RESEARCH | Article in Press

Patent-augmented knowledge graph drug repurposing for duchenne muscular dystrophy: a time-sliced evaluation

Fatih Uludağ[1] · **H. Eray CELİK**[1]

Received: 17 April 2026 / Accepted: 10 September 2026
© The Author(s) 2026

Abstract

Background Patent documents may complement curated biomedical resources, but adding patent-derived relations to a knowledge graph does not itself establish improved drug-repurposing prediction or mechanism. We evaluated patent augmentation in a historical, time-sliced Duchenne muscular dystrophy (DMD) knowledge-graph completion study with explicit temporal control, target-leakage removal, source ablation, label-blind multi-seed scoring, extraction-quality auditing, non-embedding baselines, and cutoff sensitivity.

Methods The leakage-safe UMLS 2018AA graph contained 47 816 edges. Adding 345 patent edges produced an Integrated graph with 48 161 edges, 41 335 global nodes, and 91 namespaced relations. The frozen universe comprised 3 684 candidate CUIs and 3 489 ranking units. TransE, DistMult, ComplEx, and RotatE were optimized on UMLS-only data, confirmed with three seeds, and refit with five seeds per scorable graph view. Four strict post-cutoff DMD clinical-development positives were evaluated under a positive-unlabeled framework. Four graph-topological rules and four prespecified cutoff dates were evaluated, and a domain expert in health sciences audited a 200-instance random sample of 541 model-included patent predications.

Results UMLS-only TransE achieved MRR 0.00556 versus 0.00295 for Integrated; Integrated improved one of four strict-positive ranks. The UMLS-only shortest-path baseline achieved MRR 0.0260 and recovered three positives in the Top 250, although ties were extensive. The domain-expert audit classified 106 predications as valid, 89 as invalid, and 5 as unresolved; decidable validity was 54.4% (Wilson 95% CI 47.4–61.2%). Integrated MRR increased from 0.00217 to 0.00295 across cutoffs, but remained below the time-aligned UMLS anchor at every cutoff. Whole-ranking Spearman correlations were 0.930–0.935, whereas Top-25 Jaccard overlap was 0.282–0.389. All ten mechanistic Top-10 units were classified as nonspecific graph-derived hypotheses; none had a candidate-specific patent mechanism or candidate-specific DMD evidence.

Conclusions Patent-derived relations materially perturbed KGE rankings but did not establish uniform retrospective improvement, causal patent contribution, or therapeutic efficacy. Extraction quality, temporal vocabulary alignment, model architecture, simple topology, and candidate identity materially affected the conclusions. Patent-augmented knowledge graphs should therefore be used as provenance-rich hypothesis-generation systems rather than treatment-validation tools.

Clinical trial number Not applicable

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.

Keywords Drug repurposing · Duchenne muscular dystrophy · Knowledge graph completion · UMLS · Patents · SemRep · Retrospective evaluation · Cutoff sensitivity · Non-embedding baselines · TransE

Background

Drug development remains lengthy, expensive, and high risk despite advances in molecular biology and computational medicine. Drug repurposing seeks new indications for compounds that are already approved or have progressed through part of the development pipeline, potentially reducing some of the time, cost, and uncertainty associated with de novo discovery [1, 2]. These advantages are especially relevant to rare diseases, where small patient populations, incomplete natural-history evidence, and limited commercial incentives constrain conventional development. Rare diseases collectively affect hundreds of millions of people worldwide [3].

Duchenne muscular dystrophy (DMD) is a severe, progressive muscle-wasting disorder caused by pathogenic variants in DMD that eliminate functional dystrophin in muscle. Disease progression includes skeletal-muscle degeneration, cardiomyopathy, and respiratory compromise. Both dystrophin-restoring strategies and interventions targeting secondary pathology remain areas of active development [4]. This combination of high unmet need, complex pathophysiology, and relatively sparse disease-specific evidence makes DMD an informative test case for computational hypothesis generation.

Data construction and biomedical knowledge sources

Biomedical repurposing datasets are assembled from sources with different evidential roles and error structures. Curated terminologies and ontologies such as UMLS harmonize concept identity and semantic relations, whereas SemRep automatically maps biomedical text to UMLS-normalized semantic predications [5, 6]. Heterogeneous resources such as PrimeKG integrate drugs, diseases, genes, phenotypes, pathways, and other curated relations at larger scale [7]. Patent corpora can add indication, formulation, experimental-context, and chemical disclosures, but their legal language, retrieval conventions, and automatic extraction errors differ from those of biomedical publications [8–11]. Patent-derived relations should therefore be treated as a distinct source layer whose incremental value and quality require direct evaluation.

Knowledge-graph and non-graph repurposing methods

Non-graph repurposing approaches use clinical observations, molecular and transcriptomic signatures, chemical or side-effect similarity, docking, genetic associations, and pathway evidence [2, 12]. Literature-based discovery searches for complementary concepts across disconnected literatures [13, 14], while semantic LBD and knowledge-graph completion retain typed relations and score plausible but unobserved links in a multi-relational graph [5–15]. Recent heterogeneous graph systems such as TxGNN emphasize broad integration and transfer across diseases, whereas path-based KGC methods emphasize interpretable graph routes [16, 17]. These approaches answer related but different questions. The present study does not claim a cross-method comparison with molecular, docking, or clinical-prediction systems; it tests the incremental effect of adding a patent-derived relation layer to an otherwise matched curated graph.

Retrospective evaluation and study positioning

Retrospective time slicing reserves post-cutoff evidence for evaluation while limiting model development to historically available knowledge [18]. SemRep-based KGC has combined time slicing with multiple embedding models for repurposing [19]. However, temporal validity depends on more than document dates: UMLS release timing, candidate identity, direct target leakage, normalization resources, model selection, scoring, tie-breaking,

endpoint mapping, and external validation must all be controlled. The endpoint is also positive-unlabeled, because absence of later clinical development does not establish a true negative.

Prior work establishes the value of biomedical knowledge graphs, semantic predications, patent mining, non-graph repurposing evidence, and temporal evaluation, but leaves a methodological gap: the incremental effect of a noisy patent layer is rarely tested against a matched curated baseline while target leakage, extraction quality, architecture dependence, simple topology, and cutoff sensitivity are assessed together. This study addresses that gap in DMD. Its aim was to determine how patent-derived augmentation changes rankings and what evidential claim such changes can legitimately support.

The study addressed five questions:

- Does the patent-augmented graph improve retrospective prioritization of drugs that subsequently entered DMD clinical development?
- Are patent-associated rank changes stable across embedding architectures?
- How do simple graph-topological baselines compare with the learned TransE ranking?
- What does manual review reveal about the quality of the frozen SemRep-derived patent layer?
- How sensitive are Integrated rankings and endpoint recovery to alternative historical cutoffs compatible with UMLS 2018AA?

Methods

Study design and temporal framework

We conducted a retrospective, time-sliced KGC study. The primary historical cutoff was 3 November 2018. DMD was represented by UMLS concept C0013264, and the associated DMD gene by C1414083. Graph construction, candidate definition, hyperparameter optimization (HPO), final model training, and label-blind scoring were completed before retrospective labels were joined. Candidate-specific patent attribution and external validation occurred only after discovery rankings were frozen (Fig. 1).

Historical data sources and patent provenance

The curated baseline graph used the UMLS 2018AA Active Subset. The patent corpus was retrieved from Google Patents Advanced Search with the query “Duchenne Muscular Dystrophy”. The preserved retrieval record identifies May 2022 as the acquisition month; a more granular day-level acquisition date is not asserted. The export contained 25 002 records; 18 226 English-language abstracts were retained and fetched through Python/Selenium.

