PRIMEKG-CL: A CONTINUAL GRAPH LEARNING BENCHMARK ON EVOLVING BIOMEDICAL KNOWLEDGE GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical knowledge graphs underwrite drug repurposing and clinical decision support, yet the upstream ontologies they depend on (GO, HPO, MONDO, CTD) update on independent cycles that add millions of edges and deprecate hundreds of thousands more between releases. Yet existing continual graph learning (CGL) has been studied almost exclusively on synthetic random splits of static, generic KGs, a regime that cannot reproduce the asynchronous, structured evolution real biomedical KGs undergo. To this end, we introduce **PrimeKG-CL**, a CGL benchmark built on PrimeKG (integrating 20+ biomedical databases; 129K+ nodes, 8.1M+ edges, 10 node types, 30 relation types) with a second temporal snapshot reconstructed from nine freely re-queried databases (t_0 June 2021, t_1 July 2023; 5.76M edges added, 889K removed, 7.21M persistent), six entity-type-grouped tasks tracking real $t_0 \to t_1$ additions, multimodal node features (BiomedBERT text, Morgan fingerprints, R-GCN structural), and a per-task persistent/removed/added test stratification. On three tasks (biomedical relationship prediction, entity classification, KGQA), we evaluate six CL strategies across four KGE decoders, plus LKGE, an LLM-RAG agent, and CMKL. We find that *decoder choice dominates the continual-learning strategy*: link-prediction accuracy spans roughly $20\times$ across decoders (RotatE $\gg$ DistMult $\gg$ ComplEx $\gg$ TransE) versus up to $\approx 4\times$ across CL strategies, and the two interact so that no single strategy is best across decoders (e.g., EWC markedly benefits ComplEx while Distillation degrades RotatE). Moreover, only DistMult exhibits a clear separation between persistent and deprecated knowledge ($\approx 11\times$ higher MRR on persistent vs. removed triples), indicating that standard metrics conflate retention of still-valid facts with failure to forget outdated ones; this effect is absent under RotatE. In addition, multimodal fusion improves entity classification by $+43\%$ over feature-matched continual baselines (Macro-F1 0.591 vs. 0.414, $p < 10^{-6}$), and a recent CKGE framework (IncDE) failed to scale to our 5.67M-triple base task across five attempts up to 350 GB RAM. Data, pipeline, baselines, and the stratified split will be released openly.

1 INTRODUCTION

Biomedical knowledge graphs (KGs) support drug discovery, disease understanding, and clinical decision support (Chandak et al., 2023; Huang et al., 2024). As GO, HPO, MONDO, and CTD update asynchronously, *how can models incorporate new knowledge without catastrophically forgetting?*

Existing CGL benchmarks use synthetic splits of static KGs, including FB15k-237/WN18RR (Cui et al., 2023; Liu et al., 2024a). These miss domain-specific evolution: anatomy–protein expression edges (Bgee) dominate additions, while disease–phenotype links churn as HPO is refined.

We introduce PrimeKG-CL (**Prime**KG for **C**ontinual **L**earning), the first biomedical-scale CGL benchmark with genuine temporal evolution and multimodal features. Sections 3.1 and 3.5 describe its extensible snapshot construction and features; Section 4.1 details the KGE baselines and CMKL's three-modality MoE fusion.

Across six CL methods $\times$ four KGE decoders on three tasks, we find that *decoder choice dominates the CL strategy* — link-prediction accuracy spans $\approx 20\times$ across decoders versus up to $\approx 4\times$

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

Table 1: Comparison of PrimeKG-CL with existing continual graph learning benchmarks. PrimeKG-CL is the first to offer real temporal evolution on a biomedical KG with multimodal features and multiple evaluation tasks. "Mod." = modality types provided (S: structural, T: textual, V: visual, M: molecular).

Benchmark	Domain	Real Temp.	Mod.	#Entities	#Edges	#Rel.	#Tasks	#Methods	Eval Tasks
LKGE (Cui et al., 2023)	General	✗	S	15K	310K	237	5	6	LP
Daruna et al. (Daruna et al., 2021)	Robotics	✓	S	1K	10K	9	4	4	LP
IncDE (Liu et al., 2024a)	General	✗	S	15K	310K	237	5	5	LP
ICEWS (García-Durán et al., 2018)	Political	✓	S	13K	386K	256	–	–	LP
MRCKG (Li et al., 2026)	General	✗	S,T,V	15K	99K	279	5	18	LP
PrimeKG-CL	**Biomedical**	✓	**S,T,M**	**129K+**	**8.1M+**	**30**	**6**	**10**	**LP, KGQA, NC**

across strategies — with a decoder-conditional CL interaction (no strategy is best across decoders; mismatches degrade performance) and decoder-conditional correct forgetting (DistMult separates persistent from removed knowledge; RotatE does not). Multimodal fusion yields large entity-classification gains, and IncDE fails to scale to our base task. Sections 4.2, 4.3 and 4.5 report these findings.

Contributions. (1) PrimeKG-CL, the first CGL benchmark on a real biomedical KG with genuine $t_0 \to t_1$ evolution (129K+ nodes, 8.1M+ edges, six real-evolution tasks), with multimodal features and a persistent/removed/added test stratification. (2) A 6×4 CL×decoder matrix plus LKGE, CMKL, and an LLM-RAG agent (ten methods; Joint Training as oracle), across three tasks, 5 seeds, 5 CL metrics. (3) Decoder dominance as the primary axis of variation, with a decoder-conditional CL interaction and decoder-conditional correct forgetting. (4) Open release of data, pipeline, baselines, and stratified split.

