pharmaceuticals

MDPI

Article

EviAlign: Signed Re-Interpretation of Biomedical Literature Graphs for Mechanism-Aware Drug Repurposing

Xin He [1], Xinyu Guo [1], Christophe Ye [1], Batuhan Nursal [1] and Cassie S. Mitchell [1,2,*]

[1] Laboratory for Pathology Dynamics, Georgia Institute of Technology and Emory University School of Medicine, Atlanta, GA 30332, USA; xhe383@gatech.edu (X.H.)
[2] Center for Machine Learning at Georgia Tech, Georgia Institute of Technology, Atlanta, GA 30332, USA
* Correspondence: cassie.mitchell@bme.gatech.edu

Abstract

Background/Objectives: Literature-derived knowledge graphs are attractive for drug repurposing because every edge retains its supporting publications. Their predicates, however, record which relation a sentence asserted, not whether the evidence supports a therapeutic direction. We asked whether reading those publications can align such a graph with the repurposing task. **Methods**: EviAlign re-interprets each edge of disease-centered SemMedDB subgraphs. A large language model (LLM) judge reads the supporting publications sentence by sentence, assigns a therapeutic direction or withdraws the edge, and weights each edge by evidence strength rather than publication count. Signed edges compose along drug–mediator–disease paths. Six diseases and 68,689 edges were evaluated against the original predicates read through a fixed mapping and trimmed to the same size, using curated drug–disease benchmarks, an internal sign-consistency test, and blinded rating by three domain experts. **Results**: EviAlign increased average precision by 63% over the size-matched fixed-mapping baseline. For candidates with no direct disease literature, Precision@10 increased approximately threefold (0.050 versus 0.018). Direct and mediator-derived signs agreed in 75.0% of evaluable cases, 16 percentage points above a shuffled-label control; the fixed mapping stayed at the level of its own control. Experts confirmed the re-interpreted direction in 34 of 36 reversal ratings and judged 80% of top-ranked mediator-only candidates worth investigating, versus 53% of low-ranked controls. **Conclusions**: Reading the evidence behind each edge aligns literature-derived knowledge graphs with drug repurposing and yields interpretable, mechanism-grounded hypotheses for further investigation.

Keywords: drug repurposing; literature-based discovery; signed graphs; knowledge graph; sentence-grounded re-interpretation; evidence weighting; large language models

Check for updates

Academic Editors: Francisco Rivas García, Maria José Zarzuelo Romero and Margarita López-Viota

Received: 7 October 2026
Revised: 29 October 2026
Accepted: 29 October 2026
Published: 31 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Drug repurposing seeks new therapeutic uses for approved compounds. Because their safety and pharmacokinetic profiles are already established, successful candidates can advance toward clinical evaluation substantially faster than new molecular entities [1,2]. Computational repurposing has moved from similarity heuristics toward reasoning over knowledge graphs, in which drugs, diseases, and molecular entities are nodes and their relations are edges [3]. Heterogeneous graphs such as Hetionet, and the models built on them, are now a major line of computational drug-repurposing research [4–9]. Their validation remains a concern: a highly ranked candidate may reflect biological evidence, or

artefacts of graph coverage and topology [10]. Reviews therefore emphasize interpretable, mechanism-aware approaches [11]. Path-based methods address this need by linking a candidate to the disease through an explicit biological mechanism [12–17].

Literature-derived graphs are particularly suited to mechanism-aware repurposing. Swanson showed that two disjoint literatures joined through a shared intermediate can reveal a therapeutic hypothesis [18], and later work recovered such links at scale from semantic predications [17,19]. SemRep and SemMedDB turn PubMed titles and abstracts into subject–predicate–object relations over biomedical concepts [16,20,21]. This gives path-based repurposing both of its essential inputs: mediator paths are explicit in the graph, and every edge retains the publications from which it was extracted. The paths supply the hypotheses; the publications supply the evidence.

The difficulty is that a general-purpose literature graph is not constructed to answer a therapeutic question. Previous work has improved its usefulness by removing extraction errors, representing factuality, and restricting large graphs to relations relevant to a particular prediction task [22–25]. SemRep itself illustrates the need for such processing: relation extraction remains imperfect, with entity recognition and normalization contributing substantial error [21,26]. These approaches improve the graph by filtering or qualifying its predications, but they leave a more fundamental mismatch unresolved. A predicate records the relation asserted by a sentence; drug repurposing asks whether the underlying evidence supports a therapeutic direction. A drug may therefore be linked to a disease because it was administered experimentally, used diagnostically, proposed for future study, or reported without an efficacy outcome, yet the resulting predicate can still be treated as therapeutic support. Likewise, a mediator–disease association does not by itself establish whether that mediator is harmful or protective. Repurposing methods nevertheless commonly use graph predicates directly, through fixed rules or learned weights [11,14,20], while the publications supporting the edges remain unread.

This distinction becomes especially important for mechanism-based discovery. For a drug–mediator–disease path to support repurposing, both directions must be correct: whether the drug inhibits or activates the mediator, and whether the mediator promotes or protects against the disease. An error or unresolved direction on either edge changes the therapeutic interpretation of the entire path. At the same time, the literature attached to each edge provides the evidence needed to resolve these directions. Reading that evidence at scale has become practical through biomedical language models, annotated corpora, and recent evidence that large language models can adjudicate biomedical claims [27–33].

This paper presents EviAlign, a framework that aligns a literature-derived knowledge graph to the drug-repurposing task by re-interpreting each edge from its supporting evidence. A language-model judge reads the publications supporting an edge at the sentence level and determines whether they establish one therapeutic direction, the opposing direction, or neither. These labels carry signs that compose along mediator paths, while edge weights reflect the strength of the supporting evidence rather than publication count alone. The resulting graph prioritizes candidates while retaining the specific mediator paths and literature evidence underlying each score.

Across six diseases and 68,689 edges, we evaluate four questions. First, does evidence-based re-interpretation improve recovery of curated drug–disease pairs relative to a matched-budget fixed-predicate baseline, including candidates with no direct disease literature? This baseline reads each edge's direction from its predicate and is randomly trimmed to the same number of edges as EviAlign, so that both graphs are compared at the same size. Second, does re-interpretation remove unsupported directions and recover useful directional information from relations that a fixed mapping cannot use, and does evidence weighting improve candidate prioritization beyond publication counts? Third, do independently

estimated direct and mediator routes agree on therapeutic direction more often after re-interpretation? Fourth, do the resulting rankings provide mechanistically informative hypotheses beyond what is already stated in the drug's direct disease literature? These claims are evaluated using external curated benchmarks, intrinsic sign-consistency tests, controlled mechanism comparisons, and blinded independent expert assessment. Figure 1 summarizes the complete pipeline.

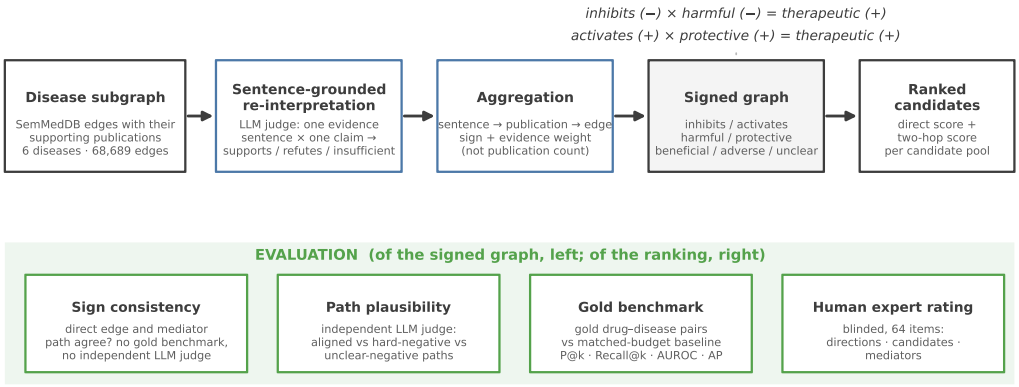


Figure 1. The EviAlign pipeline. Each edge of a disease subgraph carries its supporting publications. The LLM judge reads those publications one evidence sentence at a time and decides, per edge, whether they support one direction, the other, or neither; the verdicts are aggregated from sentences to publications to the edge, and the sentence weights accumulated along the way become the evidence weight in place of the publication count. The result is a signed graph whose signs multiply along mediator paths, from which candidates are ranked by the direct score and the two-hop score within each candidate regime. The signed graph is validated by sign consistency and by path plausibility; the ranking is validated against the gold benchmark and by blinded human expert rating.

2. Results

We evaluated EviAlign from the final drug-ranking outcome back to the evidence that produces it. The analyses first test whether evidence-grounded re-interpretation improves recovery of known drug–disease relationships and whether the resulting signed graph is internally coherent. We then isolate the contributions of edge re-interpretation and evidence weighting, examine whether the inferred mediator paths are biologically plausible and provide information beyond direct drug–disease literature, and finally evaluate the resulting interpretations and hypotheses with blinded domain experts.

Section 2.1 reports ranking performance against the gold benchmark, followed by intrinsic sign consistency in Section 2.2. Sections 2.3 and 2.4 examine the contributions of re-interpretation and evidence weighting, respectively, and Sections 2.5 and 2.6 evaluate mediator paths and individual repurposing hypotheses. Section 2.7 reports the blinded human expert evaluation.

2.1. The Ranking Recovers Gold Drug–Disease Pairs

EviAlign recovers gold drug–disease pairs more accurately than the matched-budget baseline in all three candidate regimes. Candidates are ranked within their candidate pool by the score of Section 4.5, and a candidate is counted as a hit when the gold benchmark of Section 4.7 lists it for that disease. The baseline is the matched-budget baseline of Section 4.7, redrawn 100 times. Precision@k and Recall@k measure the top of the ranking; AUROC and average precision (AP) measure the ranking as a whole. AP is the more informative of the two here, because gold drugs are a small fraction of every candidate pool.

In the two regimes that have a direct edge, the improvement is large and holds on every metric (Table 1, Figure 2A,B). Average precision rises from 0.175 to 0.307 in `direct_only` and from 0.216 to 0.353 in `overlap`, and AUROC from 0.534 to 0.705 and from 0.538 to 0.738. Eleven of these twelve comparisons exceed all 100 baseline draws and the twelfth exceeds 99. The improvement is also present in each disease taken separately: average precision beats the baseline in all six diseases in both regimes (Figure 2C; one-sided sign test $p = 0.016$ per regime). Concretely, the top 100 of a candidate pool of roughly 700 drugs now contains half of the gold pairs, against under a third for the baseline.

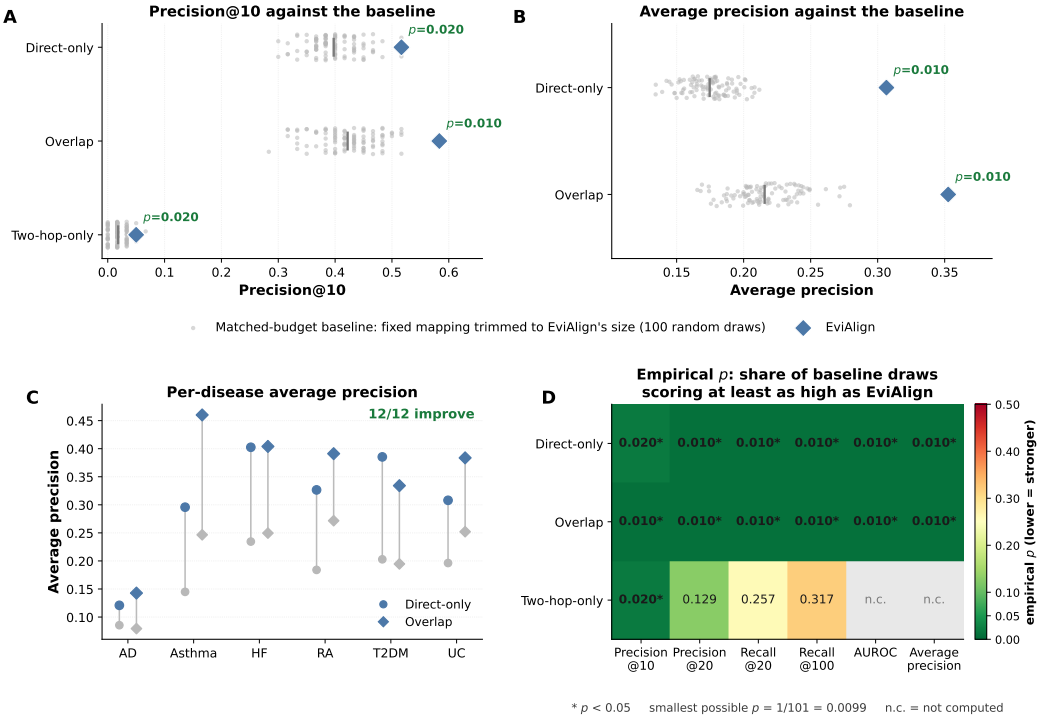


Figure 2. Ranking against the matched-budget baseline. (**A**) Precision@10 and (**B**) average precision against the distribution of 100 baseline draws; the diamond is EviAlign, grey points are individual draws, the vertical bar is their mean. (**C**) Per-disease average precision, baseline in grey and EviAlign in blue, for the two regimes in which the score ranks the whole candidate pool. (**D**) Empirical p across metrics and regimes: the share of the 100 baseline draws that score at least as high as EviAlign, computed as $(1+\#\{\text{draws} \geq \text{EviAlign}\})/101$. The smallest attainable value is 0.0099, and cells marked with an asterisk are below 0.05. AUROC and AP are not computed for `twohop_only` (n.c.), where most candidates have no aligned path and share a score of zero.