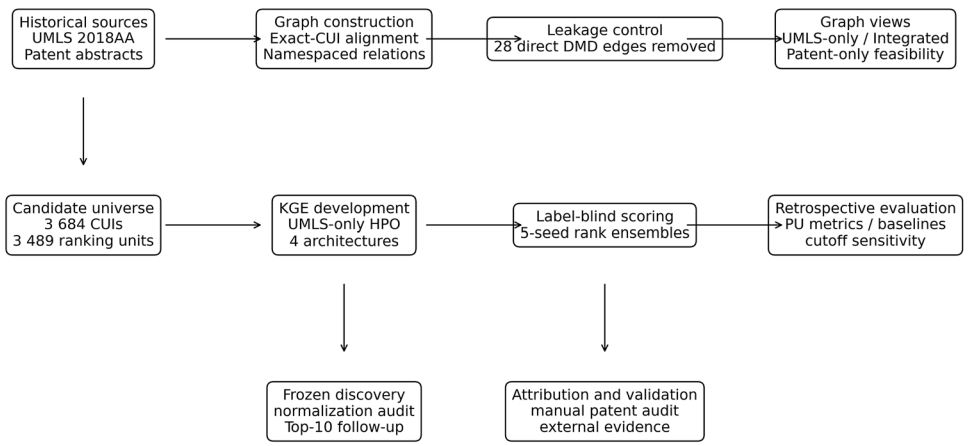


Fig. 1 Analysis workflow. Endpoint labels and candidate-specific validation were introduced only after final scores and five-seed ensembles were sealed. Patent attribution, extraction-quality auditing, and external validation were diagnostic and did not modify rankings

Local semantic extraction used SemRep 1.8 in generic processing mode with the official MetaMap 2018 runtime under OpenJDK 8. SemRep used UMLS 2018AA database and lexicon resources rather than the bundled 2018AB resources; the SKR-MedPost/tagger and word-sense-disambiguation services used ports 1795 and 5554. Historical extraction and the verified current environment use the same configuration [see Additional file 1: Table S1 and Fig. S2].

Patent temporal eligibility used publication-date metadata, not filing, priority, grant, or retrieval dates. The model-included evidence table contained 541 relation instances, which collapsed to 365 unique directed pre-leakage model triples. Twenty direct DMD therapeutic patent edges were removed, leaving 345 leakage-safe patent edges. Evidence count, source count, publication identifiers, publication dates, and internal source-edge identifiers were retained for provenance.

Data-source roles and validation boundaries

UMLS 2018AA supplied the curated baseline graph. Patent abstracts supplied an automatically generated augmentation layer: SemRep/MetaMap-derived predications entered the frozen graph without manual pre-screening. The study did not use a conventional randomly held-out DMD test set. Model selection used the common UMLS validation subset, whereas disease-level evaluation used the independently reconstructed ClinicalTrials.gov endpoint. PubMed, ClinicalTrials.gov, and regulatory records were consulted only after ranking for candidate-specific validation and did not contribute training edges. The current endpoint was not produced by SemRep or MetaMap; it used deterministic intervention-name mapping, ambiguity handling, and manual review. The later domain-expert audit quantified the quality of the frozen patent layer but did not clean the graph, alter model inputs, retrain models, or change rankings.

Entity alignment, edge identity, conflicts, and leakage control

Patent entities entered the analysis with UMLS CUIs assigned in the archived resolved layer. Cross-source alignment used exact CUI identity rather than contemporary lexical matching. Of 305 patent-catalog nodes, 293 overlapped the UMLS 2018AA catalog and 12 were patent-only nodes. UMLS and patent predicates remained source-namespaced, for example umls::may_treat and patent::treats, and source direction was preserved [see Additional file 1: Fig. S1 and Table S6].

Edge identity was one unique directed subject-predicate-object triple. Duplicate directed triples and self-loops were absent in the final views. Opposing, reverse-direction, or differently typed statements from UMLS and patents were not overwritten; they coexisted as distinct source facts. The pre-specified leakage filter removed 8 UMLS and 20 patent direct candidate-DMD therapeutic edges before model fitting.

Graph views and patent-only feasibility

Three source views were retained and are summarized in Table 1. UMLS_ONLY_SAFE and INTEGRATED_SAFE were the primary comparable scoring views. PATENT_ONLY_SAFE was retained as a source-isolation feasibility control. A literal same-task patent-only KGE arm was not defined because the UMLS-namespaced target relation umls::may_treat was absent from the patent-only vocabulary and candidate coverage was extremely sparse. Substituting patent::treats or restricting the denominator would change the estimand [see Additional file 1: Fig. S4].

Candidate universe and ranking units

The candidate universe was constructed before model fitting and independently sealed. It contained 3 684 candidate CUIs represented by 3 489 ranking units. Candidate identity was preserved at the historical UMLS concept

Table 1 Leakage-safe graph views and candidate coverage

View	Edges	Active nodes	Relations	Candidate coverage	Role
UMLS_ONLY_SAFE	47 816	41 321	69	3 684 CUIs / 3 489 units	Curated-source ablation
INTEGRATED_SAFE	48 161	41 333	91	3 684 CUIs / 3 489 units	Primary reference graph
PATENT_ONLY_SAFE	345	299	22	8 incident CUIs; 9 representatives; 1 fully covered unit	Coverage-limited feasibility

The patent-only view was not assigned same-task MRR because the target relation was absent and the full candidate denominator was not covered

level. Lexical variants represented within the same UMLS 2018AA CUI inherited that concept identity, whereas distinct generic, brand, formulation, or ingredient CUIs were not post-hoc collapsed using contemporary RxNorm or lexical resources unless such mapping was supported by the frozen historical vocabulary. Search cores used for post-ranking evidence retrieval were not graph-identity or active-ingredient mappings. Ranking-unit representatives, member CUI sets, query orientation, and deterministic tie-breaks were frozen before scoring.

Retrospective clinical-development endpoint

The primary endpoint, STRICT_NEW_POST_CUTOFF_DMD_CLINICAL_DEVELOPMENT, was rebuilt independently of model scores from a frozen ClinicalTrials.gov DMD snapshot containing 495 studies. Four ranking units met the strict-positive definition: Captopril, Tacrolimus, Bisoprolol Fumarate, and metoprolol succinate. Candidates with qualifying pre-cutoff DMD clinical history were prior-history controls rather than newly emerging positives. The remaining units were unlabeled, not verified negatives.

Training, validation, and KGE models

The UMLS graph was split into 43 314 training edges and 4 502 common validation edges. The 345 patent edges were added only to the Integrated training graph, yielding 43 659 Integrated training edges. HPO was performed exclusively on the UMLS-only training graph. The primary validation subset comprised 297 umls::may_treat edges and contained no DMD-target validation edge.

SemRep and MetaMap were historical NLP preprocessing tools used to create the patent relation layer; they were not models trained or fine-tuned in the current analysis. TransE represents a relation as a translation between head and tail embeddings. DistMult uses a diagonal bilinear score that is parsimonious but naturally symmetric for a fixed relation; ComplEx extends bilinear scoring to complex-valued embeddings and can represent asymmetric relations; and RotatE models relations as rotations in complex space, supporting patterns such as symmetry, antisymmetry, inversion, and composition [20–23]. No pretrained biomedical model or external embedding was fine-tuned. Each KGE architecture was randomly initialized and trained from scratch on the frozen graph data.