2 Related Work

Continual Learning for Knowledge Graphs. Continual learning sequentially learns from a stream of tasks without catastrophically forgetting prior knowledge (Kirkpatrick et al., 2017). LKGE (Cui et al., 2023) introduced a lifelong KG embedding framework applying EWC, experience replay, and distillation to TransE, DistMult, and RotatE (Bordes et al., 2013; Yang et al., 2015; Sun et al., 2019) on synthetic splits of FB15k-237 and WN18RR; Daruna et al. (Daruna et al., 2021) apply continual KGE to a robot-manipulation KG; IncDE (Liu et al., 2024a) uses incremental distillation; Synaptic Intelligence (Zenke et al., 2017) tracks importance online. More recent methods include FastKGE (Liu et al., 2024b) (LoRA adapters), ERPP (Yang et al., 2025) (evolutionary relation path passing), SAGE (Li et al., 2025) (scale-aware embedding), DebiasedKGE (Zhu et al., 2025) (disentangled learning), and workshop work on EWC for KG link prediction (Jhajj & Lin, 2025). A recent study (Zhao et al., 2025) rethinks CKGE benchmarks and finds that synthetic partitioning obscures pattern shifts between snapshots, directly motivating our entity-type-grouped task design. Nearly all existing continual KGE methods are evaluated on synthetically partitioned generic KGs, leaving their effectiveness on real-world biomedical KG evolution untested. Broader continual graph learning (Zhang et al., 2022) targets citation or social networks with node or graph classification, not KG-specific tasks. Concurrent work (Li et al., 2026) constructs multimodal continual KG benchmarks from DB15K, MKG-W, and MKG-Y using synthetic temporal splits; these general-domain datasets have $\leq$15K entities and no real temporal dynamics.

Biomedical Knowledge Graphs. PrimeKG (Chandak et al., 2023) integrates 20+ databases into a unified schema with 129K+ nodes across 10 biomedical entity types. TxGNN (Huang et al., 2024) leverages PrimeKG for therapeutic use prediction. Hetionet (Himmelstein et al., 2017) is an earlier integrative biomedical network. BioMedKG (Dang et al., 2025) enriches PrimeKG with biological sequences and textual descriptions, and PrimeKGQA (Yan et al., 2024) provides question-answer pairs for biomedical QA. None has been studied in a continual learning setting with real temporal evolution. Temporal KGs such as ICEWS (García-Durán et al., 2018) model event-based knowledge with explicit timestamps; in contrast, biomedical KGs undergo *structural* evolution (entities and relation types are added or removed as scientific understanding advances), which PrimeKG-CL captures.

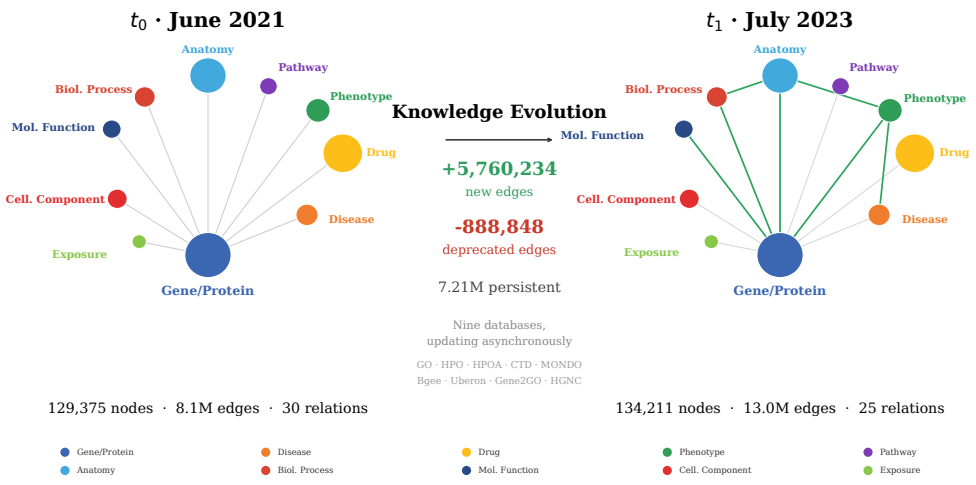

Figure 1: PrimeKG-CL knowledge evolution from t_0 (June 2021) to t_1 (July 2023): 5.76M new edges, 889K removed (t_0-only), 7.21M persistent. Node size encodes per-entity-type frequency at each snapshot.

Multimodal Knowledge Graph Embedding. Multimodal KGE combines structural, textual, and visual features: MMKGR (Zheng et al., 2023) (gated cross-modal attention), MCLEA (Lin et al., 2022) (contrastive multimodal alignment), MoSE (Zhao et al., 2022) (score-level ensemble), and OGM-GE (Peng et al., 2022) (gradient modulation). Wu et al. (Wu et al., 2022) identify the "greedy modality" phenomenon where dominant modalities suppress others. These works operate in the static setting; PrimeKG-CL enables studying multimodal KGE under continual learning, where modality-specific forgetting patterns emerge.

Knowledge Graphs and Language Models. Recent work highlights the complementarity of KGs and large language models (LLMs) (Pan et al., 2024). Our benchmark includes a RAG baseline (Lewis et al., 2020) combining KG retrieval with Qwen2.5-7B for continual KGQA.

In contrast to prior CGL benchmarks, PrimeKG-CL provides the first evaluation framework on a real-world biomedical knowledge graph with genuine temporal evolution and multimodal node features, encompassing three evaluation tracks and ten methods across four KGE decoders.

3 Benchmark Construction

PrimeKG-CL turns real biomedical knowledge evolution into a continual-learning workload (Figure 1), beyond the temporal structure of synthetic splits.