`twohop_only` is the regime in which the ranking proposes candidates rather than recovers them. Its candidates have no direct literature for the disease, so by construction they are rarely in curated resources, which are assembled from drugs whose uses are already documented. Gold drugs make up 1.2% of this candidate pool, against 5.6% and 7.5% of the other two. The gold benchmark can therefore ask only whether the few documented candidates rise to the top, and they do: EviAlign reaches Precision@10 of 0.050 against the baseline's 0.018 ± 0.014 ($p = 0.020$), three times the baseline rate, concentrated in the first ten candidates, the part of the list that would be followed up in practice. Whether the undocumented candidates are worth testing is a question the benchmark cannot answer; Sections 2.5–2.7 address it. Highly ranked `twohop_only` candidates are therefore mechanism-driven research hypotheses, not validated clinical treatments.

Table 1. Ranking against the matched-budget baseline, redrawn 100 times. The empirical p is the share of draws matching or exceeding EviAlign, computed conservatively as $(1 + \#\{\text{baseline} \geq \text{observed}\})/(N + 1)$, so 0.0099 is the smallest attainable value. AUROC and average precision (AP) summarise a complete ranking and are reported for the regimes in which the score ranks the whole candidate pool.

Regime	Metric	EviAlign	Baseline (Mean ± SD)	Empirical p
Direct-only	Precision@10	0.517	0.397 ± 0.044	0.0198
Direct-only	Precision@20	0.450	0.289 ± 0.029	0.0099
Direct-only	Recall@20	0.232	0.150 ± 0.019	0.0099
Direct-only	Recall@100	0.497	0.313 ± 0.027	0.0099
Direct-only	AUROC	0.705	0.534 ± 0.022	0.0099
Direct-only	AP	0.307	0.175 ± 0.017	0.0099
Overlap	Precision@10	0.583	0.422 ± 0.052	0.0099
Overlap	Precision@20	0.433	0.331 ± 0.032	0.0099
Overlap	Recall@20	0.282	0.207 ± 0.024	0.0099
Overlap	Recall@100	0.605	0.389 ± 0.032	0.0099
Overlap	AUROC	0.738	0.538 ± 0.028	0.0099
Overlap	AP	0.353	0.216 ± 0.024	0.0099
Two-hop-only	Precision@10	0.050	0.018 ± 0.014	0.0198
Two-hop-only	Precision@20	0.025	0.015 ± 0.008	0.1287
Two-hop-only	Recall@20	0.018	0.016 ± 0.025	0.2574
Two-hop-only	Recall@100	0.048	0.056 ± 0.053	0.3168

2.2. The Re-Interpreted Graph Is Sign-Consistent, and Consistency Is a Mediator Property

Where a drug reaches a disease both through a direct edge and through a mediator path, the graph states the therapeutic sign twice, and the two statements can be compared. EviAlign is sign-consistent on 75.0% of the 2537 triangles that carry a sign on all three edges. The fixed mapping, checked the same way over its own 19,801 triangles, is sign-consistent on 52.5% (Table 2, Figure 3B). Because the labels are skewed, some agreement follows from the label mix alone. Shuffling the labels within each edge family keeps that mix but breaks the link between the direct edge and the mediator path. It gives 59.0% ± 0.8 for EviAlign and 51.8% for the fixed mapping. EviAlign stands 16.1 percentage points, or 19.7 standard deviations, above its shuffled-label control, and above all 10,000 shuffles ($p < 10^{-4}$). The fixed mapping sits at the level its own labels would produce.

Table 2. Sign consistency. Wherever a drug has both a direct edge and a mediator path, the graph asserts the same therapeutic sign twice; the two assertions can be checked against each other without the gold benchmark or any LLM judge. The shuffled-label control shuffles labels within each edge family, preserving their distribution. Measured against its own shuffled-label control, EviAlign is 16.1 percentage points higher; the fixed mapping stays at its own control level.

Graph	Triangles	Sign-Consistent	Vs. Shuffled-Label Control
Fixed mapping	19,801	52.5%	at its own control level (51.8%)
EviAlign	2537	75.0%	+16.1 points, $p < 10^{-4}$
Shuffled-label control (EviAlign)	—	59.0% ± 0.8	—

The comparison needs no gold benchmark and no independent LLM judge. The direct edge and the mediator path are also re-interpreted from different publications, so their agreement is evidence that the signs capture pharmacological structure rather than relabelling edges. It also justifies the scoring rule of Section 4.5, since multiplying signs along a path is only meaningful on a sign-consistent graph. Sign consistency exceeds the shuffled-label control in all six diseases (Figure 3C), from 80.0% in rheumatoid arthritis, which

contributes 1017 triangles, to 65.2% in ulcerative colitis, which contributes 23. Ulcerative colitis is too small to interpret on its own and is marked as such in Figure 3C.

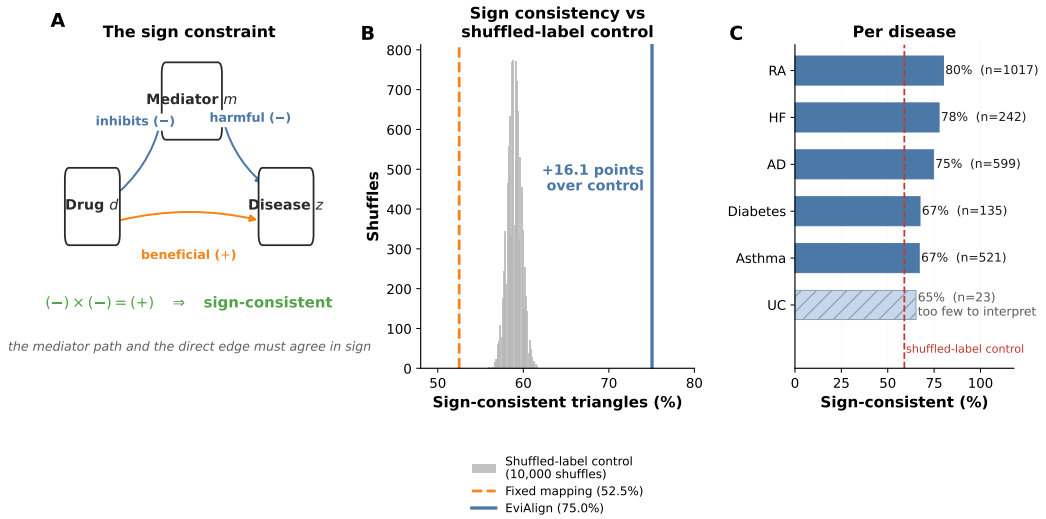


Figure 3. Intrinsic validation by sign consistency. (**A**) The constraint: where a drug reaches a disease both through a direct edge and through a mediator path, the two assert the same therapeutic sign and can be checked against each other. (**B**) Observed sign consistency against the shuffled-label control over 10,000 shuffles, with the fixed mapping shown for comparison. (**C**) Per-disease sign consistency; the dashed line is the shuffled-label control. Ulcerative colitis (23 triangles, light bar) is too small to interpret on its own.

Consistency is concentrated rather than spread evenly, which is what makes it usable: knowing the mediator then tells us which paths to trust. The fraction of a mediator's triangles that are sign-consistent, which we call its *mediator reliability*, is strongly bimodal (Figure 4A): 36% of mediators are consistent in every triangle they enter and 10% in none.

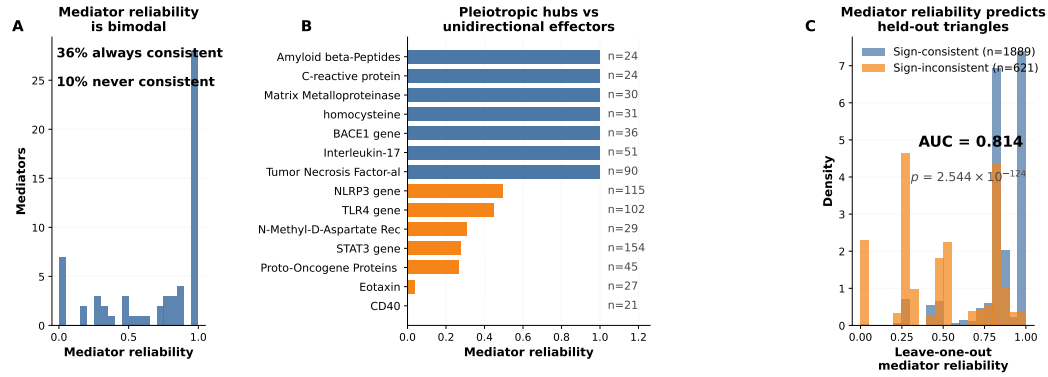


Figure 4. Sign consistency is a property of the mediator. (**A**) Mediator reliability is bimodal rather than uniform. (**B**) Mediators with at least 15 triangles, ordered by mediator reliability. (**C**) Leave-one-out mediator reliability, computed with the predicted triangle removed, separates sign-consistent from sign-inconsistent held-out triangles.

The split is biologically interpretable (Figure 4B, Table 3). Fully consistent mediators are unidirectional effectors whose therapeutic sign does not depend on context: TNF-α (90 triangles), interleukin-17 (51), BACE1 (36), homocysteine (31) and matrix metalloproteinase 9 (30). The least consistent are pleiotropic signalling hubs whose sign genuinely varies with cell type and disease stage: STAT3 (154 triangles, 28% consistent), TLR4 (102, 45%), AKT (45, 27%) and NLRP3 (115, 50%). The measure uses no information about the molecules themselves, yet it recovers a distinction that pharmacology draws independently.

Table 3. Mediators ranked by mediator reliability (at least 15 triangles). Fully consistent mediators are unidirectional effectors, whereas the least consistent are known pleiotropic signalling hubs whose therapeutic sign is context-dependent. Mediator reliability estimated from other drugs predicts the sign consistency of held-out triangles with AUC = 0.814.

Mediator	Triangles	Sign-Consistent	Interpretation
Tumor Necrosis Factor-alpha	90	100%	unidirectional effector
Interleukin-17	51	100%	unidirectional effector
BACE1 gene	36	100%	unidirectional effector
homocysteine	31	100%	unidirectional effector
Matrix Metalloproteinase 9	30	100%	unidirectional effector
CD40	21	0%	pleiotropic hub
Eotaxin	27	4%	pleiotropic hub
Proto-Oncogene Proteins c-akt	45	27%	pleiotropic hub
STAT3 gene	154	28%	pleiotropic hub
N-Methyl-D-Aspartate Receptors	29	31%	pleiotropic hub

Mediator reliability also predicts. In a leave-one-out test over the 2510 triangles whose mediator carries at least four, each triangle is scored by its mediator's reliability recomputed without that triangle. Sign-consistent triangles score 0.837 on average, against 0.495 for the rest (Figure 4C). Ranking by the score separates them with AUC = 0.814 ($p < 10^{-100}$). A mediator path through TNF-α and one through STAT3 are therefore not equally trustworthy, and the graph says which is which before any validation layer outside the graph is consulted. Used inside scoring, as a path weight or as a path filter, reliability did not improve the gold-benchmark ranking (Appendix D). Keeping it out of the score also keeps sign consistency independent of the method.

2.3. Why Re-Interpretation Is Necessary: The Predicate Is Not the Direction the Task Needs

Re-interpretation withdraws most mapped edges and reverses few (Figure 5A, Table 4): 59–91% of mapped edges are withdrawn and 1.9–4.3% are reversed. Withdrawal does not mean that the predicate was wrong. It means that the edge cannot be aligned to the direction the task needs: the predicate does not carry it, and the supporting publications do not establish it either. Three directions are needed: whether a drug improves or worsens the disease, whether it raises or lowers a mediator's activity, and whether the mediator, in its normal state, drives or restrains the disease. SemRep's predicates come from the UMLS Semantic Network, a general inventory of relation types, and none is defined as one of these directions. The fixed mapping of Table 5 is the standard reading of the predicates [14,20] and the best guess the predicate alone allows. The predicate itself lacks the information, so a method that uses it directly ranks drugs on directions the literature never stated. The predicate answers a different question from the one the task asks. Two predicates show this: `TREATS` on drug–disease edges and `PREDISPOSES` on mediator–disease edges. In both cases the predicate is correctly extracted and true, and in both cases it does not settle the direction the task needs.

On drug–disease edges, the task asks whether the disease got better after the drug. `TREATS` answers whether the drug was used for the disease. Propranolol was given to asthmatics, so SemRep records *propranolol* `TREATS` *asthma*; the asthma got worse, with FEV_1 falling 4.7%. Methacholine was given to asthmatics in 181 diagnostic challenge tests, so SemRep records *methacholine* `TREATS` *asthma*; no publication measured the disease. `TREATS` is correct in both cases. The task's answer is harm in the first and none in the second, and re-interpretation reads that answer from the result sentence: it reverses the propranolol edge to `adverse_for` and withdraws the methacholine edge.

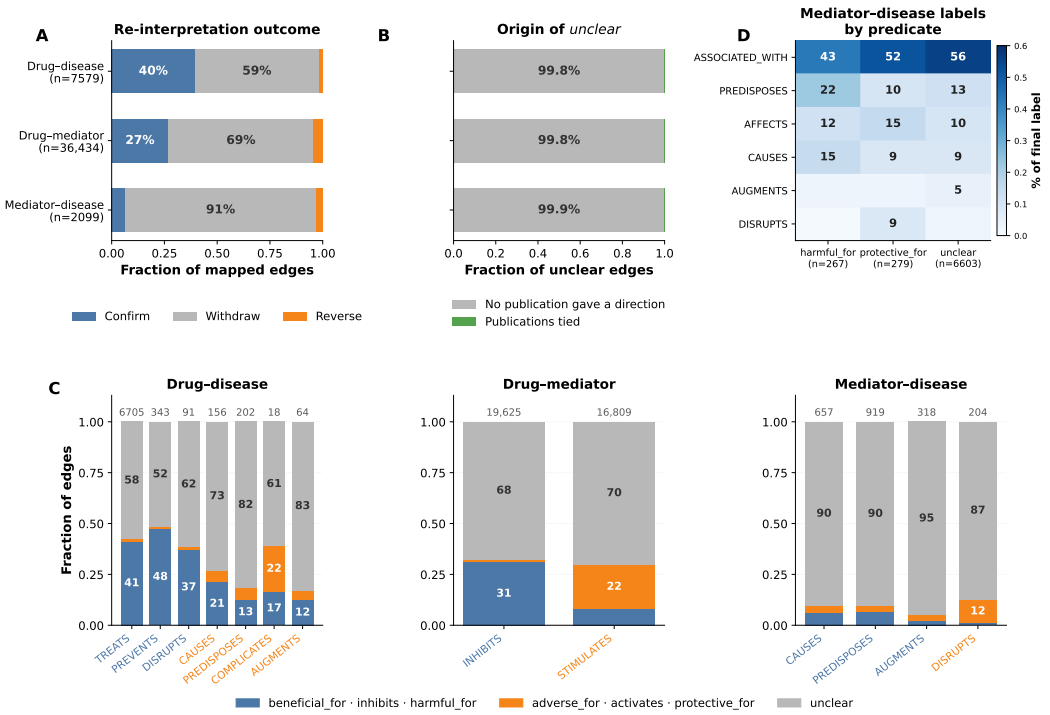


Figure 5. What re-interpretation does to the graph. (**A**) Each mapped edge either confirms the fixed mapping, withdraws its direction to `unclear`, or reverses its sign; reversal is rare in all three families. (**B**) Withdrawals arise almost entirely because no supporting publication establishes a therapeutic direction, not because publications disagree. (**C**) Redistribution of each predicate, normalised within the predicate; tick labels are coloured by the direction the fixed mapping would have assigned. (**D**) Composition of the final mediator–disease labels by predicate, with denominators over all predicates.

