CHEMTRACE: RELATION-TYPED PROVENANCE FOR MOLECULAR PRETRAINING-CORPUS AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular models are often pretrained on large chemical corpora and evaluated on reused public benchmarks. Potential overlap between benchmark molecules and pretraining records complicates the interpretation of model evaluation: exact matching misses standardized-parent, stereochemical, and tautomeric relations, whereas generic similarity cannot establish which candidate-corpus record supports a declared molecular relation. Our goal is to recover attributable molecular exposure—an auditable typed relation between a benchmark molecule and an inspectable candidate-corpus record—without requiring model access. We introduce CHEMTRACE, a provenance framework that indexes canonical identity, standardized parent, stereochemistry-stripped connectivity, and canonical tautomer form. It emits provenance certificates, versioned JSON records that bind fixed corpus, split, and intervention specifications to relation-typed retrieval evidence. Independently defined ChEMBL36 registry relations and curated Tautobase tautomer pairs yield 0.913–0.992 primary-relation recall across the three non-exact relations. CHEMTRACE recovers every changed non-exact exposure in the controlled audit, emits no certificate on 500 deliberately separated registry negatives, and remains silent across a 3,948-pair hard-neighborhood bank explicitly conditioned to be typed-relation-negative. Across six MoleculeNet datasets and 13 same-protocol audit routes, CHEMTRACE is the only evaluated method combining complete changed-only recall, a 0.0036 ± 0.0032 background hit rate, and zero decoy hits. A two-stage scaffold/seed bootstrap finds no consistently positive downstream inflation across 51 primary-estimator cells. Together, CHEMTRACE gives benchmark auditors a reproducible, record-level way to identify source-linked molecular exposure without treating chemical-neighborhood similarity as provenance; downstream model effects remain a separate empirical question.

1 INTRODUCTION

Public benchmarks standardize comparison of machine-learning models for drug discovery (Wu et al., 2018; Huang et al., 2021). Their reuse raises training–evaluation overlap concerns. Molecular language models pretrained on PubChem-, ZINC-, and literature-derived corpora are evaluated on public property tasks (Ross et al., 2022). When corpus documentation is incomplete or benchmark molecules reappear upstream, scores may combine generalization with prior exposure. Molecular audits therefore need provenance evidence alongside predictive quality (Ye et al., 2025; Bartmann et al., 2026).

The representation obstacle is immediate. Raw-string equality is brittle because one molecular graph can have multiple valid Simplified Molecular Input Line Entry System (SMILES) strings, i.e., linear encodings of the same molecular structure. Canonical identity removes serialization variance yet still misses a benchmark salt stored upstream as its parent, stereochemical variants, and linked tautomeric forms. Exact hits are interpretable, whereas exact-only matching misses exposure preserved under known record relations. Figure 1 summarizes the two failure modes and the relation-typed audit between them.

Broadening the match with fingerprints or scaffolds does not solve the problem. These representations describe molecular neighborhoods and support similarity-aware splitting (Rogers and Hahn, 2010; Bemis and Murcko, 1996; Steshin, 2023; Joeres et al., 2025). Proximity, however, is not provenance:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

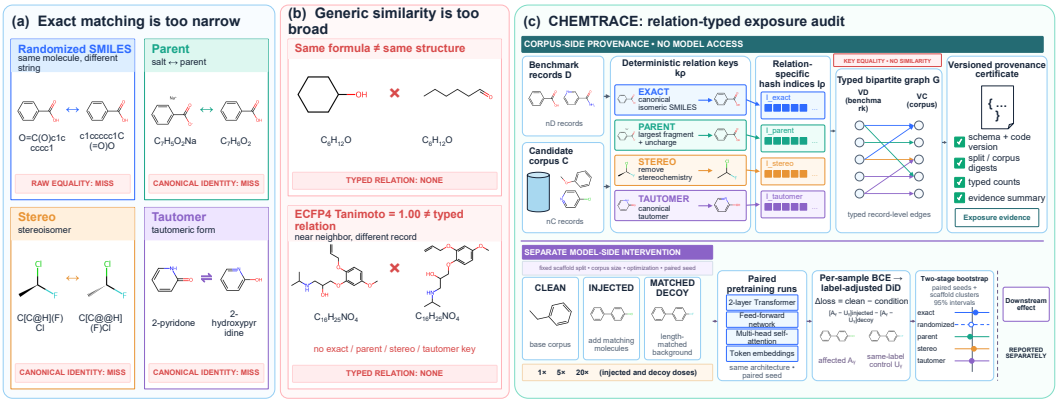

Figure 1: CHEMTRACE separates corpus exposure from model effect. String and canonical-identity checks respectively miss alternative serializations and typed relations, whereas generic similarity retrieves typed-relation-negative neighbors, including the shown ECFP4 Tanimoto-1.00 pair. Relation-specific keys yield attributable matches and versioned certificates; paired interventions estimate downstream effects separately.

isomers can share formulae, activity cliffs—highly similar molecules with discordant endpoint labels—can have near-identical fingerprints, and scaffold neighbors can differ in substituents and behavior. A similarity hit identifies a neighbor, not the record relation connecting it to a query.

An audit match must identify its supporting record and molecular relation and remain reproducible. CHEMTRACE supplies relation-specific keys while retaining fingerprints and scaffolds as neighborhood evidence. Model-side analyses can then test whether attributed exposure changes benchmark performance.

Objective and contributions. We audit a fixed, inspectable candidate pretraining-corpus snapshot to recover relation-typed exposure without overcalling chemical neighbors, then test its downstream effect under controlled pretraining. Model-side membership inference addresses the complementary setting in which only model behavior is observable. Our contributions are:

- **Relation-typed attribution.** Unlike exact-only and neighborhood-similarity checks, CHEMTRACE links non-exact molecular exposure to a specific candidate-corpus record through explicit parent, stereo, and tautomer relations and records the evidence in a versioned provenance certificate.
- **Coverage–specificity evidence.** Independent non-exact relation positives, 500 deliberately separated registry negatives, a conditional 3,948-pair hard-neighborhood bank, leave-one-route-out ablations, and a 13-route common protocol show complete changed-only recall with a 0.0036 ± 0.0032 background hit rate and zero decoy hits.
- **Downstream separation.** Controlled pretraining across six MoleculeNet datasets yields no consistently positive effect across 51 primary cells, supporting separate evaluation of corpus provenance and downstream model effect.

2 RELATED WORK

2.1 BENCHMARK CONTAMINATION AND TRAINING-DATA DETECTION

Training-data detection infers membership from model outputs, losses, or features (Shokri et al., 2017). Memorization, extraction, deduplication, and corpus documentation connect repeated records to evaluation overlap (Carlini et al., 2019; 2021; Lee et al., 2022; Kandpal et al., 2022; Dodge et al., 2021); detectors vary across modalities and optimization stages (Zhu et al., 2025; Luo et al., 2026; Wang et al., 2026). In molecular ML, Atomas uses canonical SMILES to deduplicate its PubChem-derived initial training set against downstream datasets, while MolecularIQ flags leakage risks in chemistry benchmarks (Zhang et al., 2025; Bartmann et al., 2026). CoDeC and semantic

perturbations probe behavior; KTD and Data Taggants add error control or ownership evidence (Zawalski et al., 2026; Park et al., 2026; Hu et al., 2025; Bouaziz et al., 2025). Behavioral detectors address inaccessible training records; CHEMTRACE complements them when a candidate corpus is inspectable, without relying on a signature that survives later optimization.

2.2 MOLECULAR IDENTITY, BENCHMARK LEAKAGE, AND DATA PROVENANCE

Similarity-aware evaluation. Molecular evaluation uses extended-connectivity fingerprint (ECFP) similarity and Bemis–Murcko scaffolds—ring/linker cores obtained after side-chain removal—to control redundancy (Bemis and Murcko, 1996; Rogers and Hahn, 2010). Random, temporal, and structural splits yield different estimates because redundancy can reward memorization over prospective generalization (Sheridan, 2013; Wallach and Heifets, 2018; Mayr et al., 2018; Yang et al., 2019). Lo-Hi, DataSAIL, and Hestia-GOOD stratify train–test similarity (Steshin, 2023; Joeres et al., 2025; Fernández-Díaz et al., 2025); audits identify duplicates, label errors, and activity cliffs (Northcutt et al., 2021; Schuh et al., 2026). Recent journal studies likewise show that leakage can inflate binding-affinity and drug–target benchmarks and can enter downstream biological prediction through pretrained language models (Graber et al., 2025; de la Fuente et al., 2025; Szymborski and Emad, 2026). These tools improve benchmark hygiene and out-of-distribution (OOD) evaluation; parent, stereo, and tautomer provenance require a separate attributable relation test. Their decision mechanisms serve as common-protocol neighborhood comparators in our evaluation.

Chemical identity and provenance. SMILES, InChI, and curation pipelines make identity computable, while standardization changes salts, charge, stereochemistry, and tautomers (Weininger, 1988; Heller et al., 2015; Fourches et al., 2010; Hähnke et al., 2018; Bento et al., 2020; Sayle, 2010). Production pipelines and RDKit SuperParent compose these operations but collapse the transformation path (Bento et al., 2020; RDKit contributors, 2026). ChEMBL preserves hierarchy and source metadata independent of a tested key (Mendez et al., 2019). Database provenance separates origin from transformation, while statements, datasheets, and model cards structure ML disclosure (Buneman et al., 2001; Cheney et al., 2009; Moreau and Missier, 2013; Bender and Friedman, 2018; Gebru et al., 2021; Mitchell et al., 2019). Foundation-model evaluations broaden tasks or use symbolic verification (Ross et al., 2022; Ye et al., 2025; Bartmann et al., 2026); CHEMTRACE preserves each positive's declared relation and source record. Appendix A.5 evaluates SuperParent as a collapsed reference.

3 CHEMTRACE: RELATION-TYPED MOLECULAR PROVENANCE

CHEMTRACE asks a simple corpus-side question: for each benchmark molecule, which inspectable corpus records are linked to it by a chemically meaningful relation? It answers with relation-specific keys, a source-preserving bipartite graph, and a provenance certificate bound to fixed audit inputs. Here, provenance means record attribution within the audited snapshot, not historical lineage.

3.1 PROBLEM FORMULATION AND AUDIT ASSUMPTIONS

Exposure is defined at the corpus level; model behavior is evaluated separately. Formally, let $D = \{x_i\}_{i=1}^{n}$ be a benchmark, $C = \{c_j\}_{j=1}^{m}$ a candidate pretraining corpus, and

$$\mathcal{R} = \{\text{exact}, \text{parent}, \text{stereo}, \text{tautomer}\}, \tag{1}$$

the audited relation family. For each relation $\rho \in \mathcal{R}$, a deterministic normalization k_ρ maps a valid record to a key or abstains. Its indicator is

$$M_\rho(x, c) = \mathbf{1}[k_\rho(x) \neq \varnothing \ \wedge \ k_\rho(x) = k_\rho(c)]\,. \tag{2}$$

The observable vector is $\mathbf{z}(x) = (z_\rho(x))_{\rho\in\mathcal{R}}$, where $z_\rho(x) = \max_{c\in C} M_\rho(x, c)$. Components are inclusive marginal indicators; reporting precedence never changes pair-level flags. The family is atomic: each key implements one route. Path-typed keys can represent composites without changing the graph, but arbitrary compositions are outside the evaluated family.

Audit assumptions and non-goals. CHEMTRACE passively audits an inspectable, versioned corpus snapshot using benchmark records and either corpus bytes or a reproducible key index; no model access is required. It tests whether a record supports a declared relation, not historical ingestion, actual consumption during optimization, or transformations outside that family. Hashes bind the snapshot and analysis; behavioral fields remain separate evidence.