3.1 Temporal Snapshots

We construct two temporal snapshots of PrimeKG (Chandak et al., 2023). t_0 **(June 2021)** is the original PrimeKG release (129,375 nodes, 10 entity types, 8,100,498 edges, 30 relation types; integrating 20+ databases including DrugBank, DisGeNET, GO, HPO, Reactome, MONDO).

t_1 **(July 2023)** is reconstructed by re-querying nine freely accessible databases (Bgee, CTD, GO, Gene2GO, HPO, HPOA, MONDO, Uberon, HGNC), which update asynchronously at independent cadences (GO monthly, HPO quarterly, MONDO continuously); the resulting snapshot has 134,211 nodes and 13,001,666 edges across 25 relation types. Seven databases (DrugBank, UMLS, DrugCentral, SIDER, Reactome, and protein–protein-interaction sources under restrictive licensing, plus DisGeNET under API changes) were not re-queried, so relation types sourced from these carry-forward databases have zero turnover by construction; consequently the real $t_0 \to t_1$ evolution covers additions in the freshly re-queried gene/protein, disease, phenotype, biological-process, and anatomy/pathway groups, and drug edges are carried forward unchanged (no drug-evolution task).

Table 2: Dataset statistics for the two temporal snapshots in PrimeKG-CL. The temporal difference reveals substantial real-world knowledge evolution between June 2021 and July 2023.

Statistic	t_0 **(June 2021)**	t_1 **(July 2023)**	**Difference**
Nodes	129,375	134,211	+4,836
Edges	8,100,498	13,001,666	+4,901,168
Node types	10	10	0
Relation types	30	25	-5
Added edges	–	–	+5,760,234
Removed edges	–	–	$-888,848$
Persistent edges	–	–	7,208,624
Continual tasks	6 (entity-type grouping, real $t_0 \to t_1$ evolution)		
Evaluation tracks	LP, KGQA, Node Classification		

Added/removed/persistent are the deduplicated $t_0 \to t_1$ diff; whole-graph edge totals and the subgraph diff use different scopings, so the rows are not expected to sum exactly.

The added (5.76M) and removed (889K) edges are computed on the re-queried subgraph; the 7.21M persistent count is a whole-graph total that includes frozen carry-forward edges, of which ≈3.37M lie within the re-queried subgraph (a 20.9% within-subgraph turnover). The temporal-evolution analysis (§3.2) and stratified evaluation (Table 5) are defined on this re-queried subgraph. July 2023 is the most recent point at which all nine databases could be reconstructed from archived, version-pinned releases at build time; the pipeline supports extension to additional snapshots as newer archived releases become available. Per-snapshot statistics are in Table 2.

3.2 Temporal Difference

The temporal difference between t_0 and t_1 reveals substantial real-world knowledge evolution:

- **Added edges:** 5,760,234 new triples appear in t_1 that were absent in t_0, reflecting newly discovered associations (e.g., new gene-GO annotations, updated disease-phenotype links).
- **Removed edges:** 888,848 triples present in t_0 are not observed in t_1 (e.g., gene–disease associations dropped after re-curation, updated ontology hierarchies). We characterize rather than assume retraction: for 82.1% of removed edges both endpoint entities still exist in t_1 (edge-level change, a retraction *candidate*), whereas ≈18% coincide with an endpoint disappearing (entity-level churn / ID reconciliation). A "removed" edge indicates the association is not present in the t_1 snapshot, not that it is biologically false.
- **Persistent edges:** 7,208,624 triples remain unchanged across both snapshots, providing a stable knowledge backbone.

This temporal difference is fundamentally different from synthetic CGL benchmarks that randomly partition a static graph. The additions and removals follow domain-specific patterns: anatomy–protein expression edges (Bgee) account for the largest share of added edges (≈95% of additions, reflecting a large expression-atlas refresh), with Gene Ontology functional annotations and disease–phenotype associations forming smaller strata whose links show high turnover as GO, HPO, and MONDO ontologies are refined.

Stratified evaluation on added, removed, and persistent edges. The $t_0 \to t_1$ partition directly supports stratified continual-learning analysis. In particular, the test split of the base task (1,620,099 triples) stratifies cleanly into 1,443,243 *persistent* triples (89.1%, present in both snapshots, on which a correctly updating model should *retain* its predictive ability) and 176,856 *removed* triples (10.9%, present in t_0 but deprecated in t_1, on which an ideal model should *unlearn* its initial learned association). The later tasks provide the *added* stratum: t_1-only triples grouped by entity type. We release this three-way stratification (`test_stratification.json`) with the benchmark so that future work can report per-stratum MRR, quantifying correct forgetting of deprecated knowledge alongside the standard AP/AF/BWT/REM metrics.