Models were implemented with Python 3.11.15, PyKEEN 1.11.1, PyTorch 2.12.1+cu126, NumPy 1.26.4, and Optuna 4.9.0. HPO used TPE sampling with seed 20,261,103 and 12 trials per model. Training used Adam, stochastic local closed-world assumption, unfiltered Bernoulli negative sampling, batch size 1 024, no synthetic inverse triples, and model-specific PyKEEN default losses. Early stopping was used only in HPO. Complete settings are provided in Additional file 1: Table S5.

Confirmatory evaluation used seeds 11, 23, and 37. TransE was selected by the frozen confirmatory rule. Final refits used seeds 11, 23, 37, 53, and 71, with no endpoint access, validation, or early stopping. Table 2 reports confirmatory performance [see Additional file 1: Fig. S3].

Label-blind scoring and rank ensembles

The prediction query was candidate subject - umls::may_treat - DMD object. Forty final KGE scoring runs were generated in the primary study: four architectures, two scorable graph views, and five seeds. Candidate ties were resolved by ascending global node index and ranking-unit ties by ascending frozen aggregation-unit identifier. The primary ensemble statistic was the arithmetic mean of five deterministic seed ranks, with lower values indicating higher priority. Raw logits were not averaged because independently trained models were not score-calibrated. Labels were joined only after all score and ensemble artifacts were sealed.

Evaluation metrics

For each strict positive, aggregate rank and reciprocal rank were recorded. Mean reciprocal rank was the mean of reciprocal positive ranks. Recall at K was the proportion of the four strict positives ranked at or above K. Observed positive yield at K was the number of observed positives in the first K positions divided by K, and enrichment factor compared this yield with the background positive proportion (4/3 489). "Observed positive yield" was used instead of conventional precision because unlabeled units were not verified negatives. Mean and median positive rank, Hits at 10, 25, 50, 100, and 250, and full-ranking Spearman correlation were also reported. Accuracy, specificity, ROC-AUC, PR-AUC, calibrated binary risk, and conventional significance tests were not used.

Non-embedding graph baselines

Four rules were frozen before baseline rankings were generated: representative-CUI undirected degree, undirected shortest-path distance to DMD, undirected common-neighbor count with DMD, and a UMLS-only relation-proximity rule using direct relation-label count followed by relation-labeled two-hop paths. Rankings used deterministic aggregation-unit tie-breaks. All seven view-rule combinations were produced before endpoint labels were materialized. Tie-pattern diagnostics were retained because these rules generated coarse rankings [see Additional file 1: Table S7].

Manual audit of patent predications

The audit population comprised 541 model-included resolved patent relation instances. A simple random sample of 200 instances was drawn before annotation. Eight supplemental rows ensured minimum relation-class coverage, and all 11 entity-resolution-applied instances were reviewed; after overlap, 215 unique rows were annotated. A domain expert in health sciences applied a frozen rubric covering subject alignment, object alignment, predicate support, direction, negation/context handling, and overall validity (valid, invalid, unresolved). Because the audit was conducted by a single domain expert, no inter-rater agreement statistic was computed. Labels were revealed only after the sample and rubric were sealed and were not used to modify the graph, retrain models, or change rankings. The primary estimand was the decidable validity proportion among valid plus invalid rows; Wilson intervals and finite-population sensitivity bounds were reported.

Table 2 Confirmatory validation performance

Model	Mean validation MRR	SD	Final epoch
TransE	0.088364	0.004482	85
DistMult	0.077969	0.067701	50
RotatE	0.060720	0.007256	95
ComplEx	0.000355	0.000376	5

Architecture robustness, discovery, normalization, and patent attribution

Patent-associated rank changes were examined across all four KGE architectures without re-selecting the primary model. The frozen discovery ranking was TransE on INTEGRATED_SAFE after excluding strict positives and prior-history controls. Top-25, Top-100, and mechanistic Top-10 sets were frozen. Historical normalization used only graph-native UMLS 2018AA evidence; contemporary lexical or RxNorm normalization was prohibited. Patent attribution was conducted after discovery ranks were frozen and considered direct patent adjacency, one-UMLS-hop bridges, mixed UMLS/patent paths to DMD anchors, traversal direction, reverse-context status, source layer, internal source-edge identifiers, evidence/publication counts, and publication dates. The historical extraction did not preserve a SemRep confidence field, so unsupported confidence scores were not reconstructed. Representative two- and three-edge paths were selected for provenance and interpretability rather than extended arbitrarily, because additional hops can increase semantic drift. The domain-expert audit provided quality context for the underlying patent relations; representative path classifications were made by the authors after ranking freeze and were diagnostic rather than independent clinical adjudication.

Biomedical and clinical validation

A frozen search policy evaluated five unique search cores represented in the mechanistic Top-10: nusinersen, selegiline, selegiline hydrochloride, edaravone, and cerliponase alfa. PubMed and ClinicalTrials.gov were primary validation sources; regulatory records were used when relevant. Evidence was separated into pre-cutoff, post-cutoff, and current-status categories. Candidate-specific DMD evidence was distinguished from evidence in another disease or at pharmacological-class level. Each mechanistic Top-10 ranking unit was additionally assigned to one of four DMD interpretation categories: disease-modifying, complication-related, symptomatic/supportive, or nonspecific graph-derived. Assignment to the first three categories required candidate-specific DMD evidence consistent with that role; units lacking such evidence were classified as nonspecific graph-derived hypotheses. Search cores were used only to improve evidence-retrieval recall and did not alter graph identity or ranking units. Negative targeted searches were not interpreted as proof of inefficacy or absolute absence [see Additional file 1: Tables S2-S4].

Cutoff sensitivity

UMLS 2018AA became available on 7 May 2018 and the next release, 2018AB, became available on 5 November 2018. Four cutoffs were therefore frozen within a single UMLS-release interval: 7 May, 6 July, 4 September, and 3 November 2018. For each cutoff, the Integrated snapshot contained all 47 816 leakage-safe UMLS edges plus patent edges whose earliest supporting publication date was on or before the cutoff. Cutoff-specific vocabularies were rebuilt. The frozen TransE configuration and five final seeds were used without new HPO or architecture selection. Label-free stability was sealed before endpoint labels were joined. The time-aligned UMLS anchor was trained once on the UMLS catalog vocabulary and was distinct from the primary global-vocabulary UMLS-only refit used in source ablation [see Additional file 1: Table S8].

Reproducibility and phase sealing

Candidate construction, graph policies, endpoint definitions, HPO, final refits, label-blind scoring, ensembles, baselines, audit samples, cutoff snapshots, stability analyses, and endpoint evaluations were independently validated and SHA-256 sealed. Failed or superseded runs remained preserved and were not silently overwritten. No centrality-pruning step was used in the reported analysis. The public release includes machine-readable results, input and graph-view manifests, integrity hashes, and an automated verification utility [24, 25].

Use of generative AI

OpenAI ChatGPT was used to assist with language editing, document organization, and formatting during manuscript preparation. It was not used to generate or alter the study data, execute the reported analyses, or determine the reported numerical results. Both authors reviewed and approved the final manuscript and accept full responsibility for its scientific content.

Results

Graph and candidate universe

After target-leakage filtering, UMLS_ONLY_SAFE contained 47 816 edges, 41 321 active nodes, and 69 relations. INTEGRATED_SAFE contained 48 161 edges, 41 333 active nodes, and 91 relations in a global 41 335-node catalog. PATENT_ONLY_SAFE contained 345 edges, 299 active nodes, and 22 relations. The candidate universe contained 3 684 CUIs and 3 489 ranking units. Patent-only active representative coverage was 0.258% (9/3 489), and the target relation was absent; same-task patent-only KGE was therefore structurally unavailable (Table 1).