3.2 RELATION-TYPED NORMALIZATION

Canonical identity misses standardized-parent, stereo, and tautomer relations. Generic similarity has the opposite failure mode: it also retrieves typed-relation-negative neighbors. Each audited relation therefore receives its own deterministic normalization key. We include relations that have a distinct record meaning, occur in standardization practice, and admit external or registry-anchored validation.

We implement the four routes with RDKit 2026.03.5 (RDKit contributors, 2026). The *exact* key is sanitized canonical isomeric SMILES; *parent* selects the largest fragment, uncharges it, and canonicalizes; *stereo* removes stereochemistry before non-isomeric canonicalization; and *tautomer* applies `TautomerEnumerator` before isomeric canonicalization. Appendix A gives exact calls and tie rules. Full InChIKey identity and equality of its connectivity block are direct competitors (Heller et al., 2015); randomized isomeric SMILES is a serialization control preserving the exact key.

If a pair satisfies multiple relation keys, all Boolean relation flags are retained. For reporting only, the fixed precedence exact $\rightarrow$ parent $\rightarrow$ stereo $\rightarrow$ tautomer assigns a *primary relation*. An initial parse failure or route-specific execution exception returns an empty key for that route and is logged as an abstention; CHEMTRACE never substitutes another relation's key after failure. The resulting keys define reproducible operational relation classes under the identified code and software snapshot. Appendix A specifies every transformation and stability check.

3.3 PROVENANCE GRAPH AND CERTIFICATES

Attribution requires each positive match to remain tied to its supporting candidate-corpus record. We therefore represent typed matches with a bipartite graph $G = (V_D \cup V_C, E)$ with benchmark nodes V_D, corpus-record nodes V_C, and edges

$$E = \{(x_i, c_j, \rho) : \rho \in \mathcal{R},\ M_\rho(x_i, c_j) = 1\}. \tag{3}$$

Each corpus record contributes to a relation-specific hash index I_ρ. The implementation stores key multiplicities for constant-time membership and count queries; equality joins against the fixed, ordered corpus reconstruct $\{j : k_\rho(c_j) = k_\rho(x_i)\}$ and hence E without an all-pairs comparison. Construction costs $O(m|\mathcal{R}|)$ key evaluations and querying costs $O(n|\mathcal{R}| + |E|)$. Source-indexed intervention records preserve attribution for the controlled interventions. Appendix A.3 gives executable steps and invariants.

The bipartite graph unifies evidence that would otherwise be fragmented across four independent hash lookups: one benchmark–corpus pair may support several typed relations, and one normalized key may retrieve duplicate corpus positions. The implementation represents this graph through a fixed corpus snapshot, relation-key counts, pair-level flag tables, and source-indexed intervention records, allowing record-level edges to be regenerated when required instead of duplicating them in every certificate. Relation evidence, display precedence, and downstream fields remain separate, so presentation choices do not alter the underlying typed match evidence.

Provenance-certificate output. A provenance certificate is a versioned JSON record generated from fixed split, corpus, and intervention specifications together with the typed retrieval audit. Here, *certificate* means a schema-validated, version-bound provenance record, not a statistical certificate of training membership or causal model effect. It binds code and schema versions, input digests and record counts, aggregate typed-retrieval diagnostics, a seed-level downstream summary when available, and an evidence summary. Recomputing the relation keys from the bound inputs and identified code revision regenerates the record-level graph, so the record need not duplicate every edge. Appendix H summarizes this record contract and a validated example.

4 Evaluation Protocol

Four questions organize the evaluation. RQ1 tests the keys against relation-specific validation data. RQ2 asks what typed molecular exposure exact identity misses. RQ3 tests whether similarity-based alternatives preserve provenance meaning under one protocol. RQ4 asks whether detected exposure consistently inflates downstream performance. Sections 5.1–5.4 answer these questions in order, keeping relation validity, retrieval scope, and downstream consequence separate.

4.1 Tasks, Corpora, and Controlled Exposure

The study covers six MoleculeNet molecular-property datasets spanning nine binary classification endpoints, a fixed 50,000-record corpus, three injection doses, three model families, and 13 audit routes. ChEMBL36 (ChEMBL Team, 2025) molecular hierarchy and Standard-InChI fields provide parent and stereo positives; Tautobase supplies curated tautomer pairs. Controlled exposures test retrieval and downstream effects. All 13 audit-route comparisons reuse the same splits, intervention records, corpus variants, and five background seeds; downstream effect analyses use the seed counts reported in Appendix F.

We additionally evaluate five prespecified strata. Cold-start molecules have a scaffold absent from the training partition, whereas warm-start molecules have one already present; long-tail molecules have training scaffold count at most one. Low-similarity OOD molecules have maximum background Morgan-fingerprint Tanimoto similarity below 0.70, and high-background-similarity molecules use the complementary threshold of at least 0.70. Robustness checks use a bidirectional gated recurrent unit (BiGRU) sequence encoder and a graph isomorphism network (GIN) in addition to the primary Transformer.

We fix scaffold 80/10/10 splits with seed 20260824 and select 64 label-stratified test molecules per endpoint (576 records). A 50,000-record clean corpus excludes every test-set canonical identity. At doses $d \in \{1, 5, 20\}$, a deterministic nearest-length allocator replaces background records with transformed queries. Dose-matched controls are sampled with replacement from the clean background without relation, scaffold, or similarity exclusion; the decoy hit rate therefore measures incidental relations rather than assuming their absence. Appendix D specifies allocation and tie breaking.

Intervention records store source indices, structures, hashes, tasks, relations, doses, and identifiers; splits, optimization, and seeds remain fixed. External ChEMBL36 and Tautobase pairs provide relation ground truth. Downstream analysis retains the preselected affected set, whereas retrieval uses Q_ρ^Δ; Appendix F also analyzes $A \cap Q_\rho^\Delta$. The paired contrast isolates injection relative to its decoy; clean runs diagnose baseline exposure and comparability.

4.2 Comparators, Metrics, and Statistical Analysis

Baselines are grouped by mechanism: raw-string equality and canonical identity; full or connectivity-block InChIKey equality; MinHash; ECFP similarity; Bemis–Murcko scaffolds; and four adapted split/curation rules. Literature-inspired decision rules are reimplemented under the common protocol and labeled *Adapted*; direct implementations are labeled *Self*.

We compute 2,048-bit, radius-2 Morgan fingerprints (ECFP4) without chirality and compare them by Tanimoto similarity. The common-protocol threshold route uses Tanimoto ≥ 0.80; the ≥ 0.90 route is evaluated separately in the hard-neighborhood stress analysis. SMILES-shingle MinHash uses canonical-SMILES character 3-grams, 32-component sketches, and an estimated-Jaccard threshold of 0.50. Morgan MinHash applies the same sketch size to active ECFP bits with threshold 0.80 (Broder, 1997).

Adapted comparators. The Lo-Hi-style high-similarity split calls a pair at ECFP Tanimoto ≥ 0.60. The DataSAIL-style similarity cluster calls either Bemis–Murcko scaffold equality or Tanimoto ≥ 0.70. The TDC-inspired curation rule (Huang et al., 2021) is our common-protocol adaptation: it flags canonical identity, full InChIKey identity, scaffold equality, or Tanimoto ≥ 0.80; the biomolecular-audit bundle adapted from Schuh et al. (2026) replaces full InChIKey with connectivity-block identity in the same union. All rules and thresholds are fixed before evaluation and applied identically to clean, injected, and decoy corpora.

For a relation ρ, changed-only recall is computed over queries whose injected representation changes the canonical key,

$$\mathrm{Rec}_{\Delta}(\rho) = \frac{1}{|Q_{\rho}^{\Delta}|} \sum_{x \in Q_{\rho}^{\Delta}} \mathbf{1}[\text{route retrieves the injected record}]. \tag{4}$$

where $Q_{\rho}^{\Delta} \subseteq Q$ contains queries whose injected representation changes the canonical key. For each task–relation–dose cell with query set Q, the background hit rate b is the fraction of Q called positive against the sampled clean corpus, and the decoy hit rate q is the fraction of the same Q that retrieves at least one dose-matched decoy target. Randomized SMILES uses injection recall when raw serialization changes. Changed-only recall is the primary retrieval metric; b and q diagnose incidental calls, and $S = \mathrm{Rec}_{\Delta} - b - q$ is an uncalibrated cell-level adjusted score. Higher S rewards changed-exposure recovery while penalizing incidental retrieval. Table 2 averages Rec_{Δ} and S over 78 defined cells, but b and q over all 81 semantic cells. Registry negatives test construction independence; the bank conditioned on all four CHEMTRACE rules being false tests typed-meaning preservation. Common-protocol results are mean $\pm$ SD over seeds 101, 103, 107, 109, and 113.

In Figure 2, the generic oracle is the per-cell upper envelope over scaffold and ECFP ≥ 0.70, ≥ 0.80, and ≥ 0.90 routes. Bars show raw changed-only recall; the gain subtracts dose-matched decoy calls cellwise. Its 95% percentile-bootstrap interval uses the 2.5th and 97.5th quantiles of 20,000 draws over nine task clusters, keeping each cluster's relation–dose cells together.

We operationalize downstream performance inflation as a positive label-adjusted reduction in per-sample binary-cross-entropy loss for affected molecules relative to same-label controls and matched decoys. Let $\mathscr{S}$ be the paired training-seed set, $\ell_{s,\kappa}(x)$ the loss under condition κ, and $d_s(x) = \ell_{s,\mathrm{decoy}}(x) - \ell_{s,\mathrm{injected}}(x)$. For label-specific affected and unaffected sets A_y and U_y, let $\bar{d}_{s,B}$ be the mean of $d_s(x)$ over set B. The primary label-adjusted DiD effect is

$$\widehat{\theta} = \frac{1}{|\mathscr{S}|} \sum_{s \in \mathscr{S}} \frac{\sum_y |A_y| \left(\bar{d}_{s,A_y} - \bar{d}_{s,U_y}\right)}{\sum_y |A_y|}. \tag{5}$$

Positive values mean that injection benefits affected molecules more than decoy exposure, relative to same-label controls. Equation 5 is algebraically identical to subtracting clean-to-condition affected–control contrasts; clean loss cancels from the point estimator but remains a paired diagnostic. Its 95% interval uses the 2.5th and 97.5th quantiles of 4,000 two-stage draws resampling paired seeds and, within seed, test-set scaffold clusters (Appendix F).

5 RESULTS

5.1 EXTERNAL RELATIONS VALIDATE THE NON-EXACT KEYS

Table 1 summarizes an exact-identity registry sanity check together with independently defined parent, stereo, and tautomer relations. Primary-relation recall is 0.9920 for parent, 0.9615 for stereo, and 0.9134 for tautomer; precision is at least 0.9961 across all three. The stereo key itself matches 1,927/2,000 (0.9635), with four additional positives displayed as parent under the fixed precedence. The 15 cross-positive primary-relation confusions remain positive typed certificates. Appendix A.6 summarizes rule-based diagnostics for the first 100 recorded tautomer misses.

Leave-one-route-out ablation confirms that the three non-exact relation keys contribute nonredundant coverage: external relation-union recall falls from 0.9654 to 0.5194 without parent, 0.7167 without stereo, and 0.7353 without tautomer (Appendix Table B.2).

The remaining stereo and tautomer misses quantify cross-standardization disagreement among ChEMBL Standard InChI, Tautobase, and the RDKit operational keys; an untyped similarity score would collapse these disagreements into neighborhood matches.