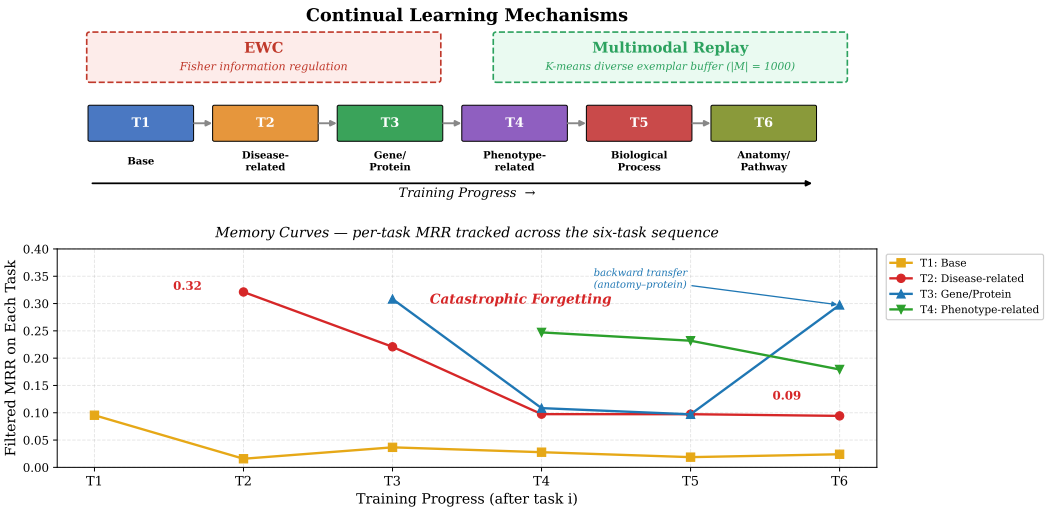

Figure 2: The six-task continual-learning sequence on PrimeKG-CL (top) and per-task memory curves for CMKL+DistMult under EWC + multimodal replay ($|M|$=1000) (bottom). Disease-related knowledge peaks at MRR ≈ 0.32 when first learned and degrades to ≈ 0.09 by the final task, a $3.4\times$ collapse despite EWC and replay mitigations.

3.3 Continual Task Sequence

We organize the continual learning task sequence using an *entity-type grouping* strategy. Rather than randomly partitioning triples into artificial time steps (as in LKGE (Cui et al., 2023)), we group the real $t_0 \to t_1$ additions by their head and tail entity types to create semantically coherent tasks. Specifically, the benchmark comprises six continual tasks: a base task on the full t_0 graph, followed by the five entity-type groups that receive genuine additions in the re-queried t_1 snapshot:

1. **Base** (t_0): the full June-2021 graph, on which the persistent/removed stratification is defined.
2. **Disease-related**: disease–disease and disease–phenotype associations (MONDO/HPO refinement).
3. **Gene/Protein**: protein–protein interactions and gene functional links.
4. **Phenotype-related**: phenotype–phenotype associations (HPO).
5. **Biological Process**: gene/protein–GO functional annotations.
6. **Anatomy/Pathway**: tissue-expression and pathway-membership edges (the largest addition source, driven by the Bgee anatomy–protein expression refresh).

Because the drug databases (DrugBank, DrugCentral, SIDER) were not re-queried, drug edges are carried forward unchanged and do not form a real-evolution task.

Each task is presented sequentially over the $t_0 \to t_1$ transition (Figure 2). The cost of getting this wrong is concrete and large: even the strongest multimodal model on this benchmark (CMKL with DistMult) sees Disease-related knowledge peak at MRR ≈ 0.32 when first learned and decay to ≈ 0.09 by the final task, a $3.4\times$ collapse on exactly the relations a clinical decision-support system would query. The grouping mirrors how upstream databases push updates (GO functional annotations, HPO phenotype links), creating a dual challenge of *task identity shifts* and *temporal adaptation* ($t_0 \to t_1$ within each task), analogous to domain-incremental CL in vision (van de Ven & Tolias, 2019).

3.4 Evaluation Tasks

PrimeKG-CL evaluates link prediction, KGQA, and node classification, reflecting missed drug candidates, stale clinical answers, and misclassified entities.

Biomedical Relationship Prediction (Link Prediction, Primary). For $(h, r, ?)$ or $(?, r, t)$, models rank all entities using filtered MRR $a_{j,i}$ (Bordes et al., 2013), with Hits@K ($K \in \{1, 3, 10\}$) in the released evaluation code. Applications include drug repurposing, target identification, and disease-gene prioritization.

Letting $a_{j,i}$ be the filtered MRR on task i after training through task j over T tasks, we report the standard CL metrics: AP $= \frac{1}{T}\sum_i a_{T,i}$, AF $= \frac{1}{T-1}\sum_{i<T} \max_j(a_{j,i} - a_{T,i})$, BWT $= \frac{1}{T-1}\sum_{i<T}(a_{T,i} - a_{i,i})$, and REM $= 1 -$ AF.

Knowledge Graph Question Answering (KGQA). Our RAG pipeline (Lewis et al., 2020) retrieves current-snapshot subgraphs for an LLM answering KGQA queries; we report Exact Match, token-level F1, and accuracy.

Biomedical Entity Classification (Node Classification). Node classification predicts PrimeKG entity types (e.g., gene/protein, disease, drug) from structural and optional multimodal features. It parallels biomedical entity recognition, assigning graph-entity types rather than locating text spans. We report Macro-F1 and CL metrics; structural-only AP≈ 0.34–0.37 indicates a non-trivial task.

3.5 Multimodal Features

PrimeKG-CL ships three node-feature modalities: **textual** (768-d BiomedBERT (Gu et al., 2021) [CLS] embeddings of entity descriptions), **molecular** (1024-bit radius-2 Morgan fingerprints (Rogers & Hahn, 2010) for drug nodes with SMILES, projected through a learned MLP), and **structural** (R-GCN (Schlichtkrull et al., 2018) embeddings capturing multi-relational neighborhood context). Together they enable study of modality-specific forgetting, a direction impossible on existing structural-only CGL benchmarks. To avoid leakage, the KGE baselines use no node features (structure only), and the R-GCN structural encoder is fit on the train+validation edges of each task with test edges held out; text and molecular features derive only from entity descriptions and SMILES, independent of the link labels. Node identities recur across the sequence (a transductive setting), which we treat as a property of continual KG evolution rather than leakage.