Table 4. Re-interpretation outcomes over the 46,112 mapped edges. Reversal is rare; the dominant outcome is withdrawal of directions the supporting publications do not establish.

Edge Family	Confirm	Withdraw	Reverse	Total
Drug–disease	2995 (39.5%)	4441 (58.6%)	143 (1.9%)	7579
Drug–mediator	9785 (26.9%)	25,067 (68.8%)	1582 (4.3%)	36,434
Mediator–disease	131 (6.2%)	1905 (90.8%)	63 (3.0%)	2099

Table 5. The fixed mapping: the predicates retained in each edge family and the direction each is assigned. Edges whose predicate is not listed are unmapped.

Edge Family	Predicates	Direction	Role in Scoring
Drug–disease	TREATS, PREVENTS, MANAGES, DISRUPTS	beneficial_for	counted in the direct score
Drug–disease	CAUSES, PREDISPOSES, COMPLICATES, AUGMENTS	adverse_for	retained, not counted
Drug–mediator	INHIBITS	inhibits	aligned with harmful_for
Drug–mediator	STIMULATES	activates	aligned with protective_for
Mediator–disease	CAUSES, PREDISPOSES, COMPLICATES, AUGMENTS	harmful_for	aligned with inhibits
Mediator–disease	PREVENTS, DISRUPTS	protective_for	aligned with activates

On mediator–disease edges, `PREDISPOSES` is correct for both APP and TREM2 in Alzheimer's disease, yet the two genes call for opposite drugs. "Rare variants in the TREM2 gene have been reported to significantly increase the risk of Alzheimer's disease" yields *TREM2* `PREDISPOSES` *Alzheimer's disease*, and the corresponding sentences for APP yield *APP* `PREDISPOSES` *Alzheimer's disease*. Both are true: `PREDISPOSES` records that variation in the gene raises the risk of the disease, and SemMedDB stores every genetic risk finding

this way, because it has no concept for a variant of a gene. The task needs something the predicate does not hold: whether raising the molecule's activity makes the disease worse or better, which decides whether a drug should inhibit or activate it. The fixed mapping reads `PREDISPOSES` as `harmful_for` for both genes: inhibit APP, inhibit TREM2. Other sentences in the same literature state which way the variants act. For APP, more processing means more Aβ and more disease. For TREM2, the risk variants are loss-of-function ("the genetic associations of TREM2 loss-of-function variants with Alzheimer disease indicate the protective roles of microglia"), so less TREM2 means more disease. Re-interpretation confirms `harmful_for` for APP, withdraws the TREM2 `PREDISPOSES` edge, and labels TREM2's directed edges `protective_for`, so a drug that activates TREM2 is the therapeutic direction. The predicate records that the gene matters, not in which direction.

Much of the graph is unusable as stored, and becomes usable only once re-interpretation gives its edges the direction the task needs. A third of the edges carry `ASSOCIATED_WITH`, `AFFECTS` or `INTERACTS_WITH`, which record that two concepts go together and nothing more, so the fixed mapping discards them. Re-interpretation reads the direction their sentences state: "targeting TNF has shown great efficacy in controlling both the inflammation and structural damage of the joints" turns *TNF* `ASSOCIATED_WITH` *rheumatoid arthritis* into `harmful_for`. These edges supply 43% of the final `harmful_for` and 52% of the final `protective_for` mediator–disease edges (Figure 5D).

The order of the two steps matters: verify first, then extend (Table 6). Re-interpreting the mapped edges alone raises sign consistency from 52.5% to 77.1% and `overlap` Precision@10 from 0.417 to 0.533. Adding the unmapped edges raises Precision@10 to 0.567 and multiplies the triangle set fivefold. The reverse order fails. Directing only the unmapped edges, while the mapped edges keep their unverified fixed-mapping labels, enlarges the triangle set from 19,801 to 31,051. Yet sign consistency stays within 1.7 points of its own control, and `overlap` Precision@10 falls from 0.417 to 0.367. A two-hop score multiplies its two edges, so one edge with a wrong or unsupported direction is enough to make the whole path wrong. New paths built on unverified edges therefore add more noise than signal. Appendix E lists further cases.

Table 6. Scope of re-interpretation. Each graph is built over the same candidate pools and evaluated against the same gold benchmark. Sign consistency is compared with a shuffled-label control computed on that graph's own triangle set, since the four graphs differ in size and in label skew. Precision@10 here uses one scoring implementation for all four graphs, so absolute values differ slightly from Table 1; the comparison within this table is like-for-like.

Edges Re-Interpreted	Triangles	Sign-Consistent	Shuffled-Label Control	P@10 Overlap	P@10 Two-Hop
None (fixed mapping)	19,801	52.5%	51.8%	0.417	0.017
Unmapped edges only	31,051	54.2%	52.5%	0.367	0.017
Mapped edges only	462	77.1%	63.1%	0.533	0.033
All (EviAlign)	2537	75.1%	59.0%	0.567	0.050

2.4. Weighting Edges by Evidence Strength Beats Counting Publications

Evidence weighting ranks better than count weighting. The two are compared on the same graph. Both arms use the re-interpreted labels and the scoring rule of Section 4.5, and both are evaluated on the gold benchmark of Section 2.1. The only difference is the edge weight: the publication count of Equation (3) in one arm, the tempered evidence weight of Equation (5) in the other.

Count weighting removes most of the ranking's resolution before scoring begins. Because 78% of signed edges rest on a single publication (Section 4.4), 82% of edges share the modal weight, and two-hop scores, being products of two such weights, tie even more

often. The evidence weight raises the number of distinct weight values from 35 to 982 in a representative disease subgraph and shrinks the largest tied block to 22% (Figure 6A).

The recovered resolution turns into better rankings, and it does so exactly where mediator paths contribute (Table 7). In `overlap`, Precision@10 rises by +0.050, average precision by +0.031 and Recall@100 by +0.034; in `twohop_only` the same three metrics rise by +0.017, +0.001 and +0.011. Precision@10 falls in no regime. `direct_only` is unchanged, as expected: the direct score is a sum of weights rather than a product, and never suffered the collapse. The gain is therefore confined to the two regimes that multiply weights along a path. The `twohop_only` ranking of Section 2.1 depends on this choice: under count weighting `twohop_only` matches the matched-budget baseline, and only evidence weighting lifts it to three times that rate. The tempering exponent of Equation (5) matters as well, and Appendix D reports its effect together with the evidence scope and the numeric channel.

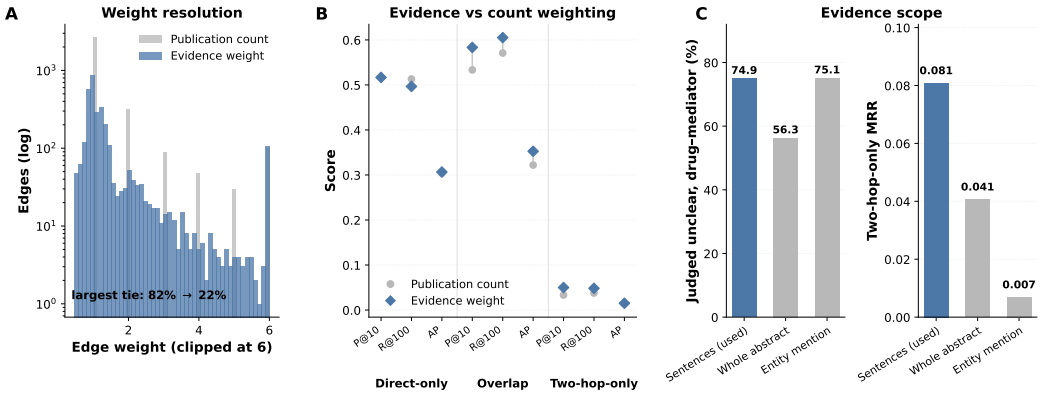


Figure 6. Effect of the weighting scheme. (**A**) Distribution of edge weights under count weighting and evidence weighting; the largest tied block shrinks from 82% to 22% of edges. (**B**) Paired comparison on the gold benchmark across the three candidate regimes and the three metrics of Table 7. (**C**) Evidence scope. The LLM judge reads the selected evidence sentences (used throughout), the whole title and abstract in one call, or only the sentences that mention an entity of the edge. (**Left**): share of drug–mediator judgements left `unclear`. (**Right**): `twohop_only` mean reciprocal rank of the first gold drug, averaged over the six diseases. Judging the whole abstract leaves fewer judgements `unclear` but ranks gold drugs lower (Appendix D).

Table 7. Evidence weighting versus count weighting on the gold benchmark, with the labels, the scoring rule and the candidate pools held fixed.

Regime	Metric	Publication Count	Evidence Weight	Δ
Direct-only	Precision@10	0.5167	0.5167	+0.0000
Direct-only	Recall@100	0.5130	0.4966	−0.0164
Direct-only	AP	0.3041	0.3065	+0.0025
Overlap	Precision@10	0.5333	0.5833	+0.0500
Overlap	Recall@100	0.5709	0.6051	+0.0342
Overlap	AP	0.3220	0.3526	+0.0306
Two-hop-only	Precision@10	0.0333	0.0500	+0.0167
Two-hop-only	Recall@100	0.0378	0.0482	+0.0105
Two-hop-only	AP	0.0138	0.0150	+0.0013

2.5. Mediator Paths Carry a Plausibility Gradient

The independent LLM judge rates aligned paths as more biologically plausible than hard-negative paths, and hard-negative paths as more plausible than unclear-negative paths. The judge never sees the labels. Sign consistency shows that the graph's signs agree with one another, not that its mediator paths are biologically plausible. *Path plausibility*

therefore has the independent LLM judge rate complete mediator paths on three points: the plausibility of each of the two edges on its own, and whether the drug can be inferred to help the disease through that mediator. We call the third the *therapeutic inference score*. The three path types of Table 8 were rated: aligned paths satisfying Equation (6), hard-negative paths that are directionally explicit but sign-inverted, and unclear-negative paths built from `unclear` edges. Within each disease the groups were subsampled to the size of the smallest.

On the therapeutic inference score the three types differ strongly (Table 9, Figure 7A; Kruskal–Wallis $H = 468.2$, $p \approx 2 \times 10^{-102}$, $n = 970$). All three pairwise orderings hold under one-sided Mann–Whitney U tests with Holm correction, with large effect sizes, reaching $\delta = 0.889$ for aligned against unclear-negative. The same ordering holds in five of the six diseases taken separately.

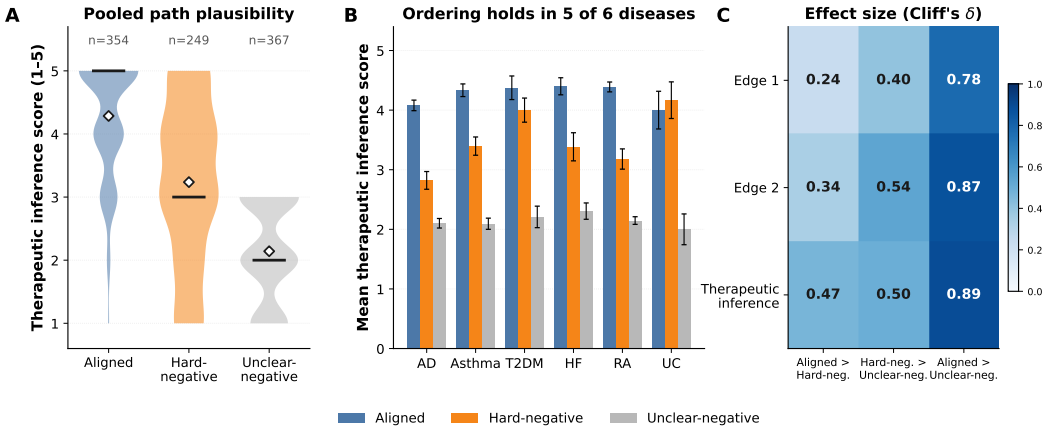


Figure 7. Path plausibility. (**A**) Pooled therapeutic inference scores by path type; horizontal bars are medians, diamonds are means. (**B**) Per-disease means with standard errors. (**C**) Cliff's δ for all pairwise contrasts, for each edge on its own and for the therapeutic inference score.

Table 8. The three path types. A mediator path is aligned when its two signs multiply to +, hard-negative when they multiply to −, and unclear-negative when at least one edge is `unclear`; the rated unclear-negative paths have `unclear` on both edges.

Path Type	$d \to m$	$m \to z$	Interpretation
Aligned	inhibits	harmful_for	signs multiply to +; counted in the two-hop score
Aligned	activates	protective_for	signs multiply to +; counted in the two-hop score
Hard-negative	inhibits	protective_for	directionally explicit, signs multiply to −
Hard-negative	activates	harmful_for	directionally explicit, signs multiply to −
Unclear-negative	unclear	unclear	no sign on either edge

Table 9. Path plausibility. One-sided Mann–Whitney U with Holm correction within each score; Cliff's δ as effect size.