Model selection

TransE achieved the highest mean confirmatory MRR (0.0884, SD 0.0045) and was frozen as the primary architecture (Table 2). Architecture selection was based on independent three-seed confirmation rather than the strongest individual HPO trial.

Primary source ablation

Strict-positive ranks were heterogeneous (Table 3; Fig. 2). Tacrolimus improved from rank 1 163 in UMLS-only to 331 in Integrated, whereas Captopril, Bisoprolol Fumarate, and metoprolol succinate ranked lower after patent augmentation. Integrated therefore improved one of four strict-positive ranks.

Aggregate primary metrics are summarized in Table 4. Integrated MRR was 0.002951 versus 0.005564 for UMLS-only. Mean positive rank favored Integrated because of the large Tacrolimus improvement, whereas median rank and Top-K recovery favored UMLS-only.

Non-embedding baselines

Baseline performance is summarized in Table 5; Fig. 3. The UMLS-only shortest-path rule achieved MRR 0.0260, placed Captopril at rank 11, and recovered three positives by Top 250. It nevertheless had mean positive rank 597.5 because Tacrolimus ranked 2 040th, and it produced only eight distinct score patterns with a largest tie group of 1 483 units. Degree, common-neighbor, and relation-proximity rules were similarly coarse.

Table 3 Primary TransE strict-positive ranks

Candidate	UMLS-only rank	Integrated rank	Delta
Captopril	315	418	-103
Tacrolimus	1 163	331	+832
Bisoprolol Fumarate	136	317	-181
metoprolol succinate	92	309	-217

Delta is UMLS-only rank minus Integrated rank; positive values favor Integrated

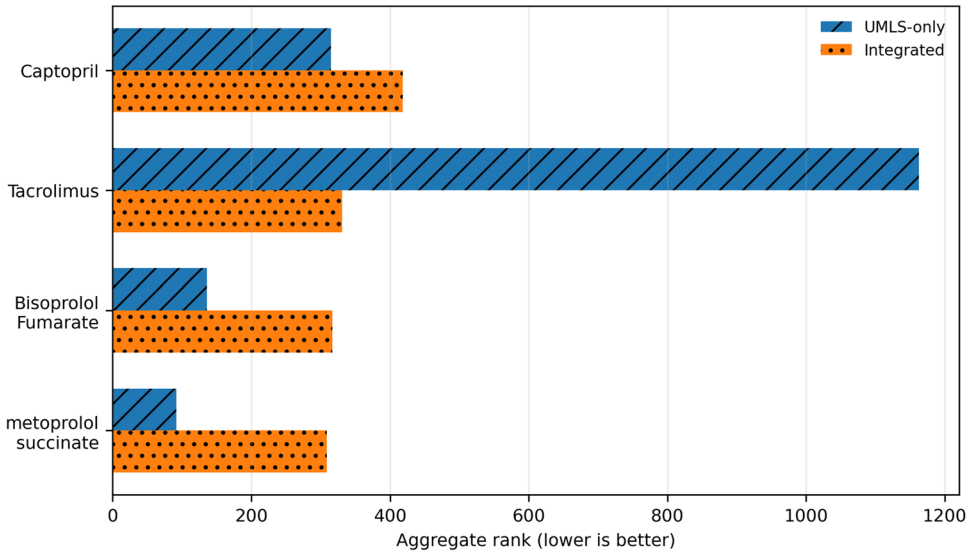


Fig. 2 Strict-positive ranks. Primary TransE aggregate ranks for the four strict post-cutoff positives. Lower rank is better

Table 4 Primary retrospective metrics

View	MRR	Mean rank	Median rank	Hits@100	Recall@100	Yield@100	Hits@250	EF@250
Integrated	0.002951	343.75	324.0	0	0.00	0.000	0	0.000
UMLS-only	0.005564	426.50	225.5	1	0.25	0.010	2	6.978

Observed positive yield is not conventional binary precision because the outcome is positive-unlabeled

Table 5 Non-embedding baseline metrics

Rule	View	MRR	Mean rank	Median rank	Hits@100	Recall@100	Hits@250	EF@250
Degree	UMLS-only	0.002765	466.75	363.0	0	0.00	1	3.489
Degree	Integrated	0.002765	466.75	363.0	0	0.00	1	3.489
Shortest path	UMLS-only	0.026029	597.50	169.5	1	0.25	3	10.467
Shortest path	Integrated	0.003250	430.25	424.5	0	0.00	1	3.489
Common neighbors	UMLS-only	0.004118	590.00	579.5	1	0.25	1	3.489
Common neighbors	Integrated	0.002451	665.75	653.5	0	0.00	1	3.489
Relation proximity	UMLS-only	0.004118	590.00	579.5	1	0.25	1	3.489

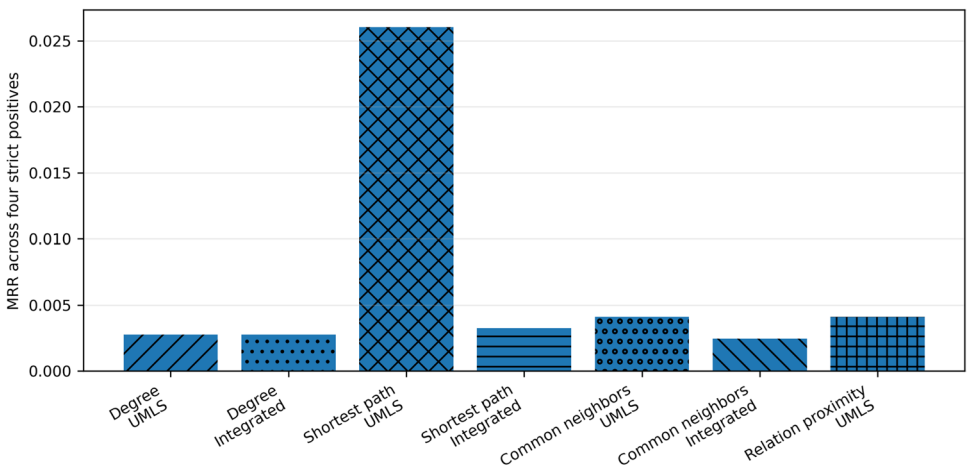


Fig. 3 Non-embedding baseline performance. MRR for seven pre-specified baseline rankings. Endpoint labels were joined only after all rankings had been generated

Table 6 Manual patent-predication audit

Component or subset	*n*	Supported/valid	Unsupported/invalid	Uncertain/unresolved
Subject concept alignment	200	182 (91.0%)	15 (7.5%)	3 (1.5%)
Object concept alignment	200	182 (91.0%)	15 (7.5%)	3 (1.5%)
Predicate semantic support	200	113 (56.5%)	84 (42.0%)	3 (1.5%)
Direction support	200	124 (62.0%)	5 (2.5%)	71 (35.5%)
Negation/context handling	200	160 (80.0%)	40 (20.0%)	0
Overall relation validity	200	106 (53.0% all rows)	89 (44.5%)	5 (2.5%)
DMD-incident subset	42	26 (61.9%)	16 (38.1%)	0
Direct therapeutic subset	41	16 (39.0%)	25 (61.0%)	0
Entity-resolution census	11	1 (9.1%)	10 (90.9%)	0

