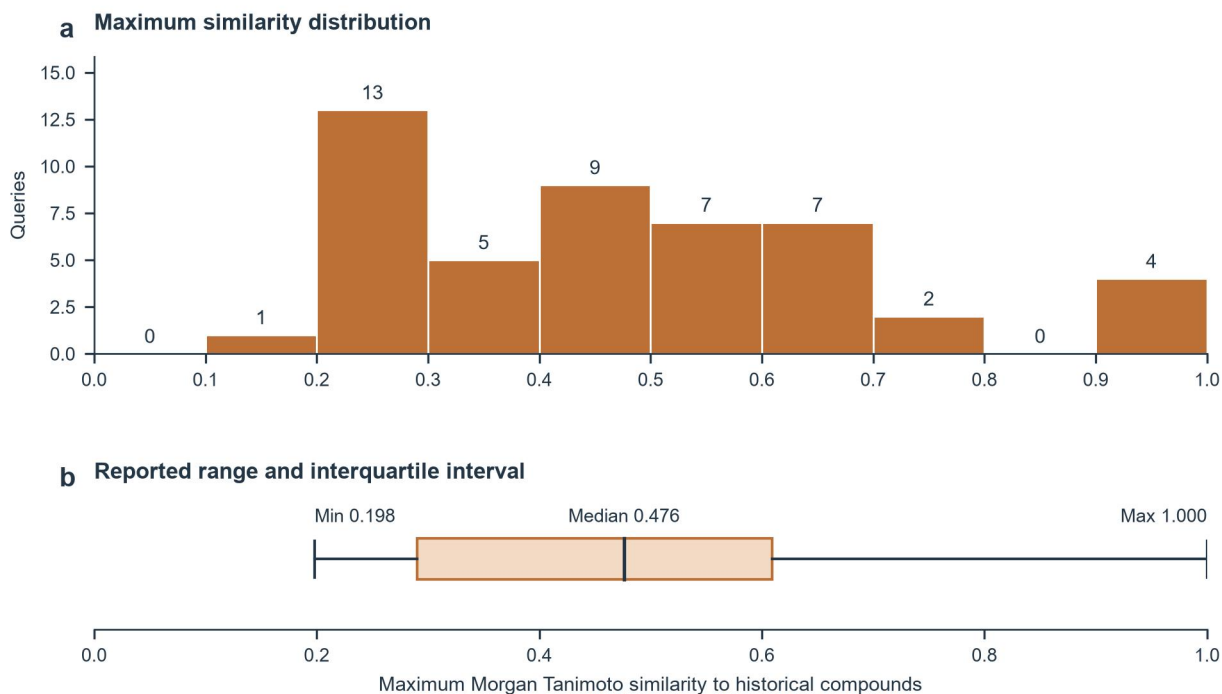
Supplementary Figures



Supplementary Figure S1. Endpoint-definition comparison within one reconstructed history. Relations, queries and targets are shown for the symmetric endpoint and fixed-358 comparator across temporal, exact scaffold-cold and three joint cold scopes. Each panel uses its own count scale and prints exact denominators.