5.2 EXACT MATCHING MISSES TYPED MOLECULAR EXPOSURE

Figure 2 quantifies retrieval on randomized representations and chemically changed parent, stereo, and tautomer queries.

Table 1: Validation of the four relation keys. Exact uses a ChEMBL36 registry self-identity sanity set; parent, stereo, and tautomer positives are defined independently from ChEMBL36 or Tautobase. Precision and recall are one-vs-rest metrics for the precedence-assigned primary relation, not corpus-wide PPV/FPR; the pair-level evaluation table retains all relation flags.

Relation	External source	Support	Precision	Recall	F1
Exact	ChEMBL36 identity	1,000	1.0000	1.0000	1.0000
Parent	ChEMBL36 hierarchy	3,000	0.9983	0.9920	0.9952
Stereo	Standard InChI	2,000	0.9979	0.9615	0.9794
Tautomer	Tautobase	1,675	0.9961	0.9134	0.9530

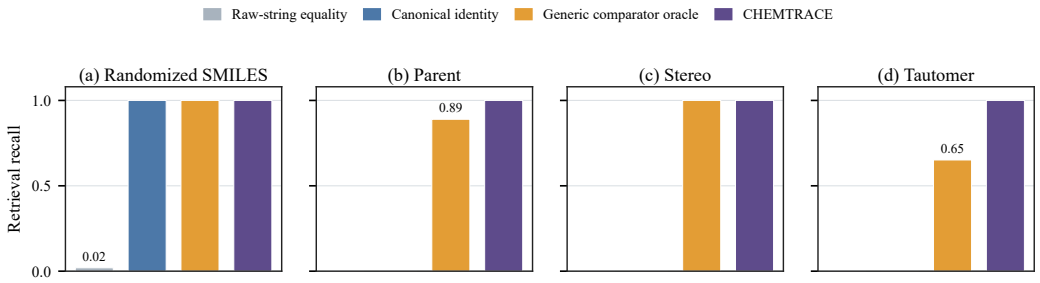

Figure 2: Exact checks fail at different levels. Raw-string equality misses serialization changes, whereas canonical identity misses chemically changed parent, stereo, and tautomer exposure. Panel (a) averages randomized-SMILES injection recall; panels (b)–(d) report raw changed-only recall after excluding canonical-identity matches. CHEMTRACE recovers all evaluated changed relations.

For randomized SMILES, raw-string equality recalls only 1.97% on average across doses, whereas canonical identity restores 100%; CHEMTRACE invokes no non-exact relation when its exact key suffices.

Across 78 changed cells, raw-string and canonical identity recall zero; CHEMTRACE reaches 1.000 versus oracle means of 0.8410 raw and 0.8408 net. Its net gain is 0.1592 (95% task-cluster bootstrap [0.1101, 0.2181]), largest on tautomers.

Thus, exact-only matching misses exposure preserved by explicit molecular-record relations. Figure 3 tests the complementary question: whether broader neighborhood rules recover these cases without overcalling typed-relation-negative molecules.

5.3 SIMILARITY IS TOO BROAD UNDER A COMMON AUDIT PROTOCOL

This typed-negative bank is a conditional semantic stress test, not an independent false-positive-rate benchmark. CHEMTRACE is silent by construction; broader routes fire as their criteria widen. The bank compares semantic overcalling, while the 500 registry pairs and background/decoy analyses supply independent negative evidence.

A separately constructed set of 500 deliberately separated ChEMBL36 registry negatives yields 0/500 CHEMTRACE certificates, providing an independent construction check rather than a hard-neighborhood specificity estimate.

CHEMTRACE retrieves all 432 controls. Conditional call rates rise from 0.38% for InChIKey connectivity to 16.79% for ECFP$\geq$ 0.90 and above 55% for the broader split/audit adaptations, despite all 3,948 pairs being typed-relation-negative.

Table 2 tests whether this coverage–specificity trade-off persists beyond the fixed stress bank by comparing CHEMTRACE with 12 alternative routes under identical intervention records, splits, and corpora.

CHEMTRACE ranks first under the common protocol with $S = 0.9963 \pm 0.0033$, complete changed-only recall, a 0.0036 ± 0.0032 background hit rate, and zero decoy hits. The descriptive per-cell oracle reaches 0.9364 ± 0.0064. None matches CHEMTRACE's coverage and low incidental retrieval.

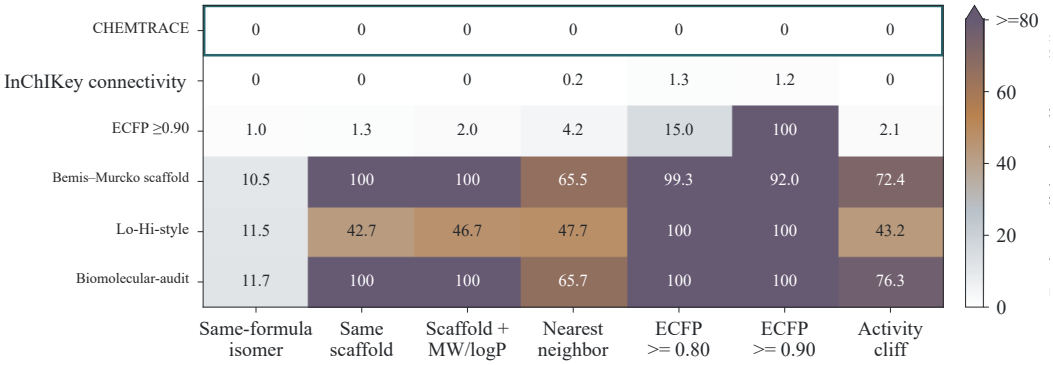

Figure 3: Conditional semantic stress test across seven typed-negative hard-neighborhood families. CHEMTRACE is silent by construction; broader neighborhood routes increasingly call pairs. These rates are not an independently sampled false-positive-rate estimate.

Table 2: Five-seed common-protocol comparison across 13 routes. $S = \mathrm{Rec}_\Delta - b - q$ is computed cellwise. Rec_Δ and S average 78 defined cells; b and q average all 81, so displayed means need not subtract exactly. Values are mean ± SD; Self denotes direct implementations and Adapted denotes same-protocol reimplementations.

Method	Implementation	Coverage	Incidental retrieval		Summary
		Rec_Δ	Background b	Decoy q	S
CHEMTRACE	Self	$\mathbf{1.0000 \pm 0.0000}$	$\mathbf{0.0036 \pm 0.0032}$	$\mathbf{0.0000 \pm 0.0000}$	$\mathbf{0.9963 \pm 0.0033}$
Raw-string equality	Self	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Canonical identity	Self	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
Full InChIKey identity	Self	0.0973 ± 0.0000	0.0038 ± 0.0053	0.0000 ± 0.0000	0.0934 ± 0.0055
InChIKey connectivity identity	Self	0.8286 ± 0.0000	0.0125 ± 0.0099	0.0000 ± 0.0000	0.8156 ± 0.0103
Morgan MinHash	Adapted	0.7387 ± 0.0000	0.0076 ± 0.0050	0.0000 ± 0.0000	0.7308 ± 0.0052
SMILES-shingle MinHash	Adapted	0.8964 ± 0.0000	0.2597 ± 0.0127	0.0156 ± 0.0000	0.6162 ± 0.0126
Lo-Hi-style high-similarity split	Adapted	0.9267 ± 0.0000	0.0465 ± 0.0155	0.0023 ± 0.0000	0.8783 ± 0.0155
Biomolecular-audit bundle	Adapted	0.9332 ± 0.0000	0.1674 ± 0.0343	0.0006 ± 0.0000	0.7639 ± 0.0342
Morgan ECFP ≥ 0.80	Self	0.7303 ± 0.0000	0.0069 ± 0.0044	0.0000 ± 0.0000	0.7231 ± 0.0046
TDC-inspired curation	Adapted	0.8627 ± 0.0000	0.1653 ± 0.0367	0.0006 ± 0.0000	0.6956 ± 0.0367
DataSAIL-style similarity cluster	Adapted	0.8584 ± 0.0000	0.1694 ± 0.0403	0.0006 ± 0.0000	0.6870 ± 0.0409
Bemis–Murcko scaffold	Self	0.7946 ± 0.0000	0.1608 ± 0.0397	0.0006 ± 0.0000	0.6321 ± 0.0400

CHEMTRACE remains highest in the cold-start, long-tail, OOD, warm-start, and high-background-similarity subsets; in the last, $S = 0.8203 \pm 0.2435$ versus 0.4826 ± 0.1073. Appendix A.5 evaluates collapsed RDKit SuperParent, which also attains $\mathrm{Rec}_\Delta = 1$ but has higher common-protocol b (0.0132 versus 0.0036), lower S (0.9863 versus 0.9963), and no relation typing.

5.4 DETECTED EXPOSURE SHOWS NO CONSISTENT PERFORMANCE INFLATION

Figure 4 reports all 51 primary effect estimates with 95% two-stage scaffold/seed bootstrap intervals and magnifies selected data-size and architecture checks.

Every primary 95% interval crosses zero under the two-stage scaffold/seed bootstrap. Across the 23 primary-protocol cells, the endpoint-clustered global mean is -0.0045 $[-0.0230, 0.0115]$. The broader 51-cell robustness grid gives a similar descriptive aggregate of -0.0004 $[-0.0140, 0.0089]$. The largest positive cell mean is 0.0500 $[-0.0310, 0.1375]$ for BBBP exact $20\times$; cell intervals are descriptive and unadjusted for multiplicity.

Appendix F reports global aggregation, practical-equivalence diagnostics, a changed-only sensitivity, and a centered-bootstrap dispersion diagnostic.

A seed-resampled adjusted-linear diagnostic produces three positive intervals, but these do not persist under the prespecified primary analysis. The primary estimator therefore supports no consistent performance inflation; retrieval-confirmed exposure and downstream behavioral effect should be reported separately.

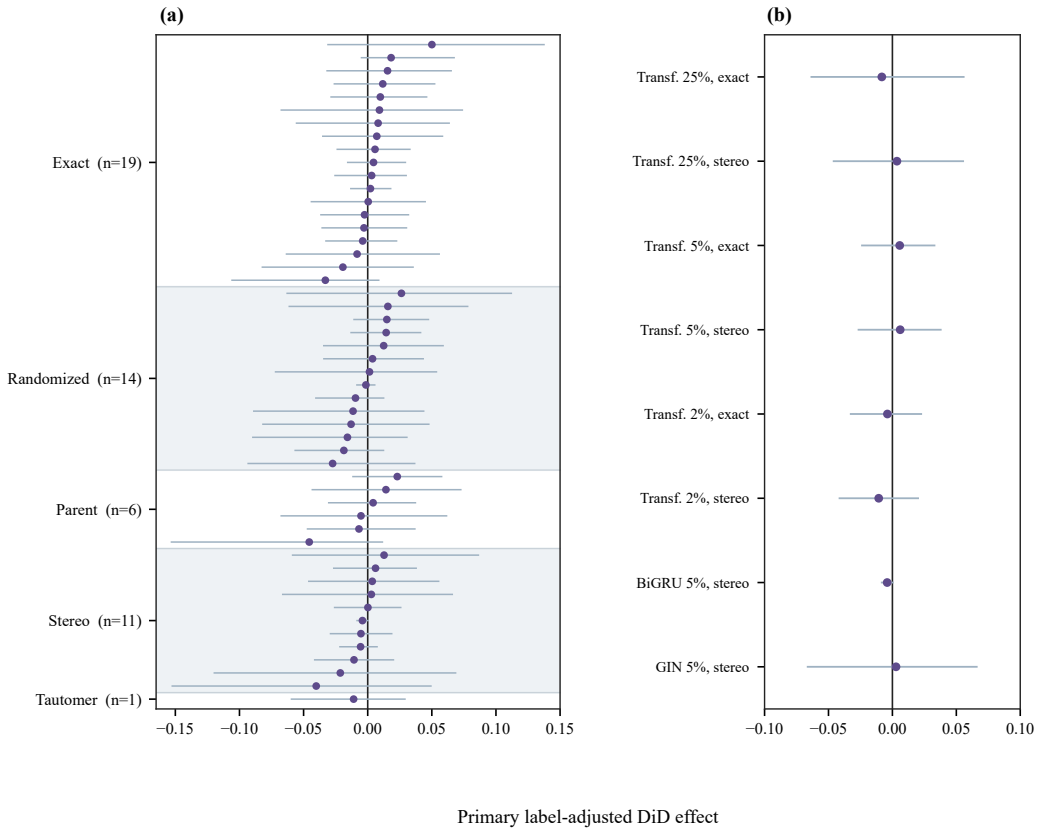

Figure 4: Downstream effects with 95% two-stage scaffold/seed bootstrap intervals. Panel (a) reports all 51 primary-estimator cells; panel (b) magnifies selected data-size and architecture checks. Every interval crosses zero.