Score	Comparison	Mean	Holm p	Cliff's δ	Kruskal H
Edge 1	Aligned > Hard-negative	3.91 vs. 3.43	2.02×10^{-8}	0.242	341.7
Edge 1	Hard-negative > Unclear-negative	3.43 vs. 2.82	1.11×10^{-21}	0.403	341.7
Edge 1	Aligned > Unclear-negative	3.91 vs. 2.82	2.60×10^{-85}	0.776	341.7
Edge 2	Aligned > Hard-negative	4.33 vs. 3.68	1.68×10^{-14}	0.340	462.4
Edge 2	Hard-negative > Unclear-negative	3.68 vs. 2.86	1.82×10^{-37}	0.538	462.4
Edge 2	Aligned > Unclear-negative	4.33 vs. 2.86	2.92×10^{-104}	0.869	462.4
Therapeutic inference	Aligned > Hard-negative	4.29 vs. 3.24	6.51×10^{-25}	0.467	468.2
Therapeutic inference	Hard-negative > Unclear-negative	3.24 vs. 2.14	1.86×10^{-27}	0.497	468.2
Therapeutic inference	Aligned > Unclear-negative	4.29 vs. 2.14	1.39×10^{-99}	0.889	468.2

Two checks show that the independent LLM judge is scoring the composition rather than the individual edges. First, hard-negative paths are locally as explicit as aligned paths and differ from them only in how the two signs compose. On that contrast the therapeutic inference score separates the groups more strongly ($\delta = 0.467$) than either edge judged alone ($\delta = 0.242$ and 0.340). The judge is therefore responding to whether the two edges compose into a mechanism, rather than to local plausibility (Figure 7C). Second, hard-negative paths score between the aligned and the unclear-negative groups rather than with either: below aligned paths because their two signs multiply to −, and above unclear-negative paths because both of their edges still state a direction.

2.6. The Ranking Is Interpretable: Every Score Decomposes into Named Mediator Paths

Every score decomposes into the named mediator paths that produced it, so each ranking decision can be audited (Figure 8, Table 10). Spironolactone is ranked third in the `twohop_only` candidate pool for rheumatoid arthritis, on paths inhibiting cytokine, TNF and matrix metalloproteinase 2. Fluorouracil is ranked fifth for Alzheimer's disease, on paths inhibiting cytokine- and NLRP3-centred mediators, together with a path activating a protective TLR4-related layer. Both are gold drugs, and neither has a direct edge, so the whole of each score comes from mediator paths. The candidates the same ranking places above them are outside the gold benchmark but carry support of the same form, such as AG-490 in heart failure through inhibition of STAT3- and interleukin-1β-centred signalling. We report those as hypotheses grounded in mediator paths, not as validated repurposing successes.

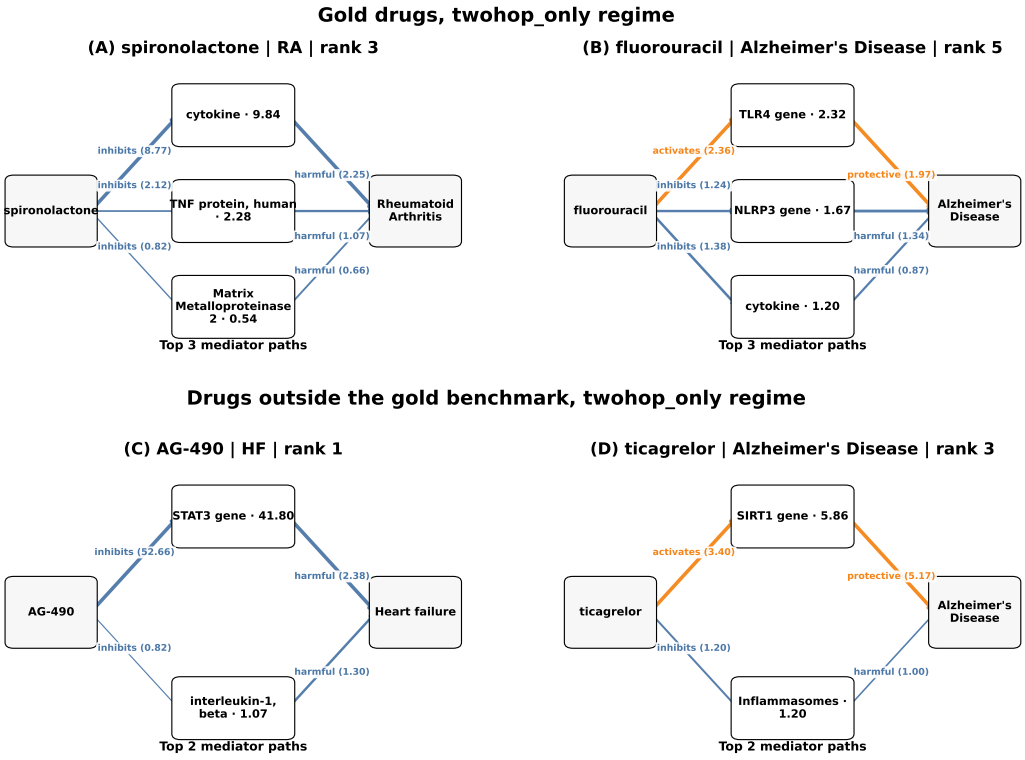


Figure 8. Representative `twohop_only` cases, ranked without any direct edge. (**A**) Spironolactone for rheumatoid arthritis (rank 3) and (**B**) fluorouracil for Alzheimer's disease (rank 5) are gold drugs; (**C**) AG-490 for heart failure (rank 1) and (**D**) ticagrelor for Alzheimer's disease (rank 3) are drugs outside the gold benchmark. Arrow labels give the edge's direction and weight; the value inside each mediator node is that mediator's contribution to the two-hop score. Blue edges are labelled `inhibits` or `harmful_for`, orange edges `activates` or `protective_for`.

Table 10. Disease-level summary of the `twohop_only` cases of Figure 8: the mediators that contribute most to the two-hop score, the gold drugs recovered, and the highest-ranked drugs outside the gold benchmark.

Disease	Major Mediators	Gold Drugs	Drugs Outside the Gold Benchmark
Rheumatoid arthritis	cytokine; TNF protein, human	spironolactone (rank 3); chlorpromazine (rank 25)	Protein Kinase C Inhibitor (rank 1); Adrenergic Agonists (rank 2)
Alzheimer's disease	cytokine; homocysteine	fluorouracil (rank 5); prednisolone (rank 7)	cynandione A (rank 1); budesonide (rank 2)
Heart failure	interleukin-1, beta; STAT3 gene	—	AG-490 (rank 1); andrographolide (rank 2)

For `overlap` candidates the explanation does not merely repeat the drug's *direct literature*, the supporting publications of its direct edge. *Mediator recovery* tests this. For each `overlap` drug we took its top five mediators and searched its direct literature for them, first by string matching and then by entity linking against UMLS [34], which also catches synonyms and abbreviations. Of 479 mediator instances, 19.0% were found by string matching and a further 21.1% only after entity linking, leaving 59.9% absent from the direct literature altogether (Figure 9, Table 11). The mediator paths that explain an `overlap` candidate therefore carry mechanism its direct literature never states, so the explanation extends the direct literature rather than restating it.

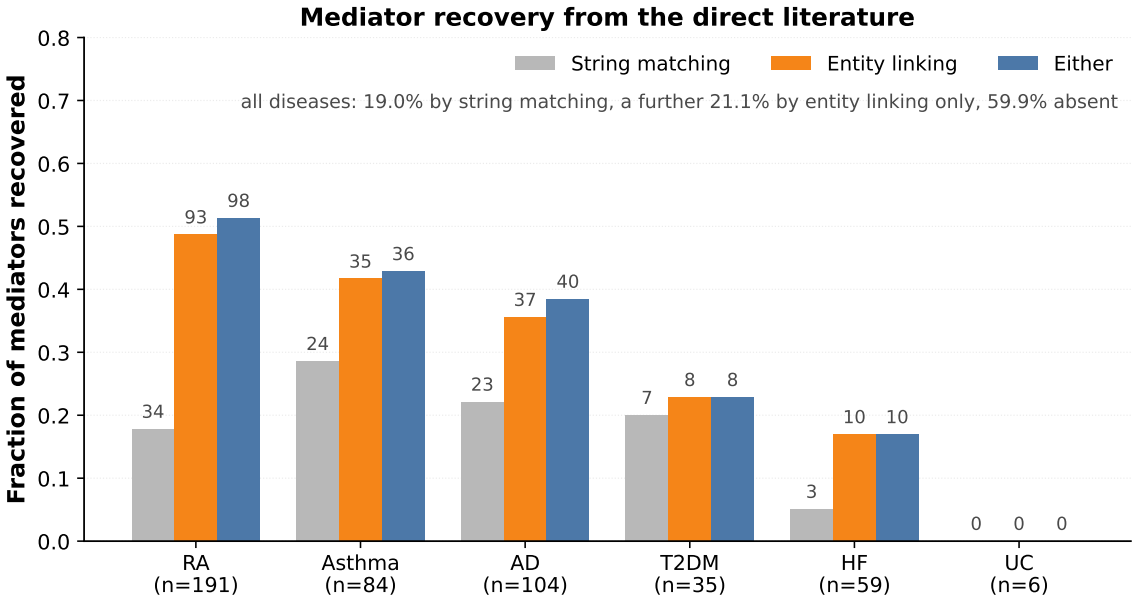


Figure 9. Mediator recovery: the share of `overlap` drugs' top mediators found in their direct literature by string matching, by entity linking, and by either. Entity linking raises recovery in every disease and leaves it incomplete in every disease.

2.7. Blinded Human Expert Rating

Three independent biomedical domain experts completed the blinded evaluation. The experts differed modestly in their overall threshold for endorsing the proposed interpretations and hypotheses. For descriptive comparison, the expert with the intermediate endorsement pattern was treated as the middle-ground reference, while the experts at either end of the observed distribution were described as relatively less conservative and

relatively more conservative. These designations describe observed differences in response threshold and were not prospectively assigned reviewer roles.

Table 11. Representative `overlap` mediators for relatively specific drugs. "entity linking" marks mediators not found in the direct literature by string matching but found by entity linking; "unrecovered" marks mediators absent from the direct literature by either method.

Disease	Drug	Mediator	Recovery	Path
Asthma	dexamethasone	TNF gene	entity linking	inhibits + harmful_for
Asthma	budesonide	Interleukin-17	entity linking	inhibits + harmful_for
Asthma	vitamin D	CD14 gene	unrecovered	activates + protective_for
HF	vitamin D	PTH protein, human	entity linking	inhibits + harmful_for
RA	dexamethasone	interleukin-1, beta	unrecovered	inhibits + harmful_for
RA	andrographolide	TNF protein, human	unrecovered	inhibits + harmful_for
T2DM	vitamin D	KL gene	unrecovered	activates + protective_for
UC	prednisolone	Kynurenine	unrecovered	inhibits + harmful_for
UC	decitabine	RIPK3 gene	unrecovered	activates + protective_for

Despite this variation in judgment threshold, agreement across the three experts was high (Table 12). The resulting judgments were robust to inter-rater variation in evidentiary threshold. Across the three experts, T1 edge re-interpretation showed 91.7% item-level unanimity and high nominal agreement (Krippendorff's $\alpha = 0.911$). The profiles supported the EviAlign interpretation on 34 of 36 reversal ratings (94.4%, 95% CI: 81.9–98.5%) and selected *Neither* on all 36 withdrawal ratings (100.0%, 95% CI: 90.4–100.0%). T2 candidate prioritization was more sensitive to the investigational threshold, with 75.0% unanimity and ordinal $\alpha = 0.704$. Experts judged 36 of 45 top-ranked `twohop_only` candidates reasonable to investigate (80.0%, 95% CI: 66.2–89.1%), against 8 of 15 bottom-ranked controls (53.3%, 95% CI: 30.1–75.2%; one-sided Mann–Whitney U, $p = 0.035$). The ranking therefore separates candidates experts would pursue from those they would not. The bottom end of the ranking is noisier. The controls were the lowest-ranked candidates that still had an aligned mediator path, so experts saw a plausible mechanism for each. The ranking is therefore most informative at the top; a low rank should not be used to exclude a candidate. Disagreements occurred primarily at the boundary between whether a biologically plausible but limited mechanism was sufficient to prioritize a candidate for further study. T3 mediator plausibility remained highly concordant despite variation in judgment threshold, with 90.0% unanimity and ordinal $\alpha = 0.902$. Mediators that the drug's direct literature never names were judged plausible or left open in 39 of 45 ratings (86.7%) and implausible in only 6 (13.3%, 95% CI: 6.3–26.2%). Recovered-mediator controls received no implausible rating, and the two groups did not differ ($p = 0.87$). The mediators that re-interpretation adds beyond the direct literature were therefore judged as credible as those the literature already states. Thus, re-interpretation and mediator-level judgments were largely stable across plausible changes in evidentiary threshold. Candidate prioritization appropriately showed greater sensitivity to investigator willingness to advance an early-stage repurposing hypothesis.

Table 12. Blinded human expert evaluation. Three independent domain experts rated the same blinded evaluation instruments. Pooled proportions are reported with Wilson 95% confidence intervals. Inter-rater concordance is summarized by item-level unanimity and Krippendorff's α, using nominal α for T1 and ordinal α for T2 and T3. Pairwise agreement is reported as exact agreement and Cohen's κ.