4 Experiments

4.1 Experimental Setup

Baselines. Ten methods spanning the major CL families: **Naive Sequential** (lower bound), **Joint Training** (oracle), regularization (**EWC** (Kirkpatrick et al., 2017), **SI** (Zenke et al., 2017)), replay (**Experience Replay** (Rolnick et al., 2019), **MIR** (Aljundi et al., 2019)), **Distillation** (Hinton et al., 2015), the architecture-plus-distillation framework **LKGE** (Cui et al., 2023), a **RAG** agent (Qwen2.5-7B) (Lewis et al., 2020), and **CMKL** (R-GCN + DistMult, MoE fusion (Zhao et al., 2022), modality-aware EWC, multimodal replay). Each applicable method is run with four KGE decoders: TransE (Bordes et al., 2013), DistMult (Yang et al., 2015), ComplEx (Trouillon et al., 2016), RotatE (Sun et al., 2019), via PyKEEN. Per-method details, and methods that exceed a 350 GB RAM budget on our 5.67M-triple base task (FastKGE (Liu et al., 2024b), ERPP (Yang et al., 2025), SAGE (Li et al., 2025), IncDE (Liu et al., 2024a)), are in Section S3.

Setup. $d = 256$, $\eta = 0.001$, batch 512, 100 epochs/task, Adam; 70/10/20 train/valid/test per task fixed across methods and seeds. EWC $\lambda = 10$; replay buffer 1,000 triples, K-means diverse selection; CMKL uses the same buffer size, MoE fusion, per-modality Fisher. λ-sweep over $50\times$ and buffer 500–5,000 produced $<$10% AP variation. Per-decoder hyperparameters are shared across CL strategies; we additionally re-tune the Naive baseline per-cell for the ComplEx decoder (learning-rate sweep) to confirm the EWC$\times$ComplEx effect is not a shared-budget artifact (Section S3). NVIDIA V100 (32 GB), 5 seeds $\{42, 123, 456, 789, 1024\}$, mean $\pm$ s.d.; significance by paired t-tests ($p < 0.05$) and 95% bootstrap confidence intervals (10k resamples).

Table 3: Continual biomedical relationship prediction on PrimeKG-CL: AP = filtered MRR over 129K+ entities, averaged over the **six real-evolution tasks** (mean ± std, 5 seeds). **Bold** = best CL-strategy row per column; underline = second (oracle and CMKL shown in separate blocks); *italic* = oracle. Decoder choice dominates CL strategy (RotatE ≫ DistMult ≫ ComplEx ≫ TransE). AF, BWT, REM in Section S2. Discussion in Section 4.2.

CL Strategy	TransE	DistMult	ComplEx	RotatE
Naive Sequential	0.007 ± 0.000	**0.096** ± 0.002	0.018 ± 0.001	0.139 ± 0.007
EWC (Kirkpatrick et al., 2017)	0.007 ± 0.000	**0.096** ± 0.001	**0.048** ± 0.003	**0.146** ± 0.005
SI (Zenke et al., 2017)	**0.009** ± 0.000	0.061 ± 0.001	0.031 ± 0.002	0.088 ± 0.004
Distillation (Hinton et al., 2015)	0.008 ± 0.000	0.052 ± 0.001	0.029 ± 0.002	0.059 ± 0.000
Experience Replay (Rolnick et al., 2019)	0.006 ± 0.000	0.085 ± 0.002	0.011 ± 0.000	0.144 ± 0.006
MIR (Aljundi et al., 2019)	0.005 ± 0.000	0.085 ± 0.001	0.011 ± 0.000	0.144 ± 0.007
Joint Training (oracle)	—	0.078 ± 0.002	0.041 ± 0.001	0.272 ± 0.001
Multimodal (DistMult only)				
CMKL (MoE, multimodal)	—	0.103 ± 0.016	—	—

AP is reported on the six real $t_0 \to t_1$ evolution tasks (§3); the oracle recovers the upper bound on RotatE and ComplEx, and sits at or slightly below Naive on the symmetric DistMult decoder. LKGE (Cui et al., 2023) (TransE) was run under its own 9-task protocol (no base snapshot) and reaches AP 0.067 ± 0.002 on its real subset; it is not directly comparable to the aligned rows above. RAG (Qwen2.5-7B) is decoder-free and reported on the KGQA track only (§4.4).

4.2 Biomedical Relationship Prediction (Link Prediction) Results

Table 3 reports filtered MRR over all 129K+ candidate entities. Contrary to the assumption that CL strategy dominates such a matrix, the column choice (the KGE decoder) moves AP by $20\times$, while the six CL strategies within any given column move it by up to $\approx 4\times$ and in opposite directions across columns. The remainder of this section unpacks this interaction.

Per-task dynamics and the difficulty of real biomedical KGs. Per-task peak MRR reaches ≥ 0.25 on several tasks (e.g., CMKL: Disease-related 0.32, Gene/Protein 0.31, Phenotype-related 0.25), confirming meaningful patterns are learned even when AP is low. Three factors keep absolute AP modest: 129K+ candidate entities (KG-FIT (Jiang et al., 2024) reports TransE MRR$= 0.048$ on a 10K-entity PrimeKG subset), 10 entity $\times$ 30 relation types of mixed semantics, and PrimeKG sparsity ($<1\%$ of possible edges) leaving many true positives outside the filter set (Sun et al., 2019).