6 Discussion: From Exposure Evidence to Evaluation Claims

Exact matching finds canonical repeats; fingerprint and scaffold methods find neighborhoods, but neither types the non-exact record relation. ChemTrace does, and external benchmarks validate its keys. Similarity-aware splits remain useful for OOD evaluation and redundancy control but answer a different question from provenance.

All 51 cell intervals, both global intervals, and the changed-only analysis cross zero. Practical equivalence holds for few cells, while centered-bootstrap dispersion does not exceed sampling uncertainty. Confirmed exposure therefore provides no evidence of uniform positive inflation. Figure 4 contains one tautomer cell; broader tautomer-specific behavioral evaluation remains open.

The measured canonical-key benchmark indexes 498,266 records in 2.506 s and serves about 6,809 queries/s; the ZINC15 proxy yields 24 typed matches among 67,228 records. Provenance is model-independent, whereas the intervention is a mechanism test, not a scaling-law study. Composite paths and larger model/corpus scales remain open.

7 Conclusion

ChemTrace provides relation-typed, record-level corpus exposure evidence. External relation validation and common-protocol comparisons support coverage and specificity; controlled pretraining shows no consistent inflation signal, separating corpus provenance from model effect.

AI USE STATEMENT

During the preparation of this work, the authors used ChatGPT (OpenAI) to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

REPRODUCIBILITY STATEMENT

Sections 3–4 specify the operational relation keys, corpus construction, evaluation protocol, and statistical estimators. Appendices A–H report normalization checks, external-relation and hard-negative construction, dataset partitions, model and optimization settings, seed-level analyses, scaling measurements, and the provenance-record schema. Fixed seeds, software versions, and cryptographic digests are reported for the principal experimental inputs and provenance record. An anonymized supplementary code package accompanies the submission and includes the implementation, fixed retrieval inputs, tables underlying the reported results, integrity checks, a SHA-256 manifest for derived analysis outputs, and documented commands for regenerating the retrieval analyses and rerunning the controlled training experiments.

REFERENCES

Christoph Bartmann, Johannes Schimunek, Mykyta Ielanskyi, Philipp Seidl, Günter Klambauer, and Sohvi Luukkonen. MolecularIQ: Characterizing chemical reasoning capabilities through symbolic verification on molecular graphs. In *International Conference on Learning Representations*, 2026.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. doi:10.1021/jm9602928.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi:10.1162/tacl_a_00041.

A. Patrícia Bento, Anne Hersey, Eloy Félix, Greg Landrum, Anna Gaulton, Francis Atkinson, Louisa J. Bellis, Marleen De Veij, and Andrew R. Leach. An open source chemical structure curation pipeline using RDKit. *Journal of Cheminformatics*, 12:51, 2020. doi:10.1186/s13321-020-00456-1.

Wassim Bouaziz, Nicolas Usunier, and El Mahdi El Mhamdi. Data taggants: Dataset ownership verification via harmless targeted data poisoning. In *International Conference on Learning Representations*, 2025.

Andrei Z. Broder. On the resemblance and containment of documents. In *Compression and Complexity of Sequences*, pp. 21–29, 1997.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, pp. 316–330, 2001. doi:10.1007/3-540-44503-X_20.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650, 2021.

ChEMBL Team. ChEMBL database release 36. EMBL–EBI, 2025. doi:10.6019/CHEMBL.database.36.

James Cheney, Laura Chiticariu, and Wang-Chiew Tan. Provenance in databases: Why, how, and where. *Foundations and Trends in Databases*, 1(4):379–474, 2009. doi:10.1561/1900000006.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of EMNLP*, pp. 1724–1734, 2014. doi:10.3115/v1/D14-1179.

Markus Dablander, Thierry Hanser, Renaud Lambiotte, and Garrett M. Morris. Sort & Slice: A simple and superior alternative to hash-based folding for extended-connectivity fingerprints. *Journal of Cheminformatics*, 16:135, 2024. doi:10.1186/s13321-024-00932-y.

Jesus de la Fuente, Guillermo Serrano, Uxía Veleiro, Mikel Casals, Laura Vera, Marija Pizurica, Nuria Gómez-Cebrián, Leonor Puchades-Carrasco, Antonio Pineda-Lucena, Idoia Ochoa, Silve Vicent, Olivier Gevaert, and Mikel Hernaez. Towards a more inductive world for drug repurposing approaches. *Nature Machine Intelligence*, 7:495–508, 2025. doi:10.1038/s42256-025-00987-y.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. In *Proceedings of EMNLP*, pp. 1286–1305, 2021. doi:10.18653/v1/2021.emnlp-main.98.

Raúl Fernández-Díaz, Thanh Lam Hoang, Vanessa Lopez, and Denis C. Shields. A new framework for evaluating model out-of-distribution generalisation for the biochemical domain. In *International Conference on Learning Representations*, 2025.

Denis Fourches, Eugene Muratov, and Alexander Tropsha. Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of Chemical Information and Modeling*, 50(7):1189–1204, 2010. doi:10.1021/ci100176x.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi:10.1145/3458723.

David Graber, Peter Stockinger, Fabian Meyer, Siddhartha Mishra, Claus Horn, and Rebecca Buller. Resolving data bias improves generalization in binding affinity prediction. *Nature Machine Intelligence*, 7:1713–1725, 2025. doi:10.1038/s42256-025-01124-5.

Volker D. Hähnke, Sunghwan Kim, and Evan E. Bolton. PubChem chemical structure standardization. *Journal of Cheminformatics*, 10:36, 2018. doi:10.1186/s13321-018-0293-8.

Stephen R. Heller, Alan McNaught, Igor Pletnev, Stephen Stein, and Dmitrii Tchekhovskoi. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics*, 7:23, 2015. doi:10.1186/s13321-015-0068-4.

Zirui Hu, Yingjie Wang, Zheng Zhang, Hong Chen, and Dacheng Tao. A statistical approach for controlled training data detection. In *International Conference on Learning Representations*, 2025.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics Data Commons: Machine learning datasets and tasks for drug discovery and development. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Roman Joeres, David B. Blumenthal, and Olga V. Kalinina. Data splitting to avoid information leakage with DataSAIL. *Nature Communications*, 16:3337, 2025. doi:10.1038/s41467-025-58606-8.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 10697–10707, 2022.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In *Proceedings of ACL*, pp. 8424–8445, 2022. doi:10.18653/v1/2022.acl-long.577.

Hengyu Luo, Raúl Vázquez, Timothee Mickus, Filip Ginter, and Jörg Tiedemann. Knowledge distillation as decontamination? Revisiting the "data laundering" concern in classification tasks. In *International Conference on Learning Representations*, 2026.

Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9:5441–5451, 2018. doi:10.1039/C8SC00148K.

David Mendez, Anna Gaulton, A. Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F. Mosquera, Prudence Mutowo, Michal Nowotka, et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 2019. doi:10.1093/nar/gky1075.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229, 2019. doi:10.1145/3287560.3287596.

Luc Moreau and Paolo Missier (editors). PROV-DM: The PROV data model. W3C Recommendation, 30 April 2013. `https://www.w3.org/TR/prov-dm/`.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *NeurIPS Datasets and Benchmarks Track*, 2021.

Jaden Park, Mu Cai, Feng Yao, Jingbo Shang, Soochahn Lee, and Yong Jae Lee. Contamination detection for VLMs using multi-modal semantic perturbations. In *International Conference on Learning Representations*, 2026.

RDKit contributors. RDKit: Open-source cheminformatics, release 2026.03.5. `https://www.rdkit.org`, 2026.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi:10.1021/ci100050t.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4:1256–1264, 2022. doi:10.1038/s42256-022-00580-7.

Roger A. Sayle. So you think you understand tautomerism? *Journal of Computer-Aided Molecular Design*, 24:485–496, 2010. doi:10.1007/s10822-010-9329-5.

Maximilian G. Schuh, Aleksandra Daniluk, and Stephan A. Sieber. Auditing widely used biomolecular benchmarks reveals systematic data inconsistencies. *Chemical Science*, 2026. doi:10.1039/D6SC01799A.

Robert P. Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–790, 2013. doi:10.1021/ci400084k.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pp. 3–18, 2017.

Simon Steshin. Lo-Hi: Practical ML drug discovery benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2023.

Joseph Szymborski and Amin Emad. A flaw in using pretrained protein language models in protein–protein interaction inference models. *Nature Machine Intelligence*, 8:197–208, 2026. doi:10.1038/s42256-025-01176-7.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Oya Wahl and Thomas Sander. Tautobase: An open tautomer database. *Journal of Chemical Information and Modeling*, 60(3):1085–1089, 2020. doi:10.1021/acs.jcim.0c00035.

Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of Chemical Information and Modeling*, 58(5):916–932, 2018. doi:10.1021/acs.jcim.7b00403.

Han Wang, Haoyu Li, Brian Ko, and Huan Zhang. On the fragility of benchmark contamination detection in reasoning models. In *International Conference on Learning Representations*, 2026.

David Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi:10.1021/ci00057a005.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi:10.1039/C7SC02664A.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, 2019. doi:10.1021/acs.jcim.9b00237.

Fei Ye, Zaixiang Zheng, Dongyu Xue, Yuning Shen, Lihao Wang, Yiming Ma, Yan Wang, Xinyou Wang, Xiangxin Zhou, and Quanquan Gu. ProteinBench: A holistic evaluation of protein foundation models. In *International Conference on Learning Representations*, 2025.

Michaël Zasso, Luc Patiny, Thomas Sander, and Christian Rufener. OpenChemLib JS: Library to manipulate chemical structures and reactions in JavaScript. Zenodo, 2021. doi:10.5281/zenodo.5139988.

Michał Zawalski, Meriem Boubdir, Klaudia Bałazy, Besmira Nushi, and Pablo Ribalta. Detecting data contamination in LLMs via in-context learning. In *International Conference on Learning Representations*, 2026.

Yikun Zhang, Geyan Ye, Chaohao Yuan, Bo Han, Long-Kai Huang, Jianhua Yao, Wei Liu, and Yu Rong. Atomas: Hierarchical adaptive alignment on molecule-text for unified molecule understanding and generation. In *International Conference on Learning Representations*, 2025.

Zhihao Zhu, Yi Yang, and Defu Lian. TDDBench: A benchmark for training data detection. In *International Conference on Learning Representations*, 2025.

A Relation Definitions and Normalization

A.1 Positive relation examples

Figure A.1 illustrates the four positive relation types used by the audit.