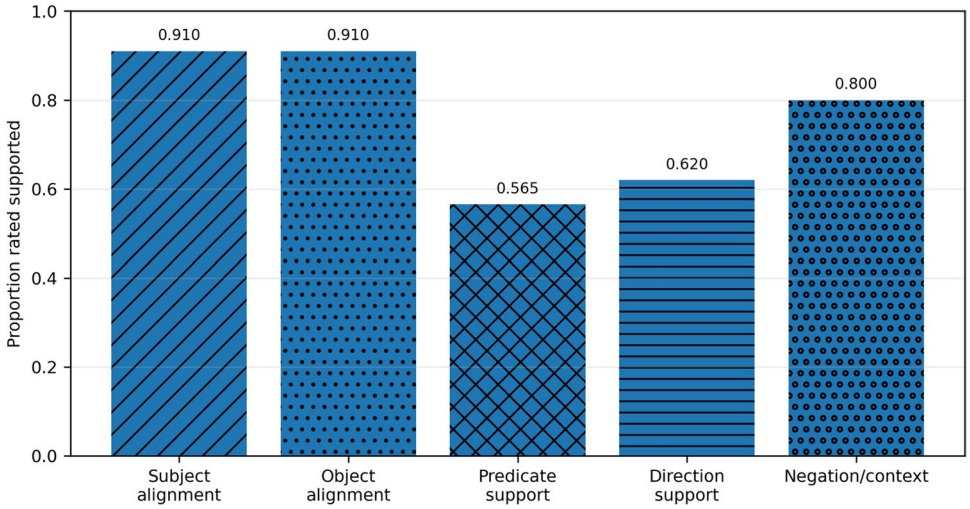


Fig. 4 Patent-predication audit. Component-level support rates in the 200-instance random sample. The audit was conducted by a single domain expert in health sciences; no inter-rater agreement statistic was computed

Patent-predication audit

Among 200 randomly sampled relation instances, 106 were valid, 89 invalid, and 5 unresolved. Decidable validity was 54.36% (Wilson 95% CI 47.35–61.20%); treating unresolved rows as all invalid or all valid gave bounds of 53.0-55.5%. Concept alignment was supported in 91.0% of rows, but predicate semantic support was 56.5%. Component and subgroup results are summarized in Table 6; Fig. 4.

Of the 89 invalid primary rows, 63 (70.8%) had multiple interacting error types. Relation-class validity was highest for taxonomic (87.3%) and procedural (83.3%) relations, but low for therapeutic (39.0%), associative (20.0%), and mechanistic (20.0%) relations. Audit outcomes were not used to modify the graph, so these findings quantify uncertainty in the frozen patent layer rather than performance after cleaning.

Architecture robustness

Patent-associated rank changes varied across architectures (Fig. 5). Mean UMLS-minus-Integrated rank difference was +82.75 for TransE, -668.75 for DistMult, +242.75 for RotatE, and +4.25 for ComplEx. No strict positive improved consistently across all architectures. Full-ranking UMLS/Integrated Spearman correlations were approximately 0.901, 0.661, 0.524, and 1.000, respectively.

Cutoff sensitivity

Integrated graph size increased from 48 127 to 48 161 edges as the cutoff advanced; eligible patent edges increased from 311 to 345. Integrated MRR rose from 0.002166 to 0.002951 and mean positive rank improved from 542.5

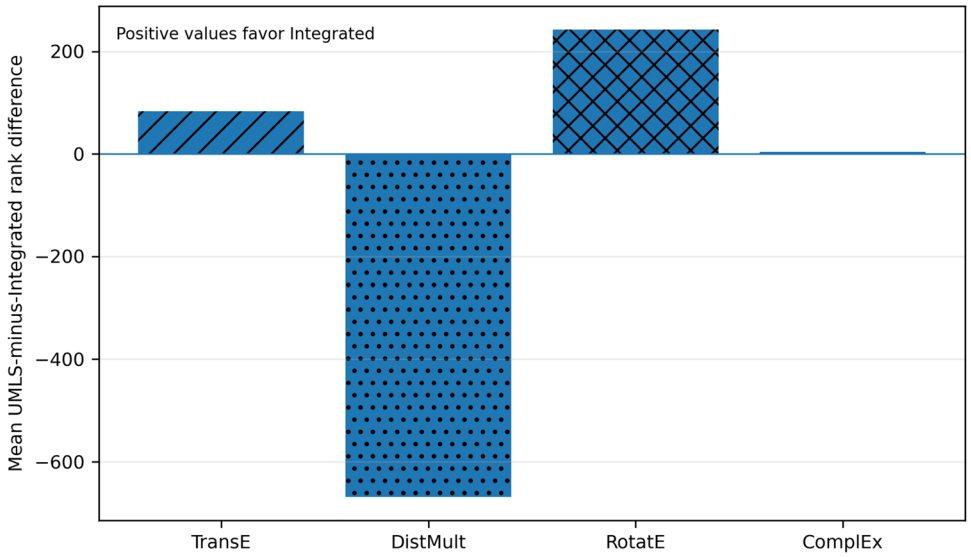


Fig. 5 Architecture-dependent rank shifts. Mean UMLS-minus-Integrated rank difference across the four strict positives. Positive values favor Integrated

Table 7 Cutoff-sensitivity metrics

Cutoff	Patent edges	Training vocabulary	Integrated MRR	UMLS MRR	Integrated mean rank	UMLS mean rank	Integrated Hits@250	UMLS Hits@250
7 May 2018	311	41 332	0.002166	0.003245	542.50	329.50	0	1
6 Jul 2018	322	41 333	0.002384	0.003245	518.50	329.50	0	1
4 Sep 2018	328	41 334	0.002869	0.003245	361.00	329.50	0	1
3 Nov 2018	345	41 335	0.002951	0.003245	343.75	329.50	0	1

The UMLS anchor was trained on the time-aligned UMLS catalog vocabulary and was distinct from the primary global-vocabulary UMLS-only refit

to 343.75. Nevertheless, the time-aligned UMLS anchor retained higher MRR and better mean positive rank at every cutoff, and no Integrated condition placed a strict positive in the Top 250. Table 7; Fig. 6 summarize aggregate cutoff results.

Individual trajectories were non-monotonic (Fig. 7). Captopril and metoprolol succinate remained UMLS-favored at every cutoff; Bisoprolol Fumarate remained Integrated-favored; and Tacrolimus became Integrated-favored only at the primary cutoff.

Whole-ranking stability across Integrated cutoffs was high, with Spearman correlations 0.9296–0.9348. Top-ranked overlap was lower: Top-25 Jaccard 0.2821–0.3889, Top-100 Jaccard 0.3605–0.4599, and Top-250 Jaccard 0.6447–0.7123 (Fig. 8). Thus, the global ranking was stable while the practical prioritization region was more cutoff-sensitive.

Discovery ranking, normalization, and attribution

After excluding strict positives and prior-history controls, 3 481 ranking units formed the unlabeled discovery set. A nusinersen injection formulation moved from UMLS rank 1 021 to Integrated rank 1, an edaravone injection moved from 2 068 to 3, and edaravone moved from 2 085 to 9. Multiple selegiline formulations occupied ranks 2 and 4–8. Because the Top-25 was selected by Integrated rank, uniformly positive deltas within that selected set were selection-induced rather than independent evidence of patent benefit.

The Top-25 contained 16 formulation CUIs without graph-native ingredient mapping and 9 concept CUIs. In the Top-100, 74 were unresolved formulations and 26 were concept CUIs. No evidence-supported UMLS 2018AA normalization was available for any Top-100 unit, so formulation concepts were retained. The list must therefore be interpreted as ranking units rather than distinct active ingredients.