Panel A. Evaluation Outcomes and Concordance	
T1 — Edge re-interpretation	
Reversal items supporting EviAlign	34/36 (94.4%) [81.9–98.5%]
Withdrawal items selecting Neither	36/36 (100.0%) [90.4–100.0%]
Unanimous items	22/24 (91.7%)
Krippendorff's α (nominal)	0.911
T2 — Repurposing candidate prioritization	
Top-ranked candidates rated Agree	36/45 (80.0%) [66.2–89.1%]
Bottom-ranked attention controls rated Agree	8/15 (53.3%) [30.1–75.2%]
Unanimous items	15/20 (75.0%)
Krippendorff's α (ordinal)	0.704
T3 — Mediator plausibility	
Novel mediators rated Agree	19/45 (42.2%) [29.0–56.7%]
Novel mediators rated Disagree	6/45 (13.3%) [6.3–26.2%]
Recovered attention controls rated Agree	8/15 (53.3%) [30.1–75.2%]
Unanimous items	18/20 (90.0%)
Krippendorff's α (ordinal)	0.902

Panel B. Pairwise Profile Concordance

Instrument	Profile comparison	Exact agreement	Cohen's κ
T1	Baseline vs. less conservative	95.8%	0.93
T1	Baseline vs. more conservative	95.8%	0.93
T1	Less vs. more conservative	91.7%	0.86
T2	Baseline vs. less conservative	90.0%	0.706
T2	Baseline vs. more conservative	85.0%	0.701
T2	Less vs. more conservative	75.0%	0.465
T3	Baseline vs. less conservative	95.0%	0.92
T3	Baseline vs. more conservative	95.0%	0.92
T3	Less vs. more conservative	90.0%	0.83

Note. T1 contained 24 publication-reading items; T2 and T3 contained 20 items each. Within each instrument, items were presented in random order and conditions were withheld from the experts; conditions were joined from the answer key at analysis time (T2: 15 top-ranked candidates and five bottom-ranked attention controls; T3: 15 mediators absent from the direct drug–disease literature and five recovered-mediator attention controls).

3. Discussion

The findings define both the opportunities and the current boundaries of evidence-grounded, mechanism-aware drug repurposing with literature-derived knowledge graphs.

3.1. Implications for Evidence-Grounded Drug Repurposing

EviAlign addresses a basic mismatch between literature-derived knowledge graphs and drug repurposing: the relation extracted from a publication is not necessarily the therapeutic interpretation needed to prioritize a candidate. A `TREATS` relation, for example, may be extracted whenever a drug is administered to patients with a disease, whether the publication reports benefit, harm, or no efficacy outcome (Section 2.3). Re-interpreting the supporting publications resolves this mismatch conservatively. EviAlign rarely reverses an existing direction (1.9–4.3% of mapped edges); instead, its dominant action is to withdraw directions that the supporting evidence does not establish. Conversely, for relations that do not encode therapeutic polarity, such as `ASSOCIATED_WITH`, re-interpretation can recover

directional evidence that a fixed mapping cannot use. Approximately half of the final signed mediator–disease evidence arises from such relations.

These changes translate into improved drug prioritization. Relative to the matched-budget fixed-mapping baseline, average precision increases by 63% in `overlap` and 75% in `direct_only`, while Precision@10 approximately triples in `twohop_only`. The latter is particularly relevant to repurposing because these candidates have no direct drug–disease literature: their prioritization arises entirely from evidence connecting the drug to disease-relevant mediators. EviAlign therefore does not require a candidate already to have been discussed for the target disease in order to rank it. These rankings should be interpreted as prioritization of mechanistically supported hypotheses for further investigation, rather than as predictions of therapeutic efficacy. Highly ranked candidates, in particular those without direct literature, are mechanism-driven research hypotheses, not validated clinical treatments.

A second finding is that the signed representation provides an intrinsic test of its own biological coherence. Where a drug reaches the same disease both directly and through a mediator, the two routes independently assert a therapeutic sign from different supporting publications. Sign consistency increases from 52.5% under the fixed mapping, approximately the level of its own shuffled-label control, to 75.0% after re-interpretation, 16.1 percentage points and 19.7 standard deviations above the corresponding control. This test requires neither a curated gold standard nor a second language model and therefore provides evidence complementary to conventional external validation.

The consistency signal is also biologically informative rather than uniformly distributed across the graph. It concentrates in comparatively unidirectional effectors such as TNF-α and BACE1 and is lower in pleiotropic signaling hubs such as STAT3. Moreover, mediator reliability estimated from other drugs predicts held-out path consistency with AUC = 0.814. This suggests that reliability can ultimately serve not only as a validation statistic but also as a guide to which mechanistic paths warrant greater trust. Together with the decomposition of every candidate score into its contributing mediator paths, this provides a route from candidate ranking to an inspectable biological hypothesis rather than a black-box prediction.

EviAlign's advantages over related work lie in three properties. The first is *task alignment*. General extraction schemas, even with directionality added [35], label relations such as positive or negative correlation. A positive correlation between IL-6 and a disease does not say whether inhibiting IL-6 would help. Large language models are also benchmarked on schemas that cover a single edge type, such as chemical–protein relations [33], which cannot form drug–mediator–disease paths. Prior SemMedDB-based repurposing forms such paths and filters out inaccurately extracted predications, but keeps each predicate's meaning [36]. A correctly extracted predicate can still be the wrong answer to the repurposing question: `TREATS` is recorded even when a drug is merely given to patients (Section 2.3). EviAlign instead labels every edge with the direction the repurposing question needs. The second is *traceability*. Knowledge graph completion learns embeddings from graph structure [9,14], so a prediction cannot be traced to the sentences supporting it. Every EviAlign score decomposes into named mediator paths with cited sentences. The third is *direction awareness*. Network proximity, path counts and LLM recall [4,37,38] reward drugs that are close to, connected with or often mentioned alongside a disease, not drugs whose evidence shows benefit. Network proximity flagged drugs that were later associated with lower disease risk in patient data and others associated with higher risk [37], and 15 of the top 100 epilepsy predictions in Hetionet were seizure-inducing [4]. EviAlign scores only paths whose evidence establishes a beneficial direction. The signs also make the graph *self-verifiable*: where a drug reaches a disease both directly and through a mediator,

the two signed routes can be checked against each other without a gold standard, a test unavailable to unsigned graphs.

The broader implication is that provenance in literature-derived knowledge graphs can be used as an active source of evidence rather than only as a record of where an edge originated. Any literature-derived graph that retains its supporting publications can, in principle, be re-interpreted for the downstream task rather than assuming that its original relation semantics are sufficient. Likewise, signed graphs containing redundant direct and compositional routes permit consistency checks analogous to the triangle test used here. External benchmarking, controlled path evaluation, and blinded expert assessment remain necessary because internal consistency alone cannot establish therapeutic correctness, but the combination provides complementary evidence for identifying and prioritizing mechanism-grounded repurposing hypotheses.

Re-interpretation is a one-time cost per graph. The six disease subgraphs required about 1.0 million LLM calls, which took 2.3 h on seven NVIDIA A40 GPUs (NVIDIA Corporation, Santa Clara, CA, USA) with batched inference in vLLM (version 0.13.0) [39]. Larger graphs remain feasible. Each verdict is a single token, judgements of edge–publication pairs that recur across subgraphs (29.7% here) can be cached, and the calls are independent and parallelize across GPUs. For much larger graphs, the judge's verdicts could train a smaller classifier that handles most edges, with the LLM reserved for uncertain cases.

3.2. Limitations

Several limitations bound the study's conclusions. First, the study includes six diseases. The ranking results therefore rely primarily on the paired matched-budget comparison rather than on a large number of independent disease replicates; with six diseases, a sign test attains a minimum p-value of 0.016. Evaluation across additional diseases and therapeutic areas will be needed to establish the generality of the observed ranking gains.

Second, sign consistency measures agreement between independent graph routes, not biological correctness, and is available only where the necessary drug–mediator–disease triangle exists. The 2537 evaluable triangles are also concentrated among better-studied drugs. We therefore treat sign consistency as an intrinsic validation layer rather than a substitute for external validation and retain the curated benchmark, controlled path evaluation, and blinded expert assessment alongside it.

Third, existing curated resources provide an incomplete benchmark for the setting in which literature-based repurposing is potentially most useful. Drugs with no direct disease literature are unlikely by construction to appear as established drug–disease pairs in a curated resource. Consequently, absolute benchmark performance is low in the `twohop_only` regime, and many of its highest-ranked candidates cannot be classified as true or false discoveries from existing gold standards. These candidates should therefore be interpreted as mechanism-grounded hypotheses requiring subsequent experimental or clinical investigation, rather than as validated repurposing successes.

Fourth, re-interpretation depends on a language-model judge. Qwen2.5-7B-Instruct produces sharply peaked token log-probabilities that do not provide a useful graded confidence measure, so EviAlign abstains through the `unclear` label rather than through a tunable confidence threshold. The principal findings were nevertheless robust to the choice of judge. Repeating re-interpretation with Llama-3.1-8B-Instruct (4-bit AWQ) again produced predominantly withdrawn directions (51–80% of mapped edges), with reversal remaining below 4%. Among edges assigned a direction by both models, agreement was 98% for drug–disease and drug–mediator edges and 93% for mediator–disease edges. The second judge also reproduced the principal structural results: sign consistency was 72.5% across 3836 triangles, 21.7 standard deviations above its shuffled-label control, and mediator

reliability predicted held-out triangle consistency with AUC = 0.816. Prompt wording changes how often the judge commits to a direction more than which direction it chooses (Appendix D).

Fifth, the LLM judge was pretrained on web-scale text that may include some of the abstracts it reads. Its verdicts could therefore partly reflect memorized knowledge, which would make the evaluation slightly optimistic. Three features limit this concern. The judge is instructed to use only the given passage. It sees one sentence and one claim, not the drug–disease question being ranked. And it withdraws most edges rather than filling in directions it might know. A judge whose training data end before the evaluated literature would test this directly.

Finally, EviAlign inherits entity recognition and linking errors from the source graph. SemRep's authors identify entity recognition and normalization as a major source of extraction error [21]; re-interpretation can withdraw an unsupported edge involving a mislinked entity but cannot repair the underlying entity assignment. Future work should therefore combine evidence re-interpretation with improved entity resolution and evaluate the framework across additional diseases.

4. Materials and Methods

The EviAlign workflow comprises construction of disease-centered literature graphs, evidence-grounded edge re-interpretation and scoring, and complementary internal, external, and expert validation. Appendix A is a terminology table for the terms used throughout the paper.

4.1. Disease-Centered Literature Subgraphs

For each target disease z we built a *disease subgraph* from the SemMedDB-derived edge table [20], which stores subject–predicate–object relations with concept identifiers, semantic types and, for each edge, the PMIDs of its *supporting publications*. Those PMIDs let us retrieve each publication's PubMed title and abstract, and these titles and abstracts are the only evidence EviAlign uses.

Each disease subgraph has three kinds of node. A *drug* d is a candidate therapeutic agent, a *mediator* m is an intermediate biomedical entity such as a gene, protein, cytokine or pathway, and z is the target disease. Correspondingly, three *edge families* are retained: drug–disease edges $(d \to z)$, which we call *direct edges*, drug–mediator edges $(d \to m)$ and mediator–disease edges $(m \to z)$. A *mediator path* is the pair of edges $d \to m$ and $m \to z$ through one mediator, which asserts that the drug acts on the disease through m. Each edge keeps the *predicate* SemRep extracted for it, such as `TREATS`; the word predicate is used only for these extracted relations, never for the labels EviAlign assigns. We studied six diseases spanning distinct pathologies—Alzheimer's disease (neurodegenerative), type 2 diabetes mellitus (metabolic), heart failure (cardiovascular), asthma (respiratory), rheumatoid arthritis and ulcerative colitis (autoimmune)—giving 68,689 edges supported by 167,947 edge–publication pairs in total.

We partition the candidate drugs of each disease into three *candidate regimes* by the support available to them, exhaustively and exclusively:

- `direct_only`: the drug has at least one direct edge $d \to z$ but no mediator path.
- `twohop_only`: the drug has no direct edge $d \to z$, but at least one mediator path $d \to m \to z$.
- `overlap`: the drug has both, that is, at least one direct edge and at least one mediator path to the same disease.

The partition matters for two reasons. `twohop_only` is the strictest test of mechanism-level reasoning, because those candidates have no direct edge and any ranking they receive comes from mediator paths alone. `overlap` is the regime in which the graph makes the same assertion twice, once through the direct edge and once through a mediator path, which makes the intrinsic validation of Section 4.8 possible. Results are reported per regime, and the candidate drugs of one disease within one regime are called its *candidate pool*.

4.2. Signed, Task-Oriented Edge Semantics

Every edge is assigned a *label* from a set of three per edge family, two *directions* and `unclear`: $(d \rightarrow m)$ from $\{\texttt{inhibits}^-, \texttt{activates}^+, \texttt{unclear}\}$, $(m \rightarrow z)$ from $\{\texttt{harmful_for}^-, \texttt{protective_for}^+, \texttt{unclear}\}$, and $(d \rightarrow z)$ from $\{\texttt{beneficial_for}^+, \texttt{adverse_for}^-, \texttt{unclear}\}$. Each direction carries a *sign*, and `unclear` carries none. The signs mean different things in different families—a minus on $(d \rightarrow m)$ means the drug suppresses the mediator, a minus on $(m \rightarrow z)$ means the mediator aggravates the disease—but they obey one algebra.

A mediator path is *aligned* when its two signs multiply to +:

$$\operatorname{sgn}(d \rightarrow m) \cdot \operatorname{sgn}(m \rightarrow z) = \begin{cases} + & \texttt{inhibits} \times \texttt{harmful_for}, \quad \texttt{activates} \times \texttt{protective_for}, \\ - & \texttt{inhibits} \times \texttt{protective_for}, \quad \texttt{activates} \times \texttt{harmful_for}. \end{cases} \tag{1}$$

A drug that suppresses a mediator which aggravates the disease, and a drug that enhances a mediator which protects against it, are therefore treated as the same therapeutic case. The other two combinations, whose signs multiply to − in Equation (1), are *hard-negative* paths: the path is just as explicit, but it points away from benefit. A path with `unclear` on either edge is an *unclear-negative* path; it has no sign at all and is never counted as aligned.