Exact scaffold absence does not ensure low fingerprint similarity

48 queries | IQR 0.290–0.609 | 4 queries at similarity ≥ 0.9

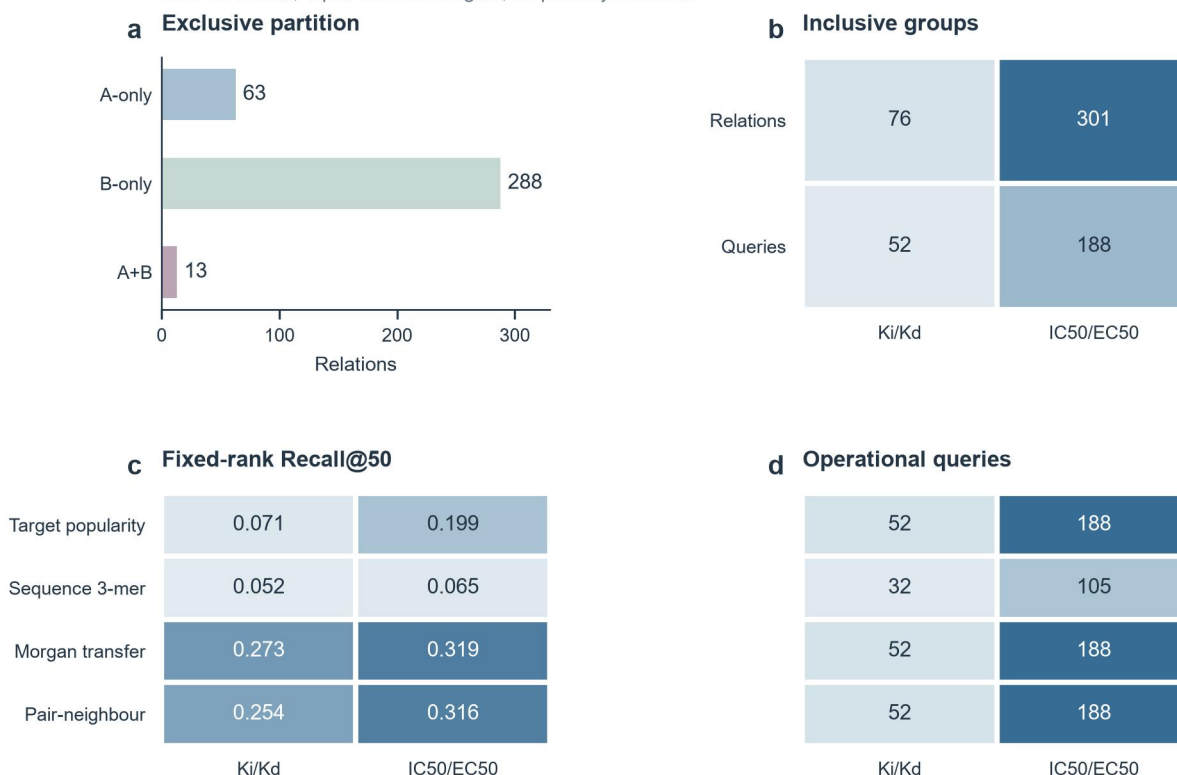


Panel a retains the ten original bins; panel b uses reported summaries only.

Supplementary Figure S2. Fingerprint similarity among 48 exact-scaffold-cold queries. (a) Maximum Morgan radius-2, 2,048-bit Tanimoto similarity to any identity-validated interval-history compound, shown in defined bins. (b) Reported range and interquartile interval; the median is 0.476, and 4 queries have maximum similarity at least 0.9. No density was fitted and no query identifiers are shown.

Measurement composition and fixed-rank sensitivity

Interval-certain; equal Tier A/B weights; no potency threshold



Inclusive groups share 13 relations; their denominators cannot be added.

Supplementary Figure S3. Measurement composition and fixed-rank sensitivity. (a) The mutually exclusive Tier-A-only, Tier-B-only and Tier-A-plus-B relation partition (63/288/13). (b) Inclusive Ki/Kd and IC50/EC50 relation and query counts; the groups overlap by 13 relations and their denominators cannot be added. (c, d) Fixed-rank Recall@50 and operational query counts for four baselines under the primary interval-certain, equal-weight reconstruction. These estimates are distinct from the main uniform-tie estimand.

Supplementary Tables workbook index

Sheet	Contents	Primary interpretation boundary
README	Workbook scope, estimands, scenario labels, missing-value rules, and rights notice	Aggregate, identifier-free companion; not a row-level data release
S1_CandidateUniverse	Date-free strict and policy-specific historical counts	Date policy follows strict gating and precedes relation-key restriction
S2_EndpointFlow	Symmetric endpoint flow and fixed-358 comparator	ChEMBL no-hit is not a negative label
S3_FactorialMetrics	Date × weight metric grid using identity-validated structures and operational denominators	NE is a status, not zero
S4_EffectDecomposition	Total date, weight, and date-by-weight contrasts	Representation-only date effects are not separately identified
S5_ScopeDenominators	Temporal, exact-scaffold-cold, and joint-scope counts	Cold scopes are scenario-local diagnostics
S6_TieUncertainty	Fixed-salt, uniform-tie, exact-bound, query-, component-, and finite-sample summaries	Tie bounds and sampling intervals answer different questions
S7_NullSummary	Observed-to-analytic-random-order enrichment	Random catalogue ordering is not an empirical negative control