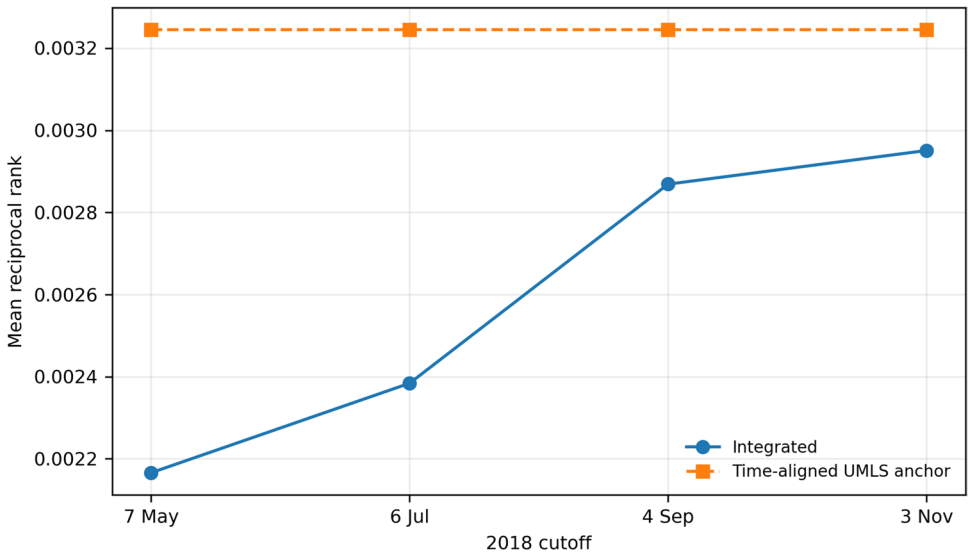


Fig. 6 Cutoff-specific MRR. Integrated MRR improved across the four prespecified cutoffs but remained below the fixed time-aligned UMLS anchor

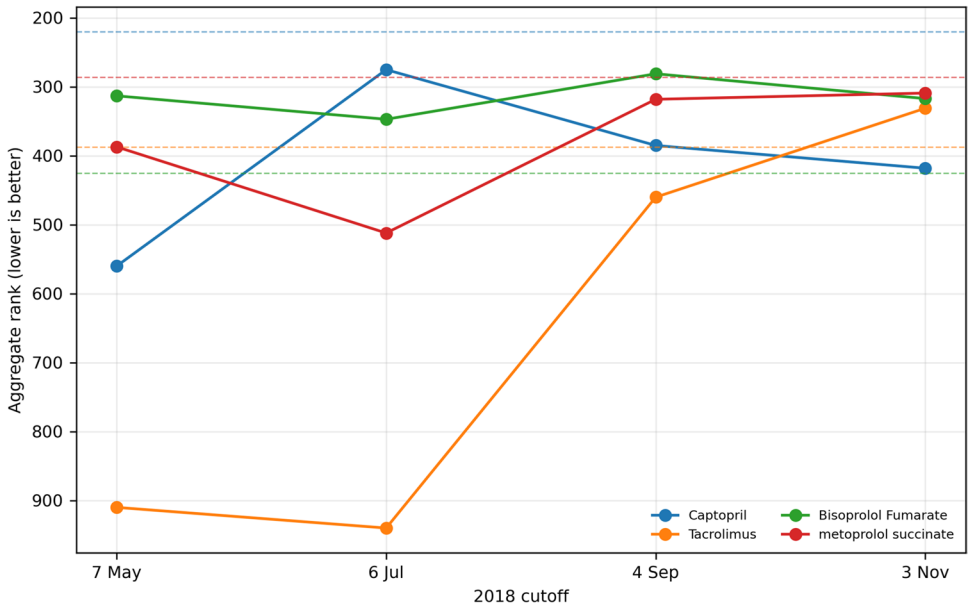


Fig. 7 Strict-positive cutoff trajectories. Solid lines show Integrated cutoff-specific ranks; dashed references show each candidate's fixed time-aligned UMLS rank. Lower rank is better

Patent attribution identified 31 one-UMLS-hop bridges and 51 mixed UMLS/patent paths linking candidate neighborhoods to DMD-related anchors, but no candidate-specific patent mechanism. Eight mechanistic Top-10 units were Tier C (directed associative DMD context); nusinersen and cerliponase alfa were Tier D (reverse-context-only or structural context). Because strict positives and prior-history controls were excluded before the discovery ranking was constructed, all mechanistic Top-10 units were unlabeled discovery units. Under the reviewer-requested DMD interpretation categories, none qualified as disease-modifying, complication-related, or symptomatic/supportive; all 10 were classified as nonspecific graph-derived hypotheses. Nusinersen and edaravone had evidence in related neuromuscular diseases, selegiline-family units had class-level MAO/MAO-B rationale without candidate-specific DMD support, and cerliponase alfa had evidence in CLN2 disease rather than DMD. Table 8 provides the consolidated interpretation, with detailed scoring, endpoint status, evidence provenance, and paths in Additional file 1: Tables S2-S4.

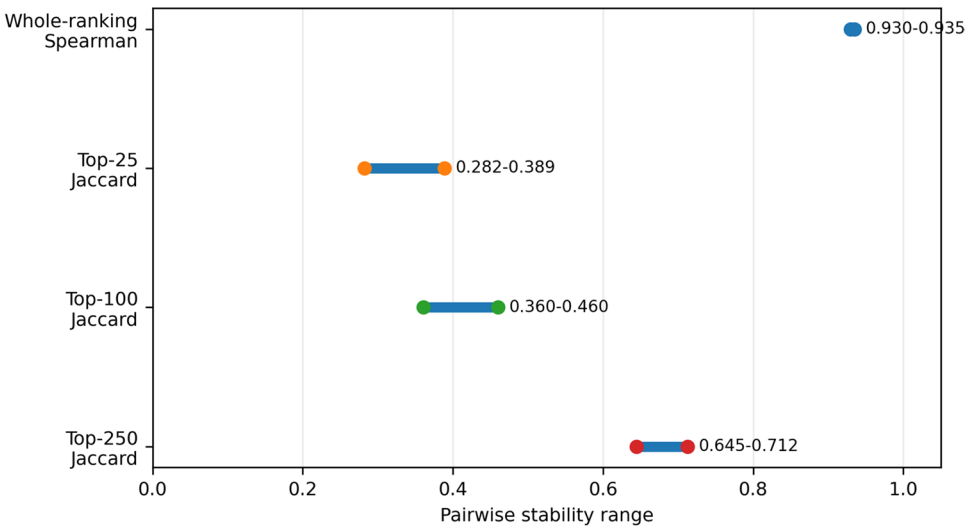


Fig. 8 Cutoff stability ranges. Minimum and maximum pairwise stability across the four Integrated cutoff conditions. The Top-25 region was substantially less stable than the complete ranking