We allow only two directions per family because the product is sensitive to nothing else. The fixed mapping of Table 5 reads `CAUSES`, `PREDISPOSES`, `COMPLICATES` and `AUGMENTS` as the same direction, and after multiplication the two-hop score could not tell them apart in any case. What separates them is the strength of the statement, which we keep in the edge weight (Section 4.4) rather than in the label: direction is carried by the sign, strength by the weight.

The same definition is used twice: it defines how paths are scored (Section 4.5) and the three path types the scoring is tested against (Table 8). Hard-negative paths are the demanding control, since their edges are as explicit and as well supported as those of aligned paths and differ only in sign.

4.3. Sentence-Grounded Re-Interpretation

Re-interpretation re-estimates the label of every edge from its supporting publications, in three steps: we select evidence sentences, judge each sentence on its own, and aggregate the verdicts up to the edge. Appendix B records the scoring constants, the exact wording of the claims put to the judge, and the aggregation rules, so that the procedure can be reproduced in full.

For each supporting publication we take the title and abstract and keep at most three *evidence sentences*. The selection heuristic favours sentences that mention both entities of the edge, that contain a directional verb, and that report numbers, and it penalises sentences taken from the title alone. The LLM, in a separate call, proposes up to three further candidates, and the two lists are merged and re-ranked (Appendix B). Each evidence sentence s receives a *sentence weight*

$$w(s) = f_{\text{rank}}(s) f_{\text{sec}}(s) f_{\text{dir}}(s) f_{\text{gr}}(s), \tag{2}$$

where f_{rank} decays with the rank at which the sentence was selected, f_{sec} favours Results and Conclusions over Background and Methods, f_{dir} is higher for statements about the two entities themselves than for indirect remarks, and f_{gr} is highest when both entities appear in the sentence. The four factors together span roughly a thirty-fold range, so a top-ranked result sentence naming both entities carries far more weight than an ungrounded background remark.

Each edge family has two competing *claims*, one per direction. For a mediator–disease edge, for example, one claim states that the mediator is harmful for the disease and the other that it is protective. We show a single sentence and a single claim to the *LLM judge* (Qwen2.5-7B-Instruct) and ask whether the sentence supports the claim, refutes it, or is insufficient to decide. The *verdict* is one of three letters, and we read it from the model's log-probabilities over those letters at temperature 0, so no free-form text has to be parsed. The verdict is the most probable letter, and a sampling temperature does not change which letter is most probable. We checked the judge against human curators: where it assigns a direction, it agrees with the manually curated Comparative Toxicogenomics Database [40] on 52 of 54 drug–mediator and 162 of 163 drug–disease publications (Appendix B.5). A sentence yields a direction only when exactly one of the two claims is supported. Asking about the two claims separately, rather than as a single three-way question, is what lets an absent direction be told apart from a contradicted one.

Verdicts are then aggregated twice, and the two levels deliberately differ. Within a publication, the evidence sentences vote for a label, weighted by Equation (2), with `unclear` competing as a third option. Across publications, only the two directions vote and `unclear` does not, so an edge is labelled `unclear` only when the two directions tie. In practice this almost always means that no publication establishes a therapeutic direction rather than that publications disagree, a distinction quantified in Section 2.3.

Every judgement is anchored to one sentence, so that each label can be traced to the sentence that supports it. Appendix D reports the corresponding ablation: judging the whole title and abstract yields fewer `unclear` labels but worse `twohop_only` rankings.

4.4. From Publication Counts to Evidence-Weighted Edges

Let $P^+(e)$ be the supporting publications of edge e that voted for its direction. The simplest choice is the *publication count*

$$w_{\text{count}}(e) = |P^+(e)|, \tag{3}$$

with $w = 0$ for `unclear` edges. An edge is *withdrawn* when its supporting publications establish neither therapeutic direction: it keeps its predicate, receives no sign, and with $w = 0$ enters no score. Withdrawal is not a judgement that the predicate is wrong; it records that the edge cannot be aligned to the direction the task needs.

Equation (3) discards the continuous information already computed in Equation (2). 78% of *signed edges*, edges whose label is a direction, are supported by exactly one publication. The count therefore assigns a single value to most of the graph, and two-hop scores, being products of two such weights, collapse further. We therefore define the *evidence weight* as the accumulated sentence weight behind the edge's direction,

$$w_{\text{evi}}(e) = \sum_{p \in P^+(e)} \sum_{s \in S(p)} w(s), \tag{4}$$

where $S(p)$ are the evidence sentences of publication p that voted for the edge's direction.

Because two-hop scores multiply two weights, applying Equation (4) at full strength squares its spread and over-sharpens mediator paths. We therefore temper the evidence weight on mediator edges only, giving the *tempered evidence weight*

$$\tilde{w}(e) = w_{\text{count}}(e)\left(\frac{w_{\text{evi}}(e)}{w_{\text{count}}(e)}\right)^{\lambda}, \qquad \lambda = \begin{cases} 0.5 & e \in (d \to m) \cup (m \to z), \\ 1 & e \in (d \to z). \end{cases} \tag{5}$$

Appendix D reports the effect of the tempering exponent λ and of replacing the path product with a harmonic mean.

4.5. Scoring Candidates from Direct and Mediator Support

Each candidate drug receives a repurposing score for the disease, made of two components. The *direct score* is the sum of the weights of the drug's direct edges labelled `beneficial_for`. The *two-hop score* is built from the drug's aligned paths, those whose signs multiply to + (Section 4.2), that is

$$(d \xrightarrow{\texttt{inhibits}} m,\ m \xrightarrow{\texttt{harmful_for}} z) \quad \text{or} \quad (d \xrightarrow{\texttt{activates}} m,\ m \xrightarrow{\texttt{protective_for}} z). \tag{6}$$

Each aligned path contributes the product of its two edge weights, and the two-hop score is the sum of these products over all aligned paths of the drug, through every mediator.

Which components exist depends on the regime, so the two are combined accordingly. Drugs in `direct_only` have no mediator path and are ranked by the direct score alone. Drugs in `twohop_only` have no direct edge and are ranked by the two-hop score alone. Drugs in `overlap` have both and are ranked by the sum. Ranking is carried out separately within each candidate pool.

The alignment rule is that of Section 4.2 and the weights are those of Section 4.4; the scoring rule introduces no further parameters.

4.6. Evaluation Design: Four Validation Layers

The preceding sections define the method; the rest of this section defines how it is evaluated. Four *validation layers* are used. The *gold benchmark*, human-curated drug–disease pairs, tests the final ranking (Section 4.7); it involves no LLM, but depends on the coverage of the curated resources. *Sign consistency* (Section 4.8) uses only the graph's own algebra and involves neither the gold benchmark nor an LLM. *Path plausibility* (Section 2.5) rates aligned, hard-negative and unclear-negative paths with an *independent LLM judge*, drawn from a different model family than the LLM judge. This layer is used only where the others cannot reach, and is always reported alongside them rather than in their place. The blinded *human expert rating* (Section 2.7) is the only layer independent of both the graph and any model.

4.7. External Validation: Matched-Budget Baseline and Gold Benchmark

To ask whether re-interpretation helps rather than merely changes the graph, we compare against the *fixed mapping* of Table 5, which assigns each edge a direction by looking up its predicate. The edges it covers are the *mapped edges*; the others, whose predicates such as `ASSOCIATED_WITH` state no direction, are unmapped. Because re-interpretation withdraws many edges, an unmatched comparison would simply reward whichever graph is larger. The *matched-budget baseline* therefore cuts the fixed-mapping graph to EviAlign's edge count within each family–direction bucket, so that both graphs use the smaller of the two counts. The random cut is redrawn 100 times. This yields a distribution of baseline scores rather than a single number. We use the fixed mapping as the baseline because it is the standard reading of SemMedDB predicates in repurposing [14,20,36]. A matched comparison therefore isolates the step EviAlign changes, whereas the methods discussed in Section 3 differ in task or supervision.

The *gold benchmark* is a set of human-curated drug–disease pairs assembled from repoDB, ClinicalTrials.gov and openFDA drug labels [41–43]. Pairs were matched by concept identifier where available and by normalised drug name otherwise. A *gold pair* is one entry of the benchmark, and a *gold drug* is a drug that has a gold pair for the disease in question. Ranking is evaluated separately within each candidate pool.

4.8. Intrinsic Validation: Sign Consistency

The signed formulation allows a validation that needs no gold benchmark. Consider a drug d with both a direct edge to disease z and a mediator path through m. The graph then asserts the therapeutic sign of the pair (d, z) twice: through the direct edge as $\mathrm{sgn}(d \to z)$, and through the mediator path as $\mathrm{sgn}(d \to m) \cdot \mathrm{sgn}(m \to z)$. We call this configuration a *triangle*, call it *sign-consistent* when the two agree, and call the share of sign-consistent triangles the graph's *sign consistency*.

Two assertions can agree by chance, and here they do so far more than half the time, because 96% of direct edges are labelled `beneficial_for`. Sign consistency is therefore compared with a *shuffled-label control*: the labels are shuffled within each edge family, which keeps each family's label distribution but breaks the link between a triangle's direct edge and its mediator path, and sign consistency is recomputed over 10,000 shuffles. Only the excess over the control is attributed to the labels. The fixed-mapping graph is compared with a control of its own, because it withdraws no edges and its labels are skewed differently. Appendix C gives the chance-agreement calculation and the shuffling procedure in full.

4.9. Protocol for Blinded Human Expert Validation

We conducted a blinded expert evaluation to test three aspects of EviAlign not fully addressed by the graph-based and benchmark analyses. Three independent domain experts rated 64 items across three instruments (T1, T2, T3) without access to EviAlign's assigned label, item condition, or ranking outcome.

T1 evaluated edge re-interpretation using 24 single-publication edges. For each item, experts read the publication title and abstract and selected which of two competing directional interpretations was supported, or *neither*. Twelve items were edges for which EviAlign reversed the direction assigned by the fixed mapping, and twelve were edges for which EviAlign withdrew the assigned direction as unsupported. For reversal items, the EviAlign and fixed-mapping interpretations were presented in randomized order. For withdrawal items, *neither* represented the EviAlign interpretation.

T2 evaluated repurposing candidate prioritization. Experts rated 20 two-hop-only drug–disease candidates as *agree*, *neutral*, or *disagree* with the statement that the drug was a reasonable candidate to investigate for the disease, after viewing the mediator paths supporting the candidate. Fifteen items were top-ranked candidates outside the curated gold set, and five bottom-ranked candidates, the lowest-ranked that still had an aligned mediator path, served as controls. T3 evaluated mediator plausibility using the same response scale. Experts judged whether a proposed mediator was a biologically plausible component of a known drug's effect in the target disease. Fifteen items were mediators identified by EviAlign but absent from the drug's direct disease-facing literature, and five mediators recovered from that literature served as controls.

Items were selected using prespecified criteria intended to make each judgment independently interpretable. Entities were required to be specifically named, and T1 was restricted to single-publication edges so that experts could review the complete evidence used for each item. Items in all three instruments were presented in random order. System labels and experimental conditions were withheld until ratings were complete. Response proportions are reported with Wilson 95% confidence intervals, together with item-level

unanimity. Inter-rater reliability was quantified using nominal Krippendorff's α for T1 and ordinal Krippendorff's α for T2 and T3; pairwise agreement was summarized by exact agreement and Cohen's κ.

5. Conclusions

EviAlign re-interprets a literature-derived graph from the publications attached to its edges. An LLM judge reads each edge's supporting publications one sentence at a time, assigns the task's direction where the publications establish it and withdraws the edge where they do not, and the accumulated sentence weight replaces the publication count as the edge weight. On six diseases and 68,689 edges the resulting graph ranks gold drug–disease pairs above the matched-budget baseline in all three candidate regimes, including the regime with no direct edges, and every score decomposes into named mediator paths.

Two conclusions reach beyond EviAlign. The direction a repurposing method needs is in the supporting publications, not in the predicate: the same predicate covers edges whose publications report benefit, harm or no outcome, and only reading them tells these apart. And a signed graph can be checked without a gold benchmark. Agreement between direct edges and mediator paths rose from 52.5% to 75.0%, 16 percentage points above a shuffled-label control, in a check that involves no gold benchmark and no independent LLM judge, and that agreement localises to mediators whose pharmacology is unidirectional.

Author Contributions: Conceptualization, X.H., X.G., C.Y., B.N. and C.S.M.; methodology, X.H., X.G., C.Y., B.N. and C.S.M.; software, X.H.; validation, X.H.; formal analysis, X.H. and C.S.M.; investigation, X.H. and C.S.M.; writing—original draft preparation, X.H. and C.S.M.; writing—review and editing, X.H. and C.S.M.; visualization, X.H.; supervision, C.S.M. All authors have read and agreed to the published version of the manuscript.

Funding: The work was supported by National Institutes of Health grant R35GM152245 (C.S.M.), National Science Foundation grant 1944247 (C.S.M.), and Chan Zuckerberg Initiative grant 253558 (C.S.M.). The funding sources had no role in study design, data collection, analysis or interpretation of the data, writing of the report, or the decision to submit the article for publication.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

CI	Confidence interval
CUI	Concept unique identifier
KG	Knowledge graph
LLM	Large language model
PMID	PubMed identifier
UMLS	Unified Medical Language System

Appendix A. Terminology

Table A1 lists the terms used throughout the paper, with the section in which each is defined.

Table A1. Terms used throughout the paper, with the section in which each is defined. No synonym is used for any of them.