RotatE dominates the decoder spectrum. The single largest effect in Table 3 is the choice of decoder: on Naive Sequential, the four decoders span $\approx 20\times$ — TransE (0.007) $\ll$ ComplEx (0.018) $<$ DistMult (0.096) $<$ RotatE (0.139, 45% above DistMult). We attribute this to PrimeKG-CL's compositional and approximately rotational regularities (gene-gene interactions, phenotype hierarchies, protein-protein symmetries) that rotational scoring captures but translational/bilinear cannot, unlike FB15k-237/WN18RR where DistMult and RotatE typically tie. TransE sits near the $\sim$129K-entity random-ranking floor and is likely under-tuned for this candidate set, so we also report the span excluding it: RotatE/ComplEx is $\approx 8\times$ and RotatE/DistMult $\approx 1.45\times$, still a large, decoder-driven gap. *Researchers must control for decoder choice when comparing CL strategies on biomedical KGs.*

The decoder-conditional CL interaction. The interaction between decoder family and CL strategy is a central benchmark finding: a CL strategy that helps one decoder can actively harm another, with effects large enough to erase the decoder's advantage (visualized in Figure 3). **EWC** is universal but decoder-sensitive in magnitude: invisible on TransE/DistMult, transformative on ComplEx ($0.018 \to 0.048$, +167%, $p < 10^{-5}$), and a small significant boost on RotatE ($0.139 \to 0.146$, +5%, $p = 0.013$), with EWC + RotatE the best configuration overall. The ComplEx gain holds under per-cell learning-rate tuning of the Naive baseline (Section S3), so it reflects a genuine decoder-conditional effect at a matched training budget rather than under-tuning. **SI and Distillation** degrade rotational decoders, dropping RotatE AP to 0.088 (-37%) and 0.059 (-58%) respectively while achieving AF≈ 0 (see Table S1): they prevent the rotational geometry from adapting. **Replay** (ER

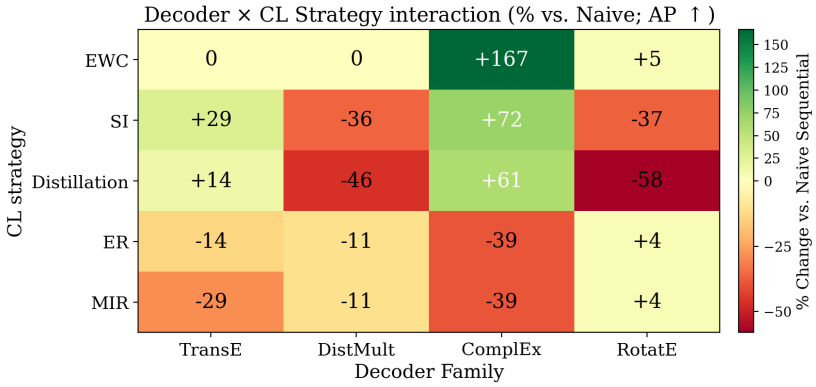

Figure 3: Decoder × CL interaction (AP relative to Naive, per column): EWC transforms ComplEx (+167%); SI/Distillation drop RotatE by 37%/58%; replay drops ComplEx by 39%.

Table 4: Continual biomedical entity classification on PrimeKG-CL (Macro-F1, mean ± std over 5 seeds). LKGE and RAG are excluded as not applicable to this track.

Method	**AP** ↑	**AF** ↓	**BWT** ↑
Naive Sequential	0.344 ± 0.004	0.011 ± 0.003	0.003 ± 0.005
Joint Training	0.370 ± 0.002	$\mathbf{0.003} \pm 0.004$	$\mathbf{0.022} \pm 0.004$
EWC (Kirkpatrick et al., 2017)	0.345 ± 0.004	0.008 ± 0.003	0.007 ± 0.004
Experience Replay (Rolnick et al., 2019)	0.344 ± 0.006	0.010 ± 0.003	0.006 ± 0.005
SI (Zenke et al., 2017)	0.362 ± 0.004	0.007 ± 0.007	0.015 ± 0.006
Distillation (Hinton et al., 2015)	0.362 ± 0.004	0.007 ± 0.007	0.015 ± 0.006
MIR (Aljundi et al., 2019)	0.362 ± 0.004	0.007 ± 0.007	0.015 ± 0.006
CMKL (MoE)	$\mathbf{0.591} \pm 0.006$	0.008 ± 0.010	$+0.003 \pm 0.008$

Feature-matched control. The CL-strategy rows above are structural-only. Giving Naive and EWC the *same* multimodal features raises them to 0.414 ± 0.001 and 0.414 ± 0.003; a frozen-text-only classifier reaches 0.246 ± 0.002 (ruling out label leakage). CMKL's $+0.177$ Macro-F1 over the feature-matched CL baselines ($+43\%$, paired $t = 77$, $p = 1.7 \times 10^{-7}$, 5 seeds) isolates the fusion architecture from feature availability.

and MIR) matches EWC on RotatE (≈ 0.144) but drops ComplEx to 0.011, 39% below Naive, consistent with reported ComplEx fragility under noisy negatives (Ruffinelli et al., 2020). There is no universal best CL strategy: decoder-agnostic CL benchmarking can be actively misleading.

Joint Training and CMKL. Joint Training, the oracle, is itself decoder-conditional: it recovers a clear upper bound on RotatE (AP$= 0.272$, nearly double the best CL configuration) and ComplEx (0.041, well above Naive's 0.018), but on the symmetric bilinear DistMult it sits at or slightly below sequential (AP$= 0.078$ vs Naive's 0.096), and adding epochs does not close the gap—pooled optimization over heterogeneous relations does not benefit DistMult (Yu et al., 2020; Riemer et al., 2019). The same-decoder multimodal CMKL (R-GCN+DistMult) achieves AP$= 0.103 \pm 0.016$ ($+7\%$ over DistMult-Naive/EWC at 0.096); it is the strongest *multimodal* method on link prediction but not the overall best (EWC+RotatE: 0.146 ± 0.005, 42% above CMKL-DistMult), indicating that decoder and modality contribute independently and that the largest LP headroom lies with the rotational decoder.