Table 8 Mechanistic Top-10 interpretation

Rank	Ranking unit	Delta	Patent context	DMD interpretation	Secondary evidence
1	5 ML nusinersen 2.4 MG/ML Injection	+1 020	Tier D; no candidate-specific mechanism	Nonspecific graph-derived	SMA evidence only
2	24 h Selegiline 0.5 MG/HR Transdermal System	+23	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	Class-level rationale only
3	100 ML edaravone 0.3 MG/ML Injection	+2 065	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	ALS/SMA evidence only
4	selegiline hydrochloride 5 MG Oral Capsule	+35	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	Class-level rationale only
5	24 h Selegiline 0.25 MG/HR Transdermal System	+37	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	Class-level rationale only
6	selegiline hydrochloride 5 MG Oral Tablet	+63	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	Class-level rationale only
7	selegiline hydrochloride	+19	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	Class-level rationale only
8	selegiline hydrochloride 1.25 MG Disintegrating Oral Tablet	+22	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	Class-level rationale only
9	edaravone	+2 076	Tier C; no candidate-specific mechanism	Nonspecific graph-derived	ALS/SMA evidence only
10	5 ML cerliponase alfa 30 MG/ML Injection	+4	Tier D; no candidate-specific mechanism	Nonspecific graph-derived	CLN2 evidence only

Tier C denotes directed associative DMD context; Tier D denotes reverse-context-only or structural context. All rows were unlabeled discovery units and were classified as nonspecific graph-derived hypotheses; secondary evidence descriptors do not imply DMD efficacy

Tier C denotes directed associative DMD context; Tier D denotes reverse-context-only or structural context. All rows were unlabeled discovery units and were classified as nonspecific graph-derived hypotheses; secondary evidence descriptors do not imply DMD efficacy.

Error analysis

Four recurrent error modes emerged: patent augmentation was not directionally stable across candidates or architectures; large uplifts could arise from disease-context topology rather than drug-specific evidence; historical identity resolution produced formulation duplication; and many patent predications lacked adequate predicate support, with especially poor performance in entity-resolution-applied instances. These findings explain why rank movement was interpreted as hypothesis generation rather than validation.

Discussion

Patent augmentation changed rankings without uniform benefit

The principal result was not uniform improvement. A small patent layer produced large candidate-specific shifts in an otherwise related ranking space. Integrated improved Tacrolimus but worsened the other three strict positives, and UMLS-only achieved higher MRR, better median rank, and greater Top-K recovery. Integrated's better mean rank was driven by one large movement. The cutoff analysis reinforced this interpretation: later patent eligibility improved Integrated aggregate metrics, but never allowed Integrated to surpass the time-aligned UMLS anchor in MRR or mean positive rank.

These findings distinguish ranking perturbation from predictive benefit. A small number of edges can reposition entities through shared neighborhoods and relation geometry even when no added edge directly encodes a drug-DMD mechanism. Patent augmentation was associated with substantial reshaping of the ranked hypothesis space, without demonstrating uniformly better retrospective recovery.

Relation to previous work and data-construction choices

Large heterogeneous biomedical graphs such as PrimeKG and systems such as TxGNN demonstrate the value of broad multi-source integration, while SemRep-based time-sliced KGC and path-based KGC illustrate temporal and explanation-oriented graph workflows [7, 16, 17, 19]. Their datasets, prediction targets, and validation designs differ from the present matched source-ablation question. Here, the distinctive contribution is not a larger general-purpose graph but a controlled comparison of a curated UMLS backbone with and without a small automatically extracted patent layer, accompanied by leakage removal, source-quality auditing, simple topology, architecture analysis, and cutoff sensitivity.

Patent-mining studies have argued that patents complement public biomedical resources [8–11], but the current audit shows why complementarity cannot be equated with reliable therapeutic knowledge. Non-graph approaches based on molecular signatures, docking, chemical similarity, genetics, pathways, or clinical observations provide orthogonal evidence [2, 12] and were not directly benchmarked because the frozen dataset does not contain the required molecular features. The findings therefore support a methodological evaluation of source augmentation, not superiority over other repurposing paradigms.

Simple topology outperformed TransE on selected retrospective metrics but was coarse

The UMLS-only shortest-path baseline outperformed the learned TransE rankings in MRR and Top-250 recall. This indicates that proximity to DMD in the curated graph captured useful retrospective signal. However, the baseline ranked one positive extremely poorly, produced very large tie groups, and ignored relation direction and type. Its high MRR was concentrated in a few favorable placements rather than broad resolution. Simple baselines should therefore accompany KGE evaluation, but unweighted distance should not be treated as a relation-aware substitute. Accordingly, the present results do not demonstrate superiority of KGE over simple graph topology for retrospective DMD candidate recovery.

Extraction quality narrowed the evidential claim

The contrast between source roles was substantive. UMLS supplied a curated terminology-based baseline, whereas the patent layer was generated automatically by SemRep/MetaMap and entered the frozen graph without manual pre-screening. In the domain-expert audit, decidable overall validity was 54.36%, direct-therapeutic validity was 39.0%, and only 1 of 11 entity-resolution-applied instances was valid (9.1%). Concept alignment was comparatively strong at 91.0%, but predicate semantic support was only 56.5%. PubMed, ClinicalTrials.gov,

and regulatory records were used only for post-ranking validation and did not repair the training graph. Because the audit was diagnostic and did not retroactively clean the graph, all ranking results include the observed extraction noise. The patent layer should therefore be viewed as a historically preserved contextual augmentation rather than a validated therapeutic relation database.

Architecture and cutoff dependence were substantive

No strict positive benefited consistently across TransE, DistMult, RotatE, and ComplEx. The same patent edges interacted differently with translational, bilinear, and complex-space geometries, so a rank increase after augmentation was not an intrinsic biological property of the source layer.

The cutoff analysis also showed that global stability can conceal local churn. Whole-ranking correlations around 0.93 appeared high, but Top-25 overlap remained below 0.40. Since practical prioritization focuses on the highest-ranked candidates, global correlation alone understated cutoff sensitivity. The analysis remained descriptive: patent-edge counts and performance co-varied over time, but the design did not identify causal effects of individual patent edges.

Rank shifts were not mechanistic explanations

Nusinersen and edaravone had some of the largest Integrated uplifts, but attribution showed associative disease-context or reverse-context paths rather than candidate-specific mechanisms. External validation identified evidence in related neuromuscular diseases rather than DMD efficacy. Under the explicit DMD interpretation framework, none of the mechanistic Top-10 units qualified as disease-modifying, complication-related, or symptomatic/supportive; all were nonspecific graph-derived hypotheses. A path can explain why graph topology changed without providing biological evidence for a predicted treatment relation.

Historical normalization was part of the design

Contemporary resources could collapse many formulation CUIs, but applying them retrospectively would import knowledge unavailable under the 2018AA design. Candidate identity was therefore treated as part of the historical data-generating process. The resulting formulation redundancy is a limitation, but post-hoc normalization would define a different study.

Methodological implications

Retrospective biomedical KGC should freeze more than the graph. Candidate definitions, target leakage, vocabularies, model selection, scoring, tie-breaks, endpoint mapping, external validation, and reporting rules should be separated and versioned. Non-embedding baselines and extraction-quality audits should be specified before endpoint access. Cryptographic sealing cannot establish biological validity, but it constrains researcher degrees of freedom and makes failed or superseded analyses visible.