Term	Meaning	Section
disease subgraph	The edges retained for one target disease from the SemMedDB-derived edge table, in three edge families.	Section 4.1
edge family	Drug–disease, drug–mediator or mediator–disease.	Section 4.1
mediator path	The two edges drug → mediator and mediator → disease through one mediator.	Section 4.1
candidate regime	Partition of a disease's candidate drugs by their support: `direct_only`, `twohop_only`, `overlap`. Rankings and metrics are computed within a regime.	Section 4.1
supporting publications	The PubMed publications behind an edge; their titles and abstracts are the only evidence used.	Section 4.1
predicate	The relation SemRep extracted for an edge, such as `TREATS`.	Section 4.1
direction	A signed label: `inhibits/activates`, `harmful_for/protective_for`, `beneficial_for/adverse_for`.	Section 4.2
`unclear` (withdrawn edge)	The label meaning the supporting publications establish no therapeutic direction; the edge keeps its predicate but receives no sign and no weight.	Section 4.2
aligned/hard-negative/unclear-negative path	A mediator path whose two signs multiply to +; to −; or that has `unclear` on at least one edge.	Section 4.2
fixed mapping	The predicate-to-direction table (Table 5); the edges it covers are the mapped edges.	Section 4.7
re-interpretation	Re-estimating an edge's direction from its supporting publications, one sentence and one claim at a time.	Section 4.3
LLM judge	Qwen2.5-7B-Instruct, which returns support, refute or insufficient for each claim.	Section 4.3
confirm/withdraw/reverse	Re-interpretation's outcome on a mapped edge: the fixed mapping's direction, no therapeutic direction (`unclear`), or the opposite direction.	Section 2.3
evidence weight	The accumulated sentence weight behind an edge's direction, used in place of the publication count.	Section 4.4
matched-budget baseline	The fixed-mapping graph cut to EviAlign's edge count per family and direction, redrawn 100 times.	Section 4.7
gold benchmark/gold pair	Curated drug–disease pairs from repoDB, ClinicalTrials.gov and openFDA; a gold drug has a gold pair for the disease.	Section 4.7
sign consistency	Share of triangles (a drug with both a direct edge and a mediator path to the disease) whose direct sign equals the product of the path signs; compared with a shuffled-label control.	Section 4.8
mediator reliability	Share of a mediator's triangles that are sign-consistent.	Section 2.2
path plausibility	Rating of aligned, hard-negative and unclear-negative paths by an independent LLM judge from a different model family.	Section 2.5
human expert rating	The blinded evaluation by human experts: three instruments, 64 items.	Section 2.7

Appendix B. Implementation Details of Sentence-Grounded Re-Interpretation

This appendix records the constants and decision rules behind Section 4.3.

Appendix B.1. Sentence Selection

Within each supporting publication, every sentence of the title and abstract is scored by the increments of Table A2, and the highest-scoring sentences are merged with up to three sentences proposed by the LLM in a separate call. The merged list is ranked by how many entities of the edge a sentence names, then by its directional terms, then by its section, and the top three are retained as evidence sentences. The directional trigger list contains terms such as *increase*, *reduce*, *elevated*, *activate*, *inhibit*, *suppress*, *block*, *promote*, *protective* and *worsen*, together with statistical markers such as *odds ratio*, *hazard ratio* and *significant*.

Table A2. Increments used to select evidence sentences. The extra bonus when both entities occur makes a sentence that names the whole edge score above two sentences that each name one endpoint.

Condition	Increment
Start entity of the edge occurs in the sentence	+3
End entity of the edge occurs in the sentence	+3
Both entities occur (additional bonus)	+4
Each directional trigger term	+1
Numeric expression (n=, decimal, fold change, p-value)	+2
Sentence taken from the title alone	−1.5

Appendix B.2. Sentence Weights

The four factors of Equation (2) take the values in Table A3. Their product ranges from 1.00, for a top-ranked direct statement in the Section 2 naming both entities, to $0.65 \times 0.35 \times 0.55 \times 0.25 \approx 0.03$ for a third-ranked, indirect, methodological sentence naming neither. Grounding is the sharpest of the four: a sentence in which neither entity is named retains only a quarter of its weight.

Table A3. Multiplicative factors in Equation (2).

Factor	Setting	Value
f_{rank} (selection rank)	first selected	1.00
	second selected	0.80
	third and later	0.65
f_{sec} (section)	Results	1.00
	Conclusions	0.95
	Title	0.75
	section not identified	0.70
	other	0.65
	Objective	0.55
	Background	0.45
	Methods	0.35
f_{dir} (directness)	direct statement about the two entities	1.00
	indirect statement	0.90
	not identified	0.75
	background remark	0.55
f_{gr} (grounding)	both entities named in the sentence	1.00
	one entity named	0.80
	neither entity named	0.25

Appendix B.3. Claims and the Sentence-Level Decision Rule

For an edge between entities A and B, the two competing claims are *"A inhibits B"* and *"A activates B"* for drug–mediator edges, *"A is harmful for B"* and *"A is protective for B"* for mediator–disease edges. For drug–disease edges they are *"A is beneficial for B"* and *"A is adverse for B"*. Each claim is presented with a one-sentence definition of its scope: inhibition and activation cover expression, activity, signalling, abundance, phosphorylation and functional effect; harmful and protective cover disease onset, progression, severity and pathology; beneficial and adverse cover treatment, alleviation and worsening of disease-relevant outcomes.

The judge returns *support*, *refute* or *insufficient* for each claim independently, and the pair of verdicts is mapped to a sentence label by Table A4. Because the two claims are asked separately, a sentence that establishes nothing (*insufficient* for both) is distinguishable from one that supports both directions, even though both outcomes are recorded as `unclear`.

Table A4. Mapping from the two claim verdicts to a sentence label. Direction 1 and direction 2 denote the negative and positive direction of the edge family concerned.

Verdict on Claim 1	Verdict on Claim 2	Sentence Label
support	not support	direction 1
not support	support	direction 2
support	support	`unclear`
not support	not support	`unclear`

Appendix B.4. Aggregation

Sentence labels are aggregated to a publication label by weighted voting. For each of the three labels, the weights of the sentences carrying it are summed, and the label with the largest sum is assigned to the publication. `unclear` competes on equal terms at this level, and an exact tie is resolved to `unclear`.

Publication labels are aggregated to an edge label by counting publications. Let n_1 and n_2 be the numbers of publications labelled with each direction; publications labelled `unclear` are not counted. The edge takes the direction with the larger count, and `unclear` when $n_1 = n_2$. This includes the case $n_1 = n_2 = 0$, in which no publication establishes a therapeutic direction, and which accounts for almost all withdrawals reported in Section 2.3.

Appendix B.5. Accuracy Against Human-Curated Directions

We compared the judge's publication labels with the Comparative Toxicogenomics Database (CTD) [40]. CTD is curated by hand: professional curators read each publication and record whether a chemical increases or decreases the expression or activity of a gene, and whether a chemical is therapeutic for a disease or implicated in causing it. Each record keeps the PubMed identifier of its source. We read these records as `activates` or `inhibits` and as `beneficial_for` or `adverse_for`. Drugs were matched to CTD chemicals by name and synonym, genes by Entrez identifier and diseases by MeSH identifier. We compared every publication that supports an edge in our graph, has a CTD record for the same drug and target, and received a direction from the judge. The judge agrees with CTD on 52 of 54 drug–mediator publications (96.3%) and on 162 of 163 drug–disease publications (99.4%).

Appendix C. Chance Agreement and the Shuffled-Label Control

If the direct edge of a triangle is positive with probability a and its mediator path is positive with probability b, two unrelated assertions still agree with probability

$ab + (1 - a)(1 - b)$. For $a = 0.96$ this is $0.04 + 0.92\,b$, so comparing sign consistency against 50% would credit the method with agreement that the label distribution supplies for free.

The shuffled-label control measures that agreement empirically. Within each edge family, the labels are randomly permuted across the edges of that family. This keeps the family's label distribution intact, so the shuffled direct edges are still 96% `beneficial_for`, but it breaks the correspondence between a triangle's direct edge and its mediator path. Sign consistency is recomputed over all triangles after each of 10,000 shuffles, and the mean and standard deviation of those values are the control reported in Section 2.2.

The graph labelled by the fixed mapping (Table 5) needs a control of its own. It withdraws no edges, so its triangle set is much larger, and its labels are less skewed, so the agreement that chance alone produces is different. Its labels are therefore shuffled on its own triangle set, and each graph is compared with its own shuffled-label control.

Appendix D. Additional Ablations and Robustness Checks

Evidence scope. We re-ran the full re-interpretation of all 167,947 edge–publication pairs under two alternative evidence scopes and pushed each through the entire pipeline. Judging the complete title and abstract in a single call lowered the `unclear` rate substantially, by 20.6 percentage points for direct edges and 18.6 for drug–mediator edges. Downstream ranking did not improve. The `twohop_only` mean reciprocal rank, the inverse rank of the first gold drug, declined in all six diseases (sign test $p = 0.031$; Figure 6C). Selecting evidence sentences by entity mention rather than by the selection score of Appendix B was worse still. This ablation also shows that a lower `unclear` rate does not by itself indicate a better graph.

Robustness of the LLM judge. The verdict is the most probable of the three answer letters, read from the log-probabilities of a single generated token, so a sampling temperature does not change it. We tested run-to-run variation, prompt wording and the number of evidence sentences on 1000 randomly drawn edges per edge family, re-interpreting all 9681 of their supporting publications with the same model. An identical re-run reproduced 99.3% of publication labels. Rewording every prompt with the same meaning changed 14.8% of them, and allowing five evidence sentences instead of three changed 6.3%. Almost all of these changes are between a direction and `unclear`: where both runs assign a direction, they disagree on 3 of 1801 and 2 of 2062 publications. Wording and the sentence cap therefore move the point at which the judge commits to a direction, not the direction it chooses.

Mediator reliability in scoring. Reliability, recomputed without each candidate's own triangles, was used as a weight on each aligned path and, separately, as a filter that removes paths through mediators whose reliability is below 0.5. Neither improved the gold-benchmark ranking: `overlap` average precision changed by at most 0.002, and `twohop_only` Precision@10 fell from 0.050 to 0.033 under both. Three quarters of the aligned paths whose reliability can be estimated already pass through mediators with reliability above 0.81.

Weighting. The two choices in Equation (5) and in the two-hop score were varied with one scoring implementation applied to every arm, so the differences rather than the absolute values are the quantity of interest. Applying the evidence weight at full strength to mediator edges ($\lambda = 1$) costs 0.017 of `twohop_only` Precision@10 and 0.005 of `overlap` average precision relative to $\lambda = 0.5$, which is the value used throughout. Replacing the product of the two edge weights with their harmonic mean costs 0.050 of `overlap` Precision@10 and 0.017 of `twohop_only` Precision@10, so the product is retained.

Numeric channel. We also evaluated a second channel that extracted sample size and statistical significance from the same supporting publications with a TrialSieve-derived token classifier [31]. Its contribution proved to be largely subsumed by evidence weighting: once the sentence weights set the edge weight directly, the additional numeric factor leaves

gold-drug rank essentially unchanged. EviAlign therefore uses the evidence weight alone, which keeps it free of a second extraction model without loss of ranking quality.

Appendix E. Withdrawn and Reversed Mapped Edges with Their Supporting Publications

Table A5 lists eight direct edges that carry the predicate `TREATS`, which the fixed mapping of Table 5 therefore labels `beneficial_for`. Re-interpretation withdraws five of them and reverses three to `adverse_for`. Table A6 lists five edges with other mapped predicates whose direction re-interpretation changes. Both tables illustrate the cases described in Section 2.3: in every row the predicate is a faithful extraction, and the therapeutic sign is decided by information the predicate does not carry.

Table A5. Direct edges carrying the predicate `TREATS`, which the fixed mapping labels `beneficial_for`, and which re-interpretation withdraws or reverses. In the first group the drug is not a treatment for the disease; in the second it is, but the publication reports no outcome; in the third the publication reports harm. In every group the predicate is a faithful extraction, because the drug was given to patients with the disease. What differs is what the supporting publications report happened.

Edge	Pubs.	What the Supporting Publications Say	Why the Fixed-Mapping Direction Does Not Hold
Withdrawn: the drug is not a treatment for the disease			
methacholine → asthma	181	it is inhaled by patients with asthma under supervision	it is a challenge agent, given to provoke bronchoconstriction for diagnosis
scopolamine → Alzheimer's	12	it is administered in Alzheimer's research	it induces the cognitive deficit, as a pharmacological model of the disease
technetium Tc-99m → Alzheimer's	1	it is given to patients with Alzheimer's disease and imaged	it is a SPECT tracer, used to measure regional brain uptake
Withdrawn: the drug is a treatment, but the publication reports no outcome			
bismuth subsalicylate → colitis	1	fifteen patients with refractory colitis were treated with it	the sentence describes the enema protocol, not whether it worked
Janus kinase inhibitor → Alzheimer's	1	it is named as a promising candidate for repurposing in the disease	a proposal for future work, not a reported result
Reversed to `adverse_for`: the publication reports harm			
propranolol → asthma	28	FEV_1 fell and airway resistance rose after propranolol in patients with asthma; β-blockers are avoided in asthma	the publication measures the bronchoconstriction the drug causes; 27 of the 28 publications yield no direction
NSAIDs → heart failure	15	NSAID use was associated with increased mortality and a doubled risk of hospitalisation for heart failure	the publications report the drug precipitating the disease, not treating it
antipsychotic agents → Alzheimer's	58	restrictive use is recommended because of increased cerebrovascular events and mortality	the publication that reports an outcome reports harm; the remaining 57 yield no direction

Table A6. Mapped edges with predicates other than `TREATS` whose direction re-interpretation changes. In each case the predicate is a correct reading of the sentence it was extracted from, and the therapeutic sign depends on a distinction that a single predicate cannot carry. Pubs. gives the number of supporting publications yielding the re-interpreted direction, the fixed-mapping direction, and no direction. Arrows point from the subject to the object of the edge.