A.2 Operational keys, controls, and baselines

Table A.1 states the implemented key and allowed interpretation for each route.

RDKit performs default sanitization with `Chem.MolFromSmiles`. Parent processing selects the largest fragment with `LargestFragmentChooser(preferOrganic=False).choose` and removes charge with `Uncharger().uncharge`. Stereo processing calls `Chem.RemoveStereochemistry`; tautomer processing calls `TautomerEnumerator().Canonicalize`; and every route ends with `MolToSmiles`. Largest-fragment ties follow RDKit's atom-count (including hydrogens), molecular-weight, and lexicographic-SMILES ordering. Here, parent denotes shared standardized-parent identity under this operation, not historical or registry parent–child lineage.

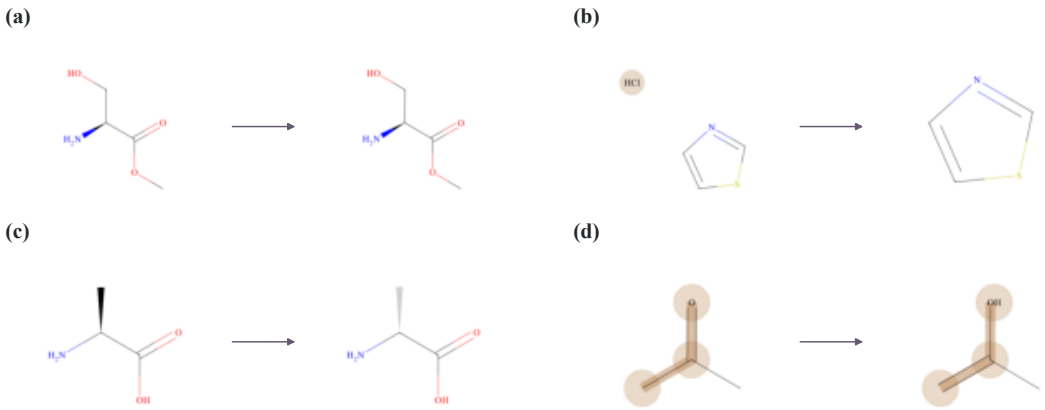

Figure A.1: Representative positive relation pairs from the external registry and curated benchmark: (a) exact identity, (b) parent, (c) stereochemistry, and (d) tautomer. Molecular coordinates are aligned within each pair so that local parent, stereochemical, or tautomeric changes remain visually attributable.

Table A.1: Operational keys, controls, and baseline representations under the fixed RDKit policy.

Route	Implemented key	Allowed interpretation
Raw-string equality	Unmodified string	Identical serialization
Canonical identity	Canonical isomeric SMILES	Toolkit-canonical record identity
Randomized control	Randomized isomeric SMILES + canonical check	Serialization-exposure control
Parent	Largest fragment + uncharge + canonicalize	Standardized-parent relation
Stereo	Remove stereochemistry + nonisomeric canonicalize	Stereo-stripped connectivity
Tautomer	Canonical tautomer + isomeric canonicalize	RDKit tautomer class
Scaffold baseline	Nonisomeric Bemis–Murcko scaffold	Scaffold neighborhood

A.3 Deterministic audit procedure

Using the relation family $\mathcal{R}$ and keys k_ρ defined in Sections 3.1–3.2, the implementation executes the following fixed procedure on benchmark and candidate-corpus bytes.

1. Parse and sanitize each record; an initial parse failure or route-specific exception produces an empty key for that route and a logged abstention, never a key substituted from another relation.
2. Compute every nonempty k_ρ for $\rho \in \mathcal{R}$ and accumulate relation-specific key multiplicities over the frozen corpus order.
3. Compute benchmark keys and query only the corresponding hash index; no similarity fallback is allowed after an abstention or empty lookup.
4. Retain all relation flags in pair-level audit tables and preserve controlled source positions in the frozen manifests; precedence selects only a displayed primary relation.
5. Emit a provenance certificate containing code-version and schema identities, split, corpus, and manifest hashes, record counts, typed retrieval summaries, and the recorded seed-level effect field.

The invariants are deterministic output for identical bytes and code identity, no positive relation without equal nonempty typed keys, reconstructibility of record-level edges from bound corpus bytes, and fixed failure behavior that leaves the relation definition unchanged.

After replacing cross-route fallbacks with abstention, a fixed-input rerun scanned 82,715 unique valid strings from the evaluated corpora, external relation pairs, registry negatives, and hard-neighborhood

bank. All four routes produced nonempty keys for every string, so the reported retrieval and validation values are unchanged.

A.4 Normalization stability

Table A.2 reports path agreement and idempotence checks under the fixed normalization policy.

The nonzero parent, stereo, and tautomer rates show that these keys are software-version-bound operational definitions; this motivates binding certificates to the reported RDKit snapshot. The audit applies each route once to the stored representation. Non-idempotence therefore leaves a fixed-snapshot audit reproducible but precludes invariance to arbitrary repeated upstream standardization; for an already standardized corpus, its stored record remains the auditable source representation.

A.5 Collapsed-standardization reference

RDKit SuperParent composes fragment, charge, isotope, stereochemistry, and tautomer normalization into one standardized structure. We evaluate equality of that structure as a collapsed-standardization reference, not as a typed-provenance method. Table A.3 reports both the fixed-pair audit and the same clean-background/decoy protocol as Table 2; unlike ChemTrace, SuperParent cannot attribute a positive to one declared transformation route.

Table A.3: Relation-preserving provenance versus collapsed standardization. External recall uses 6,675 independently sourced non-exact pairs. Common-protocol entries are five-seed means over the same 81-cell common-protocol grid used in Table 2, with 78 cells containing changed queries. The hard-bank call rate uses 3,948 typed-negative neighbors.

Method	External recall	Rec_Δ	b	q	S	Hard-bank call	Typed
ChemTrace	0.9654	1.0000	0.0036	0.0000	0.9963	0.0000	Yes
RDKit SuperParent	0.9784	1.0000	0.0132	0.0000	0.9863	0.0215	No

SuperParent slightly increases collapsed external recall and matches controlled changed recall, but its common-protocol background rate is 0.0132 ± 0.0100 versus 0.0036 ± 0.0032 for ChemTrace, yielding $S = 0.9863 \pm 0.0104$ versus 0.9963 ± 0.0033. It also calls 85 of 3,948 conditioned hard negatives and cannot identify which normalization route supports a match. Thus aggressive collapsed standardization recovers coverage, whereas ChemTrace retains route attribution with lower incidental retrieval.

A.6 Tautomer miss diagnostics

We classify the first 100 miss records in fixed table order using a deterministic rule sequence: formula/protonation, charge, scaffold/connectivity, aromaticity, then RDKit coverage/ambiguity. Figure A.2 summarizes the resulting diagnostic categories.

B External Positive Construction

B.1 Source-specific construction

Parent positives come from ChEMBL36 registry hierarchy edges, stereo positives from ChEMBL36 Standard InChI records that differ only in explicit stereo layers, and tautomer positives from curated Tautobase pairs. Table B.1 reports the resulting count-level audit; the exact row is a registry self-identity sanity check.

B.2 Gold filtering, exclusions, and negative controls

ChEMBL positives were queried from the release-36 SQLite snapshot (ChEMBL Team, 2025). Parent pairs require distinct child and parent identifiers in `molecule_hierarchy`, nonempty canonical SMILES on both sides, and unequal RDKit canonical keys. Stereo selection is independent of the

Table A.2: Normalization stability across 50,000 molecules under the fixed RDKit policy.

Check	Rate (%)
Parent–tautomer path disagreement	0.150
Stereo–parent path disagreement	0.022
Canonical non-idempotence	0.000
Parent non-idempotence	0.014
Stereo non-idempotence	0.008
Tautomer non-idempotence	0.242

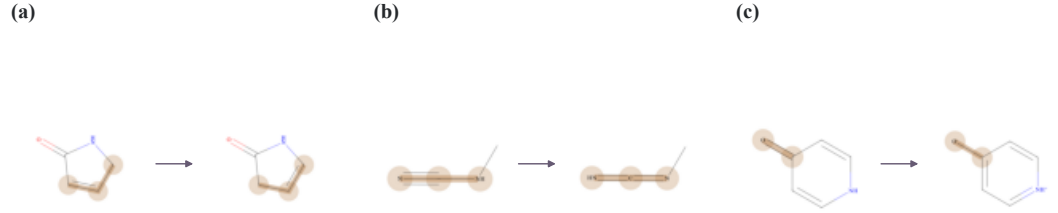

Figure A.2: First 100 recorded tautomer misses grouped by rule-based diagnostics: (a) scaffold/connectivity shift, (b) RDKit coverage/ambiguity, and (c) charge/protonation interaction, with counts 73, 20, and 7.

Table B.1: Count-level external-relation audit corresponding to Table 1.

Relation	External construction	Support	TP	FP	FN
Exact	ChEMBL36 standard InChIKey self-identity sanity set	1,000	1,000	0	0
Parent	ChEMBL36 molecule hierarchy	3,000	2,976	5	24
Stereo	ChEMBL36 Standard InChI stereo layers	2,000	1,923	4	77
Tautomer	Tautobase curated records converted by OpenChemLib	1,675	1,530	6	145

evaluated ChemTrace key: pairs must have the same stored Standard InChI after removing `/b`, `/t`, `/m`, and `/s` layers, different nonempty stereo-layer signatures, different full Standard InChIKeys, and different canonical isomeric SMILES. RDKit stereochemistry removal is applied only after selection to measure recall. ChEMBL generally deduplicates exact structures, so the exact row is a self-identity registry sanity set; parent, stereo, and tautomer rows supply the independent non-exact relation evidence.

The parent query is ordered by child and parent registry identifiers, and the first 3,000 qualifying pairs are retained. Stereo candidates are grouped in lexicographic order of the Standard InChI with stereo layers removed; within each group they are ordered by stereo signature and registry identifier, and the first 2,000 qualifying adjacent pairs are retained. These deterministic limits define the released fixed pair table.

The Tautobase release contains 1,680 unique pairs (Wahl and Sander, 2020). We decoded its DataWarrior idcodes with OpenChemLib JS 9.25.0 (Zasso et al., 2021), discarded conversion failures and empty molecules, required both sides to parse under RDKit, and removed pairs already equal by canonical identity. This left 1,675 pairs. The raw `Tautobase.dwar` SHA-256 digest is `8fabfd7123b55d30448dc6a3b20ede4558097fcae81c8fdfa0b3884d3cb58b68`. The final relation decision retains all positive flags; the order exact, parent, stereo, then tautomer is used only to select a concise display label.

Accordingly, the tautomer key denotes the equivalence class induced by the fixed RDKit `TautomerEnumerator` policy, not every chemically curated tautomerism convention. Tautobase includes curated pairs outside that operational class, especially ring–chain or connectivity-changing equilibria; these are validation misses rather than evidence that the stored pair is chemically unrelated.

For the binary certificate sanity check, 1,000 none-of-the-above pairs were reserved from the fixed hard-negative construction and accepted as conditioned negatives only when all four typed rules were false. This algorithmic set complements a distinct registry-negative set: 500 ChEMBL36 pairs were fixed using different registry identifiers, Standard InChI formulae and connectivity blocks, with at least a ten-heavy-atom formula difference, before blind ChemTrace evaluation. All 500 remain certificate-negative. The external-gold pairs have SHA-256 `97e14faff27e4a3476d331052b`

81fb82947c09e130c02644d6a667b5dfac4721; the registry-negative pairs have SHA-256 a3bc508c2fb080db7df39c34e2f1ac8065911b1df2737bb16e86d62a59dff1a3. These independent but deliberately separated registry negatives provide a construction check, not a hard-neighborhood sample. Upstream ChEMBL and Tautobase terms govern source records; audit tables contain identifiers, relation labels, and derived keys subject to those terms.