4.3 Biomedical Entity Classification (Node Classification) Results

CMKL achieves the highest AP (0.591 ± 0.006), with near-zero forgetting (AF$= 0.008 \pm 0.010$, BWT$= +0.003$). To separate architecture from feature availability, we give Naive and EWC the *same* multimodal features: they reach only 0.414, so CMKL's $+0.177$ Macro-F1 ($+43\%$, paired $t = 77$, $p < 10^{-6}$; Table 4) is an architectural gain, not merely a consequence of access to features. A frozen-text-only classifier reaches 0.246, ruling out label leakage. All methods show positive BWT, indicating that learning later associations slightly improves classification of earlier-seen entities.

Table 5: Stratified filtered MRR on the final-task test set. **Persistent** triples are in both t_0 and t_1 (a correctly updating model should retain them); **removed** triples are in t_0 but deprecated in t_1 (an ideal model should forget them: lower is better); **added** triples are new in t_1 (mean MRR across the five real added tasks). Values are mean $\pm$ std over 5 seeds.

Method	Decoder	Persistent ↑	Removed ↓	Added ↑
Naive Sequential	DistMult	0.096 ± 0.005	0.009 ± 0.000	0.099 ± 0.002
Naive Sequential	RotatE	0.106 ± 0.003	0.078 ± 0.002	0.147 ± 0.008
EWC	DistMult	0.095 ± 0.007	0.012 ± 0.001	0.098 ± 0.002
EWC	RotatE	0.106 ± 0.002	0.078 ± 0.003	0.154 ± 0.006
Joint Training	DistMult	0.213 ± 0.006	0.097 ± 0.004	0.082 ± 0.003
Joint Training	RotatE	0.231 ± 0.004	0.160 ± 0.003	0.311 ± 0.001

4.4 Biomedical Knowledge Graph Question Answering (KGQA) Results

We position the KGQA track as a stress test rather than a finished evaluation. All three configurations of Qwen2.5-7B-Instruct sit at or below token F1 $= 0.015$, with full RAG (≤ 0.015) only marginally above retrieval-only (0.003 ± 0.001) and zero-shot (0.000). Forgetting is essentially absent (AF ≤ 0.001) because there is little performance to forget. KGQA on a real, 129K-entity biomedical graph is a regime where current RAG pipelines are barely above retrieval; we therefore present this track as unsolved rather than saturated, calibrated to register future LLM-RAG improvements. Per-task breakdown and prompt templates are in Section S5.

4.5 Analysis

Per-task heterogeneity and entity-type-specific forgetting. Per-task peak MRR is concentrated on three tasks (Figure S1): Disease-related, Gene/Protein, Phenotype-related each exceed 0.1; the others remain near zero (heterogeneous *semantic* difficulty rather than scale). Forgetting is similarly concentrated (Figure S2): CMKL puts 56% on Disease-related (0.227), while Naive and EWC put 70% on Base (t_0) (0.152). Gene/protein entities make up 73% of triples, biasing performance toward gene/protein-centric tasks.

Correct forgetting is decoder-conditional. Standard AP and AF conflate two desiderata: *retaining persistent* knowledge (still true at t_1) and *unlearning deprecated* knowledge (only true at t_0). We split the base-task test set into $1{,}443{,}243$ persistent and $176{,}856$ removed triples, with a separate *added* stratum of $1{,}156{,}876$ test triples from the five real added tasks newly introduced at t_1, and report per-stratum filtered MRR over 5 seeds in Table 5. The pattern is again decoder-conditional. Under DistMult, Naive's drift yields a clean correct-forgetting signal (persistent/removed ratio $11\times$, 0.096 vs 0.009); EWC mildly over-protects deprecated edges ($8\times$), its uniform Fisher penalty unable to separate still-true from deprecated associations; Joint Training keeps both strata together ($2.2\times$). Under RotatE all three collapse to $\approx 1.4\times$: rotational geometry retains patterns uniformly, dissolving the signal DistMult-Naive achieves through drift. We release the split (`test_stratification.json`) so future methods can target a metric current baselines optimize only by accident.

5 Conclusion

We introduced PrimeKG-CL, a biomedical CGL benchmark with genuine temporal evolution and multimodal features. Decoder choice dominates the continual-learning strategy and the two interact, so ignoring the decoder family confounds comparisons; the persistent/removed/added stratification further shows that AP/AF average over correct and incorrect forgetting. Data, pipeline, baselines, and stratification will be released; Sections S7 and S8 cover limitations and availability.

AI USE STATEMENT

Generative AI tools, including Claude Code, assisted with literature review, idea suggestions used by the authors as starting points, implementation of routine experiments, and interpretation of individual experimental results. More complex experiments and synthesis of the findings across experiments involved direct human input. The benchmark was constructed from existing biomedical data sources, rather than AI-generated biomedical records. Separately, the paper evaluates language-model-based knowledge-graph question answering, as described in the methods. AI assisted with manuscript drafting and producing figures from data already verified by the authors. Codex assisted with submission formatting and manuscript review. The authors reviewed AI outputs at key stages for correctness and appropriateness and take responsibility for the final content of this work.

ETHICS STATEMENT

PrimeKG-CL is constructed entirely from publicly available biomedical databases and contains no private, patient-identifiable, or human-subjects data. Redistribution respects upstream licenses: content from permissively-licensed sources is released under CC BY 4.0, while restrictively-licensed sources are not redistributed and are instead reconstructed by a provided script under each user's own accepted terms (Section S8). The benchmark is intended for methodological research on continual graph learning; predictions from models trained on it are research artifacts and must not be used for clinical or therapeutic decisions without expert validation. In particular, a "removed" edge denotes absence from the t_1 snapshot (e.g., re-curation or ontology revision), not established biological falsity, and should not be interpreted as evidence that an association is untrue.