Edge	Predicate (Fixed Mapping)	Re-Interpreted	What the Supporting Publications Say	Pubs.
vitamin D → type 2 diabetes	`PREDISPOSES` (adverse_for)	beneficial_for	low vitamin D status is a risk factor; the evidence suggests a protective role of vitamin D	2/0/15
NSAIDs → Alzheimer's	`PREDISPOSES` (adverse_for)	beneficial_for	the risk of Alzheimer's disease in NSAID users is markedly reduced	2/0/9
TREM2 → Alzheimer's	`CAUSES` (harmful_for)	protective_for	loss-of-function variants are associated with the disease, indicating a protective role of TREM2	1/0/4
GRK5 → heart failure	`AUGMENTS` (harmful_for)	protective_for	GRK5 is upregulated in heart failure; a gain-of-function polymorphism is associated with improved outcomes	1/0/2
antioxidants → NF-κB	`STIMULATES` (activates)	inhibits	the protective mechanism of antioxidants is partly due to inhibition of NF-κB activation	27/2/28

Drug labels confirm the withdrawals and the reversals of Section 2.3 (Table A7). DrugCentral 2023 [44] and SIDER 4.1 [45] record each drug's indications, contraindications and side effects by UMLS identifier. A drug–disease pair has a *benefit record* if the disease is a labelled or off-label indication of the drug, and a *harm record* if it is a contraindication or a listed side effect. Among the 3394 direct edges whose drug is in DrugCentral, `TREATS` edges confirmed as `beneficial_for` have a benefit record in 31.2% of cases and withdrawn edges in 6.7% (odds ratio 6.3, $p < 10^{-60}$). At equal publication count the two groups have the same share of phase 2 or later trials on ClinicalTrials.gov, so the withdrawn drugs were given and trialled, but not shown to help. Edges labelled `adverse_for` have a harm record in 35.6% of cases against 13.9% for `beneficial_for` (odds ratio 3.4, $p = 3 \times 10^{-4}$). The fixed mapping's `adverse_for` edges, from `CAUSES` and its kin, have a harm record in 10.2%, no more than its `beneficial_for` edges (12.0%). Label records are incomplete, so these are comparisons between groups of edges, not accuracies.

Table A7. Drug-label reference check of the direct drug–disease edges whose drug is in DrugCentral (3394 edges, 1530 drugs). A *harm record* means the disease is a contraindication or a listed side effect of the drug; a *benefit record* means it is a labelled or off-label indication. (**A**) All direct edges, grouped by the label each method assigns. (**B**) Edges carrying the predicate `TREATS`, grouped by the outcome of re-interpretation; the last column is the share of pairs with at least one interventional trial of phase 2 or higher on ClinicalTrials.gov.

(**A**) Direct edges by assigned label

Method	Label	Edges	Harm Record	Benefit Record
Re-interpretation	`adverse_for`	45	35.6%	6.7%
	`beneficial_for`	1318	13.9%	35.1%
	`unclear`	2031	12.7%	12.9%
Fixed mapping	`adverse_for`	59	10.2%	22.0%
	`beneficial_for`	2798	12.0%	17.0%
	unmapped	537	21.0%	44.3%

(**B**) `TREATS` edges by re-interpretation outcome

Outcome	Edges	Benefit Record	Harm Record	Phase ≥ 2 Trial
Confirmed `beneficial_for`	1116	31.2%	12.8%	58.0%
Withdrawn to `unclear`	1542	6.7%	11.0%	36.6%
Reversed to `adverse_for`	38	5.3%	34.2%	57.9%

References

1. Ashburn, T.T.; Thor, K.B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **2004**, *3*, 673–683. [CrossRef]
2. Pushpakom, S.; Iorio, F.; Eyers, P.A.; Escott, K.J.; Hopper, S.; Wells, A.; Doig, A.; Guilliams, T.; Latimer, J.; McNamee, C.; et al. Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **2019**, *18*, 41–58. [CrossRef]
3. Li, J.; Zheng, S.; Chen, B.; Butte, A.J.; Swamidass, S.J.; Lu, Z. A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **2016**, *17*, 2–12. [CrossRef]
4. Himmelstein, D.S.; Lizee, A.; Hessler, C.; Brueggeman, L.; Chen, S.L.; Hadley, D.; Green, A.; Khankhanian, P.; Baranzini, S.E. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* **2017**, *6*, e26726. [CrossRef]
5. Zhu, Y.; Che, C.; Jin, B.; Zhang, N.; Shen, X.; Wei, X.; Hu, B.; Wang, X. Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Inform. J.* **2020**, *26*, 2737–2750. [CrossRef]
6. Wang, Z.; Zhou, M.; Arnold, C. Toward heterogeneous information fusion: Bipartite graph convolutional networks for in silico drug repurposing. *Bioinformatics* **2020**, *36*, i525–i533. [CrossRef]
7. Sadegh, S.; Skelton, J.; Anastasi, E.; Bernett, J.; Blumenthal, D.B.; Galindez, G.; Salgado-Albarran, M.; Lazareva, O.; Flanagan, K.; Cockell, S.; et al. Network medicine for disease module identification and drug repurposing with the NeDRex platform. *Nat. Commun.* **2021**, *12*, 6848. [CrossRef]
8. Bang, D.; Lim, S.; Lee, S.; Kim, S. Biomedical knowledge graph learning for drug repurposing by extending guilt-by-association to multiple layers. *Nat. Commun.* **2023**, *14*, 3570. [CrossRef]
9. Huang, K.; Chandak, P.; Wang, Q.; Havaldar, S.; Vaid, A.; Leskovec, J.; Nadkarni, G.N.; Glicksberg, B.S.; Gehlenborg, N.; Zitnik, M. A foundation model for clinician-centered drug repurposing. *Nat. Med.* **2024**, *30*, 3601–3613. [CrossRef]
10. Brown, A.S.; Patel, C.J. A review of validation strategies for computational drug repositioning. *Brief. Bioinform.* **2018**, *19*, 174–177. [CrossRef]
11. Perdomo-Quinteiro, P.; Belmonte-Hernández, A. Knowledge graphs for drug repurposing: A review of databases and methods. *Brief. Bioinform.* **2024**, *25*, bbae461. [CrossRef]
12. Gonzalez-Cavazos, A.C.; Tanska, A.; Mayers, M.; Carvalho-Silva, D.; Sridharan, B.; Rewers, P.A.; Sankarlal, U.; Jagannathan, L.; Su, A.I. DrugMechDB: A curated database of drug mechanisms. *Sci. Data* **2023**, *10*, 632. [CrossRef]
13. Jiménez, A.; Merino, M.J.; Parras, J.; Zazo, S. Explainable drug repurposing via path-based knowledge graph completion. *Sci. Rep.* **2024**, *14*, 16587. [CrossRef]
14. Sosa, D.N.; Derry, A.; Guo, M.; Wei, E.; Brinton, C.; Altman, R.B. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. *Pac. Symp. Biocomput.* **2020**, *25*, 463–474. [CrossRef]
15. Malas, T.B.; Vlietstra, W.J.; Kudrin, R.; Starikov, S.; Charrout, M.; Roos, M.; Peters, D.J.M.; Kors, J.A.; Vos, R.; 't Hoen, P.A.C.; et al. Drug prioritization using the semantic properties of a knowledge graph. *Sci. Rep.* **2019**, *9*, 6281. [CrossRef]

16. Kirkpatrick, A.; Onyeze, C.; Kartchner, D.; Allegri, S.; An, D.N.; McCoy, K.; Davalbhakta, E.; Mitchell, C.S. Optimizations for computing relatedness in biomedical heterogeneous information networks: SemNet 2.0. *Big Data Cogn. Comput.* **2022**, *6*, 27. [CrossRef]
17. McCoy, K.; Gudapati, S.; He, L.; Horlander, E.; Kartchner, D.; Kulkarni, S.; Mehra, N.; Prakash, J.; Thenot, H.; Vanga, S.V.; et al. Biomedical text link prediction for drug discovery: A case study with COVID-19. *Pharmaceutics* **2021**, *13*, 794. [CrossRef]
18. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **1986**, *30*, 7–18. [CrossRef]
19. Cameron, D.; Bodenreider, O.; Yalamanchili, H.; Danh, T.; Vallabhaneni, S.; Thirunarayan, K.; Sheth, A.P.; Rindflesch, T.C. A graph-based recovery and decomposition of Swanson's hypothesis using semantic predications. *J. Biomed. Inform.* **2013**, *46*, 238–251. [CrossRef]
20. Kilicoglu, H.; Shin, D.; Fiszman, M.; Rosemblat, G.; Rindflesch, T.C. SemMedDB: A PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **2012**, *28*, 3158–3160. [CrossRef]
21. Kilicoglu, H.; Rosemblat, G.; Fiszman, M.; Shin, D. Broad-coverage biomedical relation extraction with SemRep. *BMC Bioinform.* **2020**, *21*, 188. [CrossRef]
22. Kilicoglu, H.; Rosemblat, G.; Rindflesch, T.C. Assigning factuality values to semantic relations extracted from biomedical research literature. *PLoS ONE* **2017**, *12*, e0179926. [CrossRef]
23. Nian, Y.; Hu, X.; Zhang, R.; Feng, J.; Du, J.; Li, F.; Bu, L.; Zhang, Y.; Chen, Y.; Tao, C. Mining on Alzheimer's diseases related knowledge graph to identity potential AD-related semantic triples for drug repurposing. *BMC Bioinform.* **2022**, *23*, 407. [CrossRef]
24. Ratajczak, F.; Joblin, M.; Ringsquandl, M.; Hildebrandt, M. Task-driven knowledge graph filtering improves prioritizing drugs for repurposing. *BMC Bioinform.* **2022**, *23*, 84. [CrossRef]
25. Vasilakes, J.; Rizvi, R.; Melton, G.B.; Pakhomov, S.; Zhang, R. Evaluating active learning methods for annotating semantic predications. *JAMIA Open* **2018**, *1*, 275–282. [CrossRef]
26. Jeynes, J.C.G.; James, T.; Corney, M. Natural language processing for drug discovery knowledge graphs: Promises and pitfalls. In *Methods in Molecular Biology*; Springer: New York, NY, USA, 2024; Volume 2716, pp. 223–240. [CrossRef]
27. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2020**, *36*, 1234–1240. [CrossRef]
28. Nye, B.; Li, J.J.; Patel, R.; Yang, Y.; Marshall, I.; Nenkova, A.; Wallace, B. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 197–207. [CrossRef]
29. Sanchez-Graillet, O.; Witte, C.; Grimm, F.; Cimiano, P. An annotated corpus of clinical trial publications supporting schema-based relational information extraction. *J. Biomed. Semant.* **2022**, *13*, 14. [CrossRef]
30. Neumann, M.; King, D.; Beltagy, I.; Ammar, W. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*; Association for Computational Linguistics: Florence, Italy, 2019; pp. 319–327. [CrossRef]
31. Kartchner, D.; Turner, H.; Ye, C.; Al-Hussaini, I.; Nursal, B.; Lee, A.J.B.; Deng, J.; Curtis, C.; Cho, H.; Duvaris, E.L.; et al. TrialSieve: A comprehensive biomedical information extraction framework for PICO, meta-analysis, and drug repurposing. *Bioengineering* **2025**, *12*, 486. [CrossRef]
32. Yu, H.; Fan, L.; Li, L.; Zhou, J.; Ma, Z.; Xian, L.; Hua, W.; He, S.; Jin, M.; Zhang, Y.; et al. Large language models in biomedical and health informatics: A review with bibliometric analysis. *J. Healthc. Inform. Res.* **2024**, *8*, 658–711. [CrossRef]
33. Chen, Q.; Hu, Y.; Peng, X.; Xie, Q.; Jin, Q.; Gilson, A.; Singer, M.B.; Ai, X.; Lai, P.T.; Wang, Z.; et al. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nat. Commun.* **2025**, *16*, 3280. [CrossRef]
34. Bodenreider, O. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res.* **2004**, *32*, D267–D270. [CrossRef]
35. Lai, P.-T.; Wei, C.-H.; Tian, S.; Leaman, R.; Lu, Z. Enhancing biomedical relation extraction with directionality. *arXiv* **2025**, arXiv:2501.14079. [CrossRef]
36. Zhang, R.; Hristovski, D.; Schutte, D.; Kastrin, A.; Fiszman, M.; Kilicoglu, H. Drug repurposing for COVID-19 via knowledge graph completion. *J. Biomed. Inform.* **2021**, *115*, 103696. [CrossRef]
37. Cheng, F.; Desai, R.J.; Handy, D.E.; Wang, R.; Schneeweiss, S.; Barabási, A.-L.; Loscalzo, J. Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **2018**, *9*, 2691. [CrossRef]
38. Yan, C.; Grabowska, M.E.; Dickson, A.L.; Li, B.; Wen, Z.; Roden, D.M.; Stein, C.M.; Embí, P.J.; Peterson, J.F.; Feng, Q.; et al. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. *npj Digit. Med.* **2024**, *7*, 46. [CrossRef]
39. Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C.H.; Gonzalez, J.; Zhang, H.; Stoica, I. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*; Association for Computing Machinery: Koblenz, Germany, 2023; pp. 611–626. [CrossRef]

40. Davis, A.P.; Wiegers, T.C.; Sciaky, D.; Barkalow, F.; Strong, M.; Wyatt, B.; Wiegers, J.; McMorran, R.; Abrar, S.; Mattingly, C.J. Comparative Toxicogenomics Database's 20th anniversary: Update 2025. *Nucleic Acids Res*. **2025**, *53*, D1328–D1334. [CrossRef]
41. Brown, A.S.; Patel, C.J. A standard database for drug repositioning. *Sci*. *Data* **2017**, *4*, 170029. [CrossRef]
42. Zarin, D.A.; Tse, T.; Williams, R.J.; Califf, R.M.; Ide, N.C. The ClinicalTrials.gov results database—Update and key issues. *N*. *Engl*. *J*. *Med*. **2011**, *364*, 852–860. [CrossRef]
43. Kass-Hout, T.A.; Xu, Z.; Mohebbi, M.; Nelsen, H.; Baker, A.; Levine, J.; Johanson, E.; Bright, R.A. openFDA: An innovative platform providing access to a wealth of FDA's publicly available data. *J*. *Am*. *Med*. *Inform*. *Assoc*. **2016**, *23*, 596–600. [CrossRef]
44. Avram, S.; Wilson, T.B.; Curpan, R.; Halip, L.; Borota, A.; Bora, A.; Bologa, C.G.; Holmes, J.; Knockel, J.; Yang, J.J.; et al. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Res*. **2023**, *51*, D1276–D1287. [CrossRef]
45. Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res*. **2016**, *44*, D1075–D1079. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.