Relation precision and recall in Tables 1 and B.1 are one-vs-rest metrics over the precedence-assigned primary relations. The pair-level table also retains each route's Boolean flag: for example, the stereo key matches 1,927 of 2,000 stereo positives, including four displayed as parent. The five parent, four stereo, and six tautomer primary-relation confusions are cross-positive relation confusions that remain positive typed certificates.

B.3 Leave-one-route-out coverage

Table B.2 tests whether the externally sourced non-exact positive coverage is attributable to a single route.

Table B.2: Leave-one-route-out recall on 6,675 independently sourced parent, stereo, and tautomer positives.

Retained certificate union	Detected	Recall
All four routes	6,444 / 6,675	0.9654
Without exact	6,444 / 6,675	0.9654
Without parent	3,467 / 6,675	0.5194
Without stereo	4,784 / 6,675	0.7167
Without tautomer	4,908 / 6,675	0.7353

C Hard-Negative Construction

C.1 Bank construction and similarity severity

This bank is a conditional semantic stress test and does not estimate an independently sampled false-positive rate. The construction draws queries from a fixed pool of at most 6,000 records and candidates from fixed clean-corpus (15,000), task (20,000), and external ChEMBL/ZINC (200,000) pools. Every retained pair is negative under all four typed relations. The seven family labels use the following sampling rules: same-formula constitutional isomers share a formula but differ in InChIKey connectivity; same-scaffold pairs share a Bemis–Murcko scaffold; property-matched scaffold pairs additionally satisfy $|\Delta \mathrm{MW}| \leq 25$ Da and $|\Delta \log P| \leq 1.0$; the two ECFP families require radius-2, 2,048-bit Morgan Tanimoto of at least 0.80 or 0.90; nearest neighbors are the highest-similarity candidate in the fixed benchmark, clean, or external pool without a threshold; and activity cliffs have opposite labels on the same task plus either Tanimoto at least 0.70 or scaffold equality. MW and logP are computed with RDKit 2026.03.5 using `Descriptors.MolWt` and `Crippen.MolLogP`, respectively. Each label is capped at 600 pairs; the realized counts in that order are 600, 600, 600, 600, 513, 600, and 435 (3,948 total). Families may overlap; each label records the rule used to sample that pair. Figure C.1 shows the resulting similarity distributions, and Table C.1 reports route-level operating characteristics on the fixed bank.

The 432 positive controls come from the smallest-dose ($1\times$) parent, stereo, and tautomer manifests of the primary controlled setting. We deduplicate by sample identifier and canonical target and exclude canonical-identity matches, leaving 151 parent, 173 stereo, and 108 tautomer controls.

For the Morgan Tanimoto row, a binary call in Table C.1 requires a score of at least 1.0 (equivalently, Tanimoto $= 1.00$); the ECFP≥ 0.90 row uses its stated cutoff. Bank precision treats the 432 typed controls as positives and the 3,948 hard neighbors as negatives.

C.2 Molecular examples

Figure C.2 shows representative pairs from the hard-negative families.

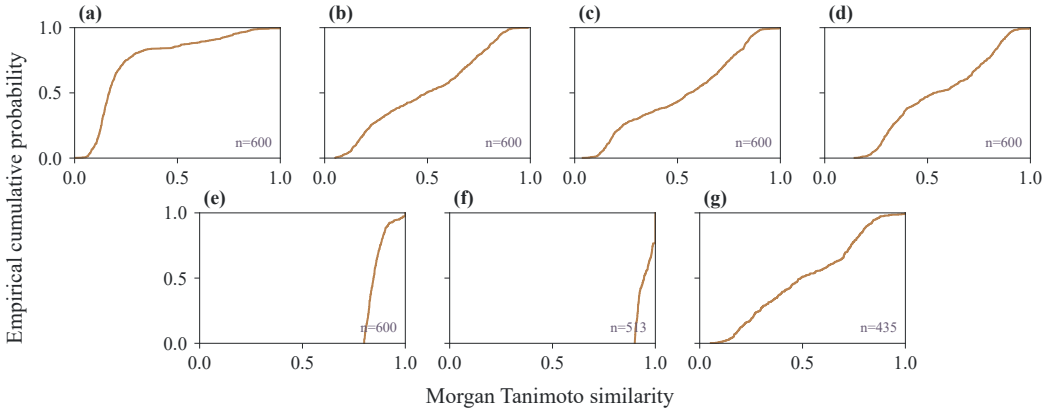

Figure C.1: Fingerprint-similarity distributions for seven hard-negative families: (a) same-formula isomer, (b) same scaffold, (c) property-matched scaffold, (d) nearest neighbor, (e) ECFP$\geq$ 0.80, (f) ECFP$\geq$ 0.90, and (g) activity cliff.

Table C.1: Operating characteristics on the fixed bank of 432 typed positive controls and 3,948 typed-negative hard chemical neighbors.

Method	Group	Positive-control recall (%)	Conditional bank precision (%)	Negative-pair call rate (%)
CHEMTRACE	Proposed	**100.00**	**100.00**[†]	**0.00**
InChIKey connectivity	Direct	91.44	96.34	0.38
Morgan ECFP Tanimoto = 1.00	Similarity	55.56	61.54	3.80
Morgan ECFP $\geq$ 0.90	Similarity	58.10	27.46	16.79
Lo-Hi-style high-similarity split	Split control	93.52	15.56	55.52
Biomolecular-audit bundle	Audit inspired	96.30	11.80	78.75

[†]Bank negatives were conditioned to be negative under all four CHEMTRACE keys. The resulting 100% conditional precision is a construction property, not an independently estimated PPV or false-positive rate. Independent negative evidence is supplied by the separately constructed ChEMBL36 registry negatives and the common-protocol background/decoy analyses.

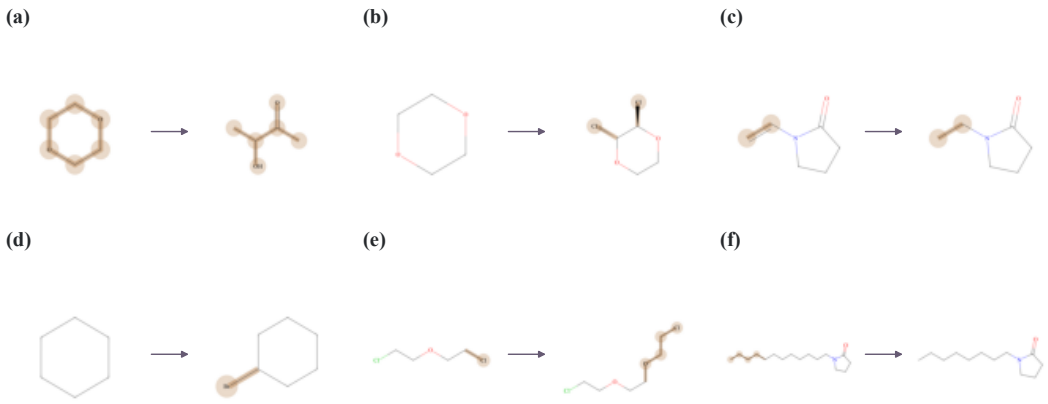

Figure C.2: Representative hard negatives: (a) constitutional isomer, (b) same Bemis–Murcko scaffold, (c) property-matched scaffold, (d) nearest neighbor, (e) ECFP Tanimoto at least 0.90, and (f) same-task activity cliff.

D Full Experimental Protocol

D.1 Evaluation inventory

Table D.1 summarizes the datasets, exposure conditions, audit routes, and repetitions used in the full evaluation.

Table D.1: Evaluation inventory for the common-protocol audit and controlled-effect analyses.

Scope item	Value	Count
Public benchmarks	BACE, BBBP, HIV, ClinTox (toxicity/FDA), Tox21 (AhR/ARE/MMP), and SIDER (gastrointestinal)	6 datasets / 9 endpoints
Primary controlled corpus	Clean, injected, and matched-decoy variants	50,000 records
Exposure conditions	Exact, randomized, parent, stereo, tautomer; matched decoy	5 routes + control
Injection doses	Per-affected-record multiplicity	$1\times$, $5\times$, $20\times$
Model families	Transformer; BiGRU and GIN stress checks	3
Common-protocol comparison	Identity, similarity, and adapted split/curation routes	13 routes
Repetition	Background seeds 101, 103, 107, 109, 113	5 seeds
Positive-relation evaluation	ChEMBL36 and Tautobase pairs	7,675 pairs
Independent negative check	ChEMBL36 registry pairs	500 pairs
Conditional stress bank	432 typed positives; 3,948 typed-negative neighbors	4,380 pairs
Downstream primary estimator	Label-adjusted difference-in-differences (DiD)	51 cells

D.2 Datasets, splits, and preprocessing

All nine endpoints are binary classification tasks from fixed DeepChem-format MoleculeNet mirrors. They cover BACE inhibition, blood–brain-barrier penetration (BBBP), HIV activity, ClinTox toxicity and FDA-approval labels, Tox21 aryl hydrocarbon receptor (AhR), antioxidant response element (ARE), and mitochondrial membrane potential (MMP) endpoints, and the SIDER gastrointestinal adverse-event endpoint. The selected target columns are BACE `Class`, BBBP `p_np`, HIV `HIV_active`, ClinTox `CT_TOX` and `FDA_APPROVED`, Tox21 `NR-AhR`, `SR-ARE`, and `SR-MMP`, and SIDER `Gastrointestinal disorders`. We materialize each endpoint separately, remove rows whose selected endpoint label is missing, parse and sanitize the molecular record, and then construct a scaffold partition using preparation seed 20260824. Table D.2 reports the post-filter denominators consumed by training. The concatenated split table has SHA-256 `9edb0c7151a7d89dcade72cb12703a76e20e34499d2aa21ed689d3681edc9391`, and the affected-record table has SHA-256 `3f08169adeef4ffdf241cfa65cf47c794e5f3d53830fa1bb6358221d19ca2798`. Every affected subset contains 64 label-stratified test records with canonical-SMILES character length at most 158. The intervention analysis uses per-sample binary cross-entropy to construct affected-versus-control loss contrasts.

After RDKit parsing and canonicalization, canonical-structure deduplication, and exclusion of every task test-set canonical identity, the clean-corpus source pool contained 38,322 unique structures from fixed MoleculeNet mirrors (HIV, Tox21, BACE, BBBP, and ClinTox) (Wu et al., 2018) and the five cleaned data files distributed with the Sort & Slice repository (Dablander et al., 2024) at commit `ec080049710e`: Lit-PCBA ESR-ant, PostEra SARS-CoV-2 Mpro, Ames mutagenicity, AqSolDB, and MoleculeNet lipophilicity. The pool was ordered by a stable hash with seed 20260824; seeded randomized-SMILES variants filled the remaining 11,678 record positions to fix a 50,000-record corpus.

For every affected query and repetition, the decoy allocator searches this fixed clean pool by absolute canonical-SMILES character-length difference, checking the shorter length before the longer length on a tie. A stable hash of sample identifier, repetition, candidate length, and seed selects among candidates; candidates remain available, so sampling is with replacement. Dose d inserts d indepen-

dently allocated decoy records. No exact, parent, stereo, tautomer, scaffold, or fingerprint-similarity exclusion is applied. Background positions are selected without replacement by the same nearest-length policy. Thus *length-matched background decoy* is the operational control, and q records any incidental typed match.

Table D.2: Fixed post-filter endpoint partitions and affected-set sizes used by the primary controlled setting.