Limitations

This study has several limitations. Only four ranking units met the primary endpoint, making aggregate metrics highly sensitive to individual rank movements, and ClinicalTrials.gov entry is an imperfect proxy shaped by biological, strategic, commercial, and exploratory factors rather than therapeutic efficacy. The endpoint was positive-unlabeled, so unlabeled candidates were not confirmed negatives and conventional discrimination metrics were not interpretable in their usual sense. The patent layer was small and noisy: the leakage-safe view contained 345 edges, did not represent the complete patent literature, and achieved 54.36% decidable overall validity, with

only 39.0% validity in the direct-therapeutic subset. The audit was conducted by a single domain expert in health sciences and therefore lacked an inter-rater reliability estimate. Sparse patent-only coverage and absence of the frozen target relation prevented a commensurate full-universe patent-only KGE result. Associative, structural, and ontological edges could influence embeddings without encoding therapeutic causality; rank shifts depended on architecture; and seed-based analyses quantified initialization variability under a fixed train/validation construction but did not assess uncertainty from alternative graph splits. Historical UMLS 2018AA identity also prevented post-hoc active-ingredient collapse for many formulation CUIs, and no mechanistic Top-10 unit had a candidate-specific patent mechanism or DMD validation. Representative path classifications were diagnostic author interpretations informed by frozen provenance and audit findings rather than independent clinical adjudication. Cutoff sensitivity was limited to one UMLS 2018AA availability interval, while baseline rankings were tie-dominated and sensitive to deterministic tie-breaking. The outputs were computational hypotheses without experimental validation, and generalizability is limited by the single disease, one UMLS release family, one patent layer, and four embedding architectures. Finally, the complete raw Google Patents snapshot is not redistributed; graph-level reconstruction begins from documented frozen resolved predications and requires licensed UMLS 2018AA inputs.

Conclusions

Integrating a leakage-safe patent relation layer with historical UMLS knowledge materially changed selected DMD drug-repurposing ranks, but the changes were neither uniformly beneficial across retrospective positives nor stable in direction across architectures. UMLS-only TransE outperformed Integrated in MRR and Top-K recovery, and the time-aligned UMLS anchor remained stronger than every Integrated cutoff in MRR and mean positive rank. A simple UMLS shortest-path baseline recovered several positives strongly but was coarse and tie-dominated. Manual auditing revealed substantial predicate and entity-resolution error. Attribution showed that large discovery shifts reflected associative disease context rather than candidate-specific patent mechanisms, and all mechanistic Top-10 units were classified as nonspecific graph-derived hypotheses because none had candidate-specific DMD evidence. In this study, adding the patent-derived layer was associated with substantial perturbation of the ranked hypothesis space, but rank movement should not be equated with mechanism, efficacy, or causal source contribution. Robust retrospective graph-based repurposing requires temporal vocabulary alignment, label-blind scoring, simple baselines, extraction-quality auditing, source-specific provenance, and explicit positive-unlabeled interpretation.

Abbreviations

CUI	Concept unique identifier
DMD	Duchenne muscular dystrophy
EF	Enrichment factor
HPO	Hyperparameter optimization
KGC	Knowledge graph completion
LBD	Literature-based discovery
MRR	Mean reciprocal rank
NLP	Natural-language processing
PU	Positive-unlabeled
UMLS	Unified Medical Language System
WSD	Word-sense disambiguation

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1186/s13040-026-00605-6.

Acknowledgements The authors thank Omer Turkmen, a domain expert in health sciences, for conducting themanual audit of patent predications and applying the frozen rubric for concept alignment, predicate support, direction, context handling, and overall validity.

Author contributions FU conceptualized the study, developed the methodology, performed data collection and preprocessing, constructed the knowledge graph, conducted the experiments, analyzed the results, and wrote the original manuscript. HEC supervised the study, contributed to the interpretation of the findings, and provided critical revisions to the manuscript. Both authors reviewed and approved the final version of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate Not applicable. The study did not involve human participants, human tissue, animals, or identifiable individual-level data.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3:673–83. https://doi.org/10.1038/nrd1468.
2. Pushpakom S, Iorio F, Eyers PA, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019;18:41–58. https://doi.org/10.1038/nrd.2018.168.
3. Nguengang Wakap S, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28:165–73. https://doi.org/10.1038/s41431-019-0508-0.
4. Duan D, Goemans N, Takeda S, Mercuri E, Aartsma-Rus A. Duchenne muscular dystrophy. Nat Rev Dis Primers. 2021;7:13. https://doi.org/10.1038/s41572-021-00248-3.
5. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inf. 2003;36:462–77. https://doi.org/10.1016/j.jbi.2003.11.003.
6. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32:D267–70. https://doi.org/10.1093/nar/gkh061.
7. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023;10:67. https://doi.org/10.1038/s41597-023-01960-3.
8. Rodriguez-Esteban R, Bundschus M. Text mining patents for biomedical knowledge. Drug Discov Today. 2016;21:997–1002. https://doi.org/10.1016/j.drudis.2016.05.002.
9. Shalaby W, Zadrozny W. Patent retrieval: a literature review. Knowl Inf Syst. 2019;61:631–60. https://doi.org/10.1007/s10115-018-1322-7.
10. Southan C, Varkonyi P, Muresan S. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. J Cheminform. 2009;1:10. https://doi.org/10.1186/1758-2946-1-10.
11. Robson B, Li J, Dettinger R, Peters A, Boyer SK. Drug discovery using very large numbers of patents. General strategy with extensive use of match and edit operations. J Comput Aided Mol Des. 2011;25:427–41. https://doi.org/10.1007/s10822-011-9429-x.

12. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science. 2008;321:263–6. https://doi.org/10.1126/science.1158140.
13. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspect Biol Med. 1986;30:7–18. https://doi.org/10.1353/pbm.1986.0087.
14. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artif Intell. 1997;91:183–203. https://doi.org/10.1016/S0004-3702(97)00008-8.
15. Nickel M, Murphy K, Tresp V, Gabrilovich E. A review of relational machine learning for knowledge graphs. Proc IEEE. 2016;104:11–33. https://doi.org/10.1109/JPROC.2015.2483592.
16. Huang K, Chandak P, Wang Q, et al. A foundation model for clinician-centered drug repurposing. Nat Med. 2024;30:3601–13. https://doi.org/10.1038/s41591-024-03233-x.
17. Jimenez A, Merino MJ, Parras J, Zazo S. Explainable drug repurposing via path-based knowledge graph completion. Sci Rep. 2024;14:16587. https://doi.org/10.1038/s41598-024-67163-x.
18. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. J Biomed Inf. 2009;42:633–43. https://doi.org/10.1016/j.jbi.2008.12.001.
19. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. J Biomed Inf. 2021;115:103696. https://doi.org/10.1016/j.jbi.2021.103696.
20. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. Adv Neural Inf Process Syst. 2013;26.
21. Yang B, Yih WT, He X, Gao J, Deng L. Embedding entities and relations for learning and inference in knowledge bases. International Conference on Learning Representations. 2015.
22. Trouillon T, Welbl J, Riedel S, Gaussier E, Bouchard G. Complex embeddings for simple link prediction. Proc Mach Learn Res. 2016;48:2071–80.
23. Sun Z, Deng ZH, Nie JY, Tang J. RotatE: knowledge graph embedding by relational rotation in complex space. International Conference on Learning Representations. 2019.
24. Uludağ F. Patent-Augmented knowledge graph drug repurposing for duchenne muscular dystrophy. Version v1.1.0. Zenodo. 2026. https://doi.org/10.5281/zenodo.21876856.
25. Uludağ F. Patent-Augmented knowledge graph drug repurposing for duchenne muscular dystrophy [software]. GitHub; 2026. https://github.com/fatihuludag-lab/dmd-patent-augmented-kg-repurposing. Accessed 12 Aug 2026.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Fatih Uludağ[11] · H. Eray CELİK[11]

✉ Fatih Uludağ[1]
fatihuludag@yyu.edu.tr

[1] Department of Econometrics, Faculty of Economics and Administrative Sciences, Van Yüzüncü Yıl University, Van, Türkiye, Turkey