Sheet	Contents	Primary interpretation boundary
S8_PotencyAssay	Potency and measurement membership and fixed-rank sensitivities	Assay biological level remains unknown
S9_Top100Exhaustive	Pair-neighbour rank and top-50 membership comparison	Local Recall@50 stability is not full-rank equivalence
S10_RightsManifest	Artifact classes, permitted aggregation, and restricted/link-only fields	External identifier-bearing source content is not redistributed
S11_BuildReceipts	Software versions, protocol, hashing, and execution records	Documents the computational environment and workflow
S12_QACrossChecks	Cross-table invariants and validation checks	Automated checks assess internal consistency, not external validation
S13_EndpointExclusions	Aggregate candidate inclusion/exclusion characteristics	Target family was unavailable; no-hit is not a negative
S14_MeasurementOverlap	Exclusive Tier partition and inclusive measurement groups	Inclusive groups overlap and must not be added
S15_ScaffoldSimilarity	Maximum Morgan similarity summary and histogram	Exact scaffold absence does not imply low similarity
Audit_DateInputs	Field $\times$ precision $\times$ era $\times$ inclusion, electronic/print concordance, and cutoff-distance audit	Relation counts across row strata are non-additive
Audit_Structure	Complete-InChIKey identity paths, exceptions, and duplicate sensitivity	Resolution paths and duplicate sensitivity are reported separately
S16_ScopeTransitions	Retained, exited, and entered cold-scope membership counts	Transitions are aggregate and history-specific
S17_FixedKeyAblation	Three-condition fixed-relation-key attribution plus target-frequency redistribution	Evaluated at equal Tier weights
S18_PairedRobustness	Common-query contrasts, paired resampling, leave-one-component-out, and cold composition identity	Decomposition is descriptive and non-causal
S19_Top100Tie	Top-100 versus exhaustive fixed-salt and uniform-tie Recall@50	Equality is local to the primary cell and cutoff
S20_JointScoreSupport	Positive-candidate and relevant-target score support in joint scopes	Structural zero support differs from observed ranking failure
Audit_Reproducibility	Source-input matrix, bounded workflow receipt, software tests, and repository checksums	Bounded checks do not constitute independent full replication

Abbreviations

C31, ChEMBL version 31; IC50, half-maximal inhibitory concentration; EC50, half-maximal effective concentration; K_d, dissociation constant; K_i, inhibition constant; MRR, mean reciprocal rank; NDCG, normalized discounted cumulative gain; NPASS, Natural Product Activity and Species Source database; PMID, PubMed identifier.

References cited in the supplement

Reference numbers follow the main manuscript. The entries cited in this supplement are reproduced below for convenience; gaps in numbering are intentional.

27. Blackshaw J, Gaulton A, Bento AP, De Veij M, Mendez Lopez D, Bosc N, et al (2022) ChEMBL database release 31 [dataset]. EMBL-EBI. <https://doi.org/10.6019/chembl.database.31>.
28. National Library of Medicine. PubMed User Guide. <https://pubmed.ncbi.nlm.nih.gov/help/>. Accessed 4 September 2026.
33. McSherry F, Najork M (2008) Computing information retrieval performance measures efficiently in the presence of tied scores. In: Macdonald C, Ounis I, Plachouras V, Ruthven I, White RW (eds) Advances in Information Retrieval. ECIR 2008. Lecture Notes in Computer Science, vol 4956. Springer, Berlin, Heidelberg, pp 414–421. https://doi.org/10.1007/978-3-540-78646-7_38.
34. Efron B (1981) Nonparametric standard errors and confidence intervals. Can J Stat 9(2):139–158. <https://doi.org/10.2307/3314608>.

35. Field CA, Welsh AH (2007) Bootstrapping clustered data. *J R Stat Soc B* 69(3):369–390. <https://doi.org/10.1111/j.1467-9868.2007.00593.x>.
36. Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26(4):404–413. <https://doi.org/10.1093/biomet/26.4.404>.
37. Hu C (2026) Source-aware temporal reconstruction exposes policy dependence in natural-product target-retrieval benchmarks: v1.2.0 [software]. Zenodo. <https://doi.org/10.5281/zenodo.22343339>.
38. Hu C (2026) Identifier-free aggregate outputs for source-aware temporal natural-product target retrieval: v1.2.0 [dataset]. Zenodo. <https://doi.org/10.5281/zenodo.22343340>.