Endpoint	Train	Validation	Test	Affected	Endpoint digest (first 12 hex)
BACE	1,211	151	151	64	`85e75720b7d7`
BBBP	1,575	260	204	64	`d082c88ab86b`
ClinTox toxicity	1,228	102	149	64	`c638b61320be`
ClinTox FDA approval	1,228	102	149	64	`8a1754eda64f`
HIV	32,460	4,548	4,112	64	`459a7535d8f1`
SIDER gastrointestinal	1,189	95	143	64	`b61d73e88f4c`
Tox21 AhR	5,703	184	655	64	`31f567070624`
Tox21 ARE	5,063	178	584	64	`ff0978d8dc19`
Tox21 MMP	5,040	182	582	64	`efd20a6ea3a4`

D.3 Model and optimization details

The lightweight controlled model measures the corpus intervention under a fixed architecture and training recipe. The primary encoder consumes character-level SMILES with a 70-token vocabulary and a maximum length of 160. It has two Transformer encoder layers (Vaswani et al., 2017), hidden dimension 160, four attention heads, a 640-dimensional feed-forward block, GELU activations, and dropout 0.1. The robustness checks use a bidirectional gated recurrent unit (BiGRU) sequence encoder (Cho et al., 2014) and a graph isomorphism network (GIN) (Xu et al., 2019). The Transformer encoder has 655,680 parameters; the masked-language-model head adds 11,270 parameters and the binary downstream head brings the fine-tuned model to approximately 681,601 parameters. Table D.3 records the complete optimization and execution settings.

Table D.3: Optimization, robustness-model, software, and hardware settings.

Component	Fixed setting
Masked pretraining	15% mask; 4 epochs; bs=256; 50,000 records/condition
Fine-tuning	$\leq$ 80 epochs; bs=128; patience=12; primary train fraction=0.10
Optimization	AdamW; lr=3×10^{-4}; wd=0.01; constant schedule; clip=1.0
Architecture checks	BiGRU character encoder; 5-layer GIN; dim=256; dropout=0.1; masked-atom pretraining
Repetition	Detector: 1,000-record background samples over 5 seeds; downstream: 4/12/24 paired seeds (Table F.2)
Software snapshot	Python 3.10.12; PyTorch 2.13.0+cu130; RDKit 2026.03.5; NumPy 2.2.6; pandas 2.3.3; SciPy 1.15.3; scikit-learn 1.7.2
Execution machine	2×AMD EPYC 9654; 384 threads; 8×RTX 4090 D 24 GB; one GPU/job

The primary Transformer experiments use a 10% deterministic subset of each training partition; low-resource BBBP checks use fractions 0.25, 0.05, and 0.02. The BiGRU and GIN checks use 12 seeds (307, 331, 353, 379, 401, 421, 443, 467, 491, 509, 541, and 563), while one low-resource Transformer setting was evaluated with 24 paired seeds. Pretraining and fine-tuning reuse the same seed within each clean, injected, and decoy comparison. Early stopping selects the minimum validation loss without consulting affected test outcomes.

E Special-Scenario and Seed-Level Results

E.1 Special scenarios

Table E.1 compares ChemTrace with the strongest per-scenario comparator using the same S definition, with b and q computed within each scenario.

Table E.1: Five-seed CHEMTRACE performance and strongest comparator in special audit scenarios.

Scenario	CHEMTRACE S	Strongest comparator	Comparator S
All non-exact changed relations	0.9963 ± 0.0033	Lo-Hi-style split	0.8783 ± 0.0155
Cold-start scaffold absent	0.9975 ± 0.0026	Biomolecular-audit bundle	0.9149 ± 0.0131
High background similarity (≥ 0.70)	0.8203 ± 0.2435	InChIKey connectivity	0.4826 ± 0.1073
Long-tail scaffold count ≤ 1	0.9965 ± 0.0033	Lo-Hi-style split	0.8834 ± 0.0217
OOD low similarity (< 0.70)	0.9978 ± 0.0025	Lo-Hi-style split	0.8994 ± 0.0108
Warm-start scaffold present	0.9795 ± 0.0182	InChIKey connectivity	0.8572 ± 0.0777

E.2 PER-REPEAT ALL-ROUTE ORACLE COMPARISON AND STABILITY

For the all-route comparator oracle defined in Section 5.3, five paired repeats give 0.9963 ± 0.0033 for CHEMTRACE versus 0.9364 ± 0.0064 for the oracle, a paired gain of 0.0598 ± 0.0040 with 95% t interval [0.0549, 0.0648]. Figure E.1 shows all five paired repeats.

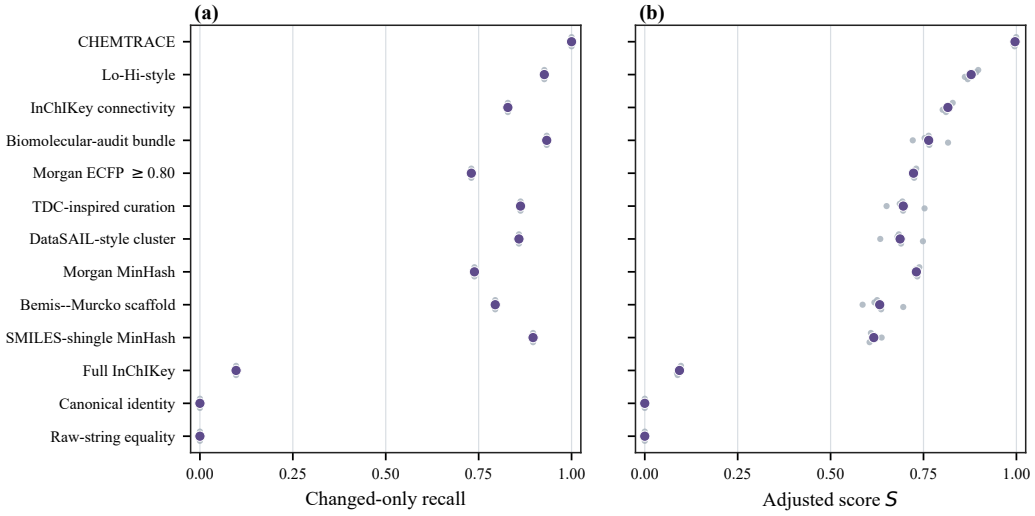

Figure E.1: Per-seed stability of the common-protocol provenance comparison. Panel (a) shows changed-only recall, and panel (b) shows the background- and decoy-adjusted score S. Individual repeat points expose background-sampling variability that is compressed by the mean and standard deviation in Table 2.

F DOWNSTREAM EFFECT ROBUSTNESS

F.1 ESTIMATOR ROBUSTNESS, GLOBAL AGGREGATION, AND PRACTICAL-EQUIVALENCE DIAGNOSTICS

Across all 51 primary-estimator cells, every two-stage scaffold/seed bootstrap interval crosses zero. The reported primary 95% intervals are percentile-bootstrap intervals formed from the 2.5th and 97.5th empirical quantiles of the 4,000 draws. Three seed-resampled adjusted-linear diagnostic intervals are positive and three are negative. The label-adjusted estimator is the prespecified primary analysis; adjusted-linear results are reported as sensitivity checks.

Practical-equivalence analysis uses two descriptive margins, $\epsilon = 0.01$ and 0.02, and classifies a cell as equivalent when its complete primary 95% interval lies within $[-\epsilon, \epsilon]$. At $\epsilon = 0.01$, two of 51 cells meet this criterion (one randomized and one stereo); at $\epsilon = 0.02$, three do (one exact, one randomized, and one stereo). This leaves 49 and 48 cells for which practical equivalence is not established at the corresponding margins; none has an interval wholly beyond either margin.

Table F.1 adds endpoint-clustered global summaries. Each draw resamples endpoint clusters and uses the paired-seed/scaffold draw for every included cell; the equal-cell mean is descriptive because the robustness grid is intentionally unbalanced. The primary-protocol changed-only analysis retains between 5 and 41 of the 64 affected records per non-exact cell. All nine of its cell-level intervals cross zero, as do all 18 changed-only intervals across the full robustness grid. Thus the fixed-affected-set result is not an artifact of including transformation-invariant affected records.

Table F.1: Global and canonical-changing downstream sensitivities. Intervals use 4,000 endpoint-clustered draws with the paired seed/scaffold bootstrap nested within cells.

Analysis	Cells	Endpoints	Affected population	Global mean [95% CI]
All reported cells	51	9	Fixed affected set	-0.0004 [-0.0140, 0.0089]
Primary protocol	23	9	Fixed affected set	-0.0045 [-0.0230, 0.0115]
Primary non-exact sensitivity	9	6	$A \cap Q_\rho^\Delta$	-0.0353 [-0.1151, 0.0312]
All non-exact sensitivities	18	6	$A \cap Q_\rho^\Delta$	-0.0128 [-0.0768, 0.0227]
Parent sensitivities	6	4	$A \cap Q_\rho^\Delta$	-0.0317 [-0.1531, 0.0429]
Stereo sensitivities	11	4	$A \cap Q_\rho^\Delta$	-0.0088 [-0.0711, 0.0206]
Tautomer sensitivity	1	1	$A \cap Q_\rho^\Delta$	0.0569 [-0.0103, 0.1947]

A centered-bootstrap dispersion diagnostic compares the observed variance of the 51 cell point estimates (2.78×10^{-4}) with the variance induced after centering every cell's bootstrap distribution at a common effect. The centered null variance averages 6.97×10^{-4}, and the one-sided bootstrap probability of a null variance at least as large as observed is 0.9985. Cross-cell spread is therefore not distinguishable from the reported sampling uncertainty and is not interpreted as evidence of effect heterogeneity.

Tables F.2 and F.3 report every estimate and its seed count, separating primary-protocol cells from targeted stress-test settings. The Setting column states the architecture, affected-set variant, or training fraction directly. Point estimates and adjusted-linear diagnostics derive from the 51-cell result table; primary intervals were recomputed from the same 772 per-seed/condition prediction outputs using the 4,000-draw scaffold/seed analysis.

Table F.2: Primary-protocol Transformer effect cells. Primary intervals use the 4,000-draw scaffold/seed bootstrap; adjusted-linear diagnostics use paired-seed resampling.

Setting	Task	Relation/dose	Seeds	Primary label-adjusted DiD effect	Adjusted-linear diagnostic (seed)
Transformer	BACE	randomized/1	4	0.0263 [-0.0628, 0.1120]	0.0198 [-0.0442, 0.0803]
Transformer	BBBP	randomized/20	12	0.0144 [-0.0130, 0.0412]	0.0148 [0.0001, 0.0290]
Transformer	Tox21 MMP	parent/1	12	0.0142 [-0.0432, 0.0726]	0.0221 [-0.0050, 0.0477]
Transformer	Tox21 AhR	randomized/1	12	0.0125 [-0.0343, 0.0588]	0.0089 [-0.0116, 0.0256]
Transformer	HIV	exact/20	12	0.0117 [-0.0260, 0.0522]	0.0176 [-0.0132, 0.0501]
Transformer	Tox21 ARE	exact/20	12	0.0098 [-0.0286, 0.0460]	0.0072 [-0.0156, 0.0286]
Transformer	Tox21 AhR	exact/20	12	0.0091 [-0.0674, 0.0738]	0.0113 [-0.0344, 0.0516]
Transformer	SIDER gastro.	exact/1	12	0.0045 [-0.0156, 0.0294]	0.0058 [-0.0065, 0.0173]
Transformer	ClinTox tox.	exact/1	4	0.0021 [-0.0132, 0.0179]	0.0010 [-0.0044, 0.0056]
Transformer	HIV	randomized/20	4	0.0014 [-0.0719, 0.0536]	0.0051 [-0.0228, 0.0300]
Transformer	Tox21 MMP	exact/1	12	0.0004 [-0.0439, 0.0448]	0.0008 [-0.0179, 0.0182]
Transformer	HIV	parent/5	4	-0.0052 [-0.0675, 0.0615]	-0.0087 [-0.0559, 0.0460]
Transformer	BBBP	stereo/5	12	-0.0053 [-0.0290, 0.0187]	-0.0019 [-0.0183, 0.0168]
Transformer	ClinTox FDA	stereo/1	4	-0.0056 [-0.0216, 0.0073]	-0.0037 [-0.0076, -0.0002]
Transformer	BBBP	parent/5	4	-0.0068 [-0.0470, 0.0368]	-0.0053 [-0.0313, 0.0206]
Transformer	SIDER gastro.	randomized/1	12	-0.0095 [-0.0404, 0.0124]	-0.0124 [-0.0372, 0.0039]
Transformer	SIDER gastro.	tautomer/5	4	-0.0110 [-0.0595, 0.0291]	-0.0080 [-0.0408, 0.0148]
Transformer	Tox21 MMP	randomized/5	4	-0.0114 [-0.0888, 0.0437]	0.0047 [-0.0039, 0.0132]
Transformer	Tox21 ARE	randomized/1	4	-0.0158 [-0.0896, 0.0306]	-0.0010 [-0.0172, 0.0152]
Transformer	Tox21 ARE	stereo/5	4	-0.0214 [-0.1196, 0.0685]	-0.0353 [-0.0943, -0.0020]
Transformer	BACE	exact/20	4	-0.0330 [-0.1058, 0.0086]	-0.0018 [-0.0377, 0.0353]
Transformer	Tox21 MMP	stereo/20	4	-0.0402 [-0.1524, 0.0493]	-0.0129 [-0.0515, 0.0309]
Transformer	ClinTox FDA	parent/20	4	-0.0456 [-0.1531, 0.0115]	-0.0366 [-0.1011, 0.0027]

Table F.3: Affected-set, low-resource, and architecture stress cells. Unqualified settings use the Transformer; the Setting column directly names the affected-set, training-fraction, or alternate-architecture variant. Primary intervals use the 4,000-draw scaffold/seed bootstrap; adjusted-linear diagnostics use paired-seed resampling.

Setting	Task	Relation/dose	Seeds	Primary label-adjusted DiD effect	Adjusted-linear diagnostic (seed)
82 affected	BBBP	exact/20	12	0.0500 [−0.0310, 0.1375]	0.0336 [−0.0102, 0.0792]
128 affected	Tox21 MMP	parent/1	12	0.0230 [−0.0115, 0.0576]	0.0188 [0.0009, 0.0370]
BiGRU, 5% train	BBBP	exact/20	12	0.0182 [−0.0049, 0.0674]	0.0073 [−0.0035, 0.0250]
82 affected	BBBP	randomized/20	12	0.0157 [−0.0612, 0.0779]	0.0081 [−0.0237, 0.0386]
128 affected	Tox21 ARE	exact/20	4	0.0155 [−0.0317, 0.0651]	0.0118 [−0.0012, 0.0248]
128 affected	BBBP	randomized/20	4	0.0149 [−0.0107, 0.0474]	0.0134 [−0.0011, 0.0321]
82 affected	BBBP	stereo/20	12	0.0127 [−0.0586, 0.0863]	−0.0063 [−0.0401, 0.0270]
128 affected	HIV	exact/20	4	0.0081 [−0.0555, 0.0635]	−0.0100 [−0.0366, 0.0139]
128 affected	Tox21 AhR	exact/20	4	0.0071 [−0.0350, 0.0583]	0.0027 [−0.0103, 0.0193]
5% train	BBBP	stereo/20	24	0.0061 [−0.0265, 0.0378]	−0.0035 [−0.0215, 0.0140]
5% train	BBBP	exact/20	24	0.0057 [−0.0238, 0.0329]	−0.0023 [−0.0198, 0.0163]
Isolated 1×	Tox21 MMP	parent/1	12	0.0042 [−0.0304, 0.0372]	0.0019 [−0.0124, 0.0173]
5% train	BBBP	randomized/20	12	0.0038 [−0.0342, 0.0433]	0.0059 [−0.0169, 0.0299]
25% train	BBBP	stereo/20	12	0.0035 [−0.0459, 0.0553]	0.0209 [0.0037, 0.0380]
128 affected	Tox21 MMP	exact/20	12	0.0031 [−0.0255, 0.0300]	0.0022 [−0.0177, 0.0202]
GIN, 5% train	BBBP	stereo/20	12	0.0028 [−0.0663, 0.0659]	0.0125 [−0.0020, 0.0250]
128 affected	Tox21 MMP	stereo/20	12	0.0002 [−0.0258, 0.0258]	−0.0010 [−0.0233, 0.0219]
BiGRU, 5% train	BBBP	randomized/20	12	−0.0014 [−0.0084, 0.0055]	−0.0002 [−0.0036, 0.0028]
Isolated 1×	Tox21 MMP	exact/1	12	−0.0025 [−0.0365, 0.0318]	−0.0020 [−0.0175, 0.0156]
128 affected	Tox21 MMP	exact/1	12	−0.0029 [−0.0355, 0.0302]	−0.0033 [−0.0220, 0.0130]
2% train	BBBP	exact/20	12	−0.0039 [−0.0326, 0.0225]	−0.0014 [−0.0145, 0.0097]
BiGRU, 5% train	BBBP	stereo/20	12	−0.0041 [−0.0083, 0.0003]	−0.0025 [−0.0061, 0.0011]
25% train	BBBP	exact/20	12	−0.0083 [−0.0634, 0.0557]	0.0234 [−0.0092, 0.0554]
2% train	BBBP	stereo/20	12	−0.0107 [−0.0413, 0.0201]	−0.0131 [−0.0293, −0.0005]
GIN, 5% train	BBBP	randomized/20	12	−0.0130 [−0.0817, 0.0476]	0.0004 [−0.0178, 0.0166]
128 affected	Tox21 AhR	randomized/1	4	−0.0186 [−0.0566, 0.0123]	−0.0105 [−0.0254, 0.0019]
GIN, 5% train	BBBP	exact/20	12	−0.0194 [−0.0822, 0.0354]	−0.0008 [−0.0207, 0.0193]
25% train	BBBP	randomized/20	12	−0.0274 [−0.0932, 0.0366]	−0.0055 [−0.0397, 0.0219]

Primary-estimator intervals have median half-width 0.0443 (interquartile range (IQR) 0.0299–0.0636) and uneven coverage (19 exact, 14 randomized, six parent, 11 stereo, one tautomer). The three positive and three negative seed-resampled adjusted-linear intervals remain sensitivity results. We report cell-level intervals because task, relation, dose, seed count, and architecture vary across the evaluated cells.

G Real-Corpus Proxy Audit and Scalability

G.1 Measured retrieval scaling

The benchmark concatenates the frozen 50,000-record clean corpus, the available ZINC15 cache, and ChEMBL36 canonical SMILES, removes duplicate strings in insertion order, and takes nested prefixes. Scaling runs request 48 CPU workers on the 384-thread host; Table G.1 reports the corresponding wall-clock measurements.

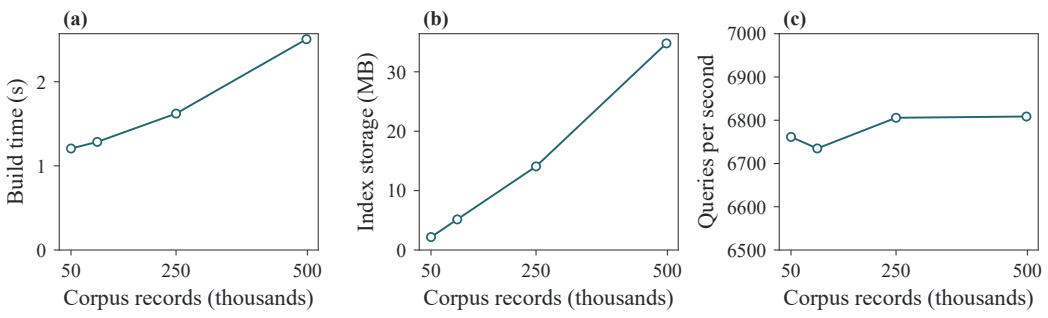

Figure G.1: Measured canonical-key scaling from 50,000 to 498,266 string-deduplicated records: (a) build time, (b) estimated key-set storage, and (c) query throughput.

Table G.1: Measured canonical-key index scaling on the available string-deduplicated mixed corpus.

Requested size	String-deduplicated records used	Build wall time (s)	Molecules/s	Estimated key storage (MB)	Queries/s
50,000	50,000	1.206	41,446	2.18	6,761
100,000	100,000	1.285	77,820	5.15	6,735
250,000	250,000	1.622	154,176	14.08	6,806
500,000	498,266	2.506	198,842	34.79	6,809

Query throughput includes key normalization plus set membership for the first 512 affected records repeated 30 times. Estimated storage is the total UTF-8 key bytes plus 16 bytes per set entry; 776.7 MB is the separately observed peak process resident set size (RSS). Canonical, stereo, and scaffold routes were fully materialized at 498,266 records. Parent and tautomer normalization were capped at 100,000 and their larger-size projections are excluded from Table G.1 and Figure G.1.

G.2 ZINC15 PROXY-CORPUS AUDIT

We precompute all four keys for every post-filter task row and build relation-specific sets from the hash-identified 250,000-record ZINC15 cache. Exact–parent–stereo–tautomer display precedence produces mutually exclusive counts of 4, 0, 18, and 2 matches among 67,228 endpoint records, with 67,204 unmatched. Repeated molecules across endpoints remain separate evaluation records; the query uses corpus records alone.

H MACHINE-READABLE PROVENANCE RECORD

The record format binds the evaluated split, corpus, and intervention specification to relation-typed retrieval evidence and, when available, a seed-level downstream summary. A JSON Schema checks required fields, formats, and internal count constraints. Input digests and version identifiers make the evaluated material explicit, while record-level edges can be regenerated from the bound inputs and code revision rather than copied into the record. Here, *certificate* means a schema-validated audit record: it supports provenance inspection but is not by itself evidence that a model used a matched record or that exposure caused a downstream effect.

H.1 VALIDATED EXAMPLE

Table H.1 summarizes an anonymized primary-protocol record for the stereo-$5\times$ intervention. It covers 2,039 BBBP molecules (1,575/260/204 train/validation/test), 64 affected test molecules, and 12 paired seeds, and validates against schema `2.0.0`.

Table H.1: Selected scientific fields from an anonymized, schema-valid primary-protocol provenance record.

Record component	Reported value
Input binding	BBBP: 2,039 records (1,575/260/204); clean corpus: 50,000 records; stereo-$5\times$ corpus: 50,000 records; abbreviated digests `d082c88ab86b`, `84d3d2fd46eb`, and `34acae0d2446`
Intervention scope	64 affected test molecules; 320 manifest rows; 64 sources and 64 inserted structures; digest `98c85fa44312`
Relation evidence	$\mathrm{Rec}_{\Delta} = 1.0000$ ($n = 31$); pre-injection affected-query match fraction $= 0.046875$; incremental fraction $= 0.953125$
Downstream evidence	Mean-minus-decoy -0.00528 $[-0.02199, 0.01232]$ over 12 seeds; the interval crosses zero
Validation	Schema `2.0.0`; record digest `d76da523f9ba`

Complete digests remain in the machine-readable record. Tables F.2–F.3 report the two-stage scaffold/seed reanalysis used for the main downstream result.