REPRODUCIBILITY STATEMENT

To support reproducibility, the paper fully specifies the experimental setup: the two-snapshot construction and entity-type task design (Section 3), the 70/10/20 train/validation/test splits with fixed seeds, and all decoder and continual-learning hyperparameters (Section 4.1, Section S3). All reported results are mean $\pm$ standard deviation over five seeds $\{42, 123, 456, 789, 1024\}$ on a single NVIDIA V100 (32 GB), using filtered MRR over the full 129K+ entity set with the PyKEEN convention. The data, construction pipeline, task splits, persistent/removed/added stratification metadata, and reference implementations of all baselines and CMKL will be released openly (Section S8).

REFERENCES

Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximally interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023.

Yuanning Cui, Yuxin Wang, Zequn Sun, Wenqiang Liu, Yiqiao Jiang, Kexin Han, and Wei Hu. Lifelong embedding learning and transfer for growing knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4217–4224, 2023.

Tien Dang, Viet Thanh Duy Nguyen, Minh Tuan Le, and Truong-Son Hy. Multimodal contrastive representation learning in augmented biomedical knowledge graphs. *arXiv preprint arXiv:2501.01644*, 2025.

Angel Daruna, Mehul Gupta, Mohan Sridharan, and Sonia Chernova. Continual learning of knowledge graph embeddings. *IEEE Robotics and Automation Letters*, 6(2):1128–1135, 2021.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4816–4821, 2018.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

Gaganpreet Jhajj and Fuhua Lin. Elastic weight consolidation for knowledge graph continual learning: An empirical evaluation. *arXiv preprint arXiv:2512.01890*, 2025.

Pengcheng Jiang, Lang Cao, Cao Xiao, Parminder Bhatia, Jimeng Sun, and Jiawei Han. KG-FIT: Knowledge graph fine-tuning upon open-world knowledge. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114 (13):3521–3526, 2017.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Linyu Li, Zhi Jin, Yichi Zhang, Dongming Jin, Yuanpeng He, Haoran Duan, Gadeng Luosang, and Nyima Tashi. When modalities remember: Continual learning for multimodal knowledge graphs. *arXiv preprint arXiv:2604.02778*, 2026.

Yifei Li, Lingling Zhang, Hang Yan, Tianzhe Zhao, Zihan Ma, Muye Huang, and Jun Liu. SAGE: Scale-aware gradual evolution for continual knowledge graph embedding. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2572–2584, 2022.

Jiajun Liu, Wenjun Ke, Peng Wang, Ziyu Shang, Jinhua Gao, Guozheng Li, Ke Ji, and Yanhe Liu. Towards continual knowledge graph embedding via incremental distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 8759–8768, 2024a.

Jiajun Liu, Wenjun Ke, Peng Wang, Jiahao Wang, Jinhua Gao, Ziyu Shang, Guozheng Li, Zijie Xu, Ke Ji, and Yining Li. Fast and continual knowledge graph embedding via incremental LoRA. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 2198–2206, 2024b.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8238–8247, 2022.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations*, 2019.

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

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. You CAN teach an old dog new tricks! on training knowledge graph embeddings. In *Proceedings of the 8th International Conference on Learning Representations*, 2020.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference (ESWC 2018)*, pp. 593–607. Springer, 2018.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *Proceedings of the 7th International Conference on Learning Representations*, 2019.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2071–2080, 2016.

Gido M van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019.

Nan Wu, Stanisław Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 24043–24055. PMLR, 2022.

Xi Yan, Patrick Westphal, Jan Seliger, and Ricardo Usbeck. Bridging the gap: Generating a comprehensive biomedical knowledge graph question answering dataset. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pp. 1198–1205, 2024.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Jing Yang, Xinfa Jiang, Xiaowen Jiang, Yuan Gao, Laurence T Yang, Shaojun Zou, and Shundong Yang. From knowledge forgetting to accumulation: Evolutionary relation path passing for lifelong knowledge graph embedding. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1197–1206, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Xikun Zhang, Dongjin Song, and Dacheng Tao. Cglb: Benchmark tasks for continual graph learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13006–13021, 2022.

Tianzhe Zhao, Jiaoyan Chen, Yanchi Ru, Qika Lin, Yuxia Geng, Haiping Zhu, Yudai Pan, and Jun Liu. Rethinking continual knowledge graph embedding: Benchmarks and analysis. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2025.

Yu Zhao, Xiangrui Cai, Yike Wu, Haiwei Zhang, Ying Zhang, Guoqing Zhao, and Ning Jiang. Mose: Modality split and ensemble for multimodal knowledge graph completion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10527–10536, 2022.

Shangfei Zheng, Weiqing Wang, Jianfeng Qu, Hongzhi Yin, Wei Chen, and Lei Zhao. MMKGR: Multi-hop multi-modal knowledge graph reasoning. In *Proceedings of the 39th IEEE International Conference on Data Engineering (ICDE)*, pp. 96–109, 2023.

Junlin Zhu, Bo Fu, and Guiduo Duan. DebiasedKGE: Towards mitigating spurious forgetting in continual knowledge graph embedding. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM)*, 2025. doi: 10.1145/3746252.3761387.

PrimeKG-CL: A Continual Graph Learning Benchmark on Evolving Biomedical Knowledge Graphs

Supplementary Material

S1 Experimental Details

All experiments use NVIDIA V100 GPUs (32 GB) with PyTorch 2.0 and PyKEEN (exact version pinned in the released environment specification). Key hyperparameters: embedding dimension 256, learning rate 0.001, batch size 512, 100 epochs per task, 50,000 sampled triples per epoch. For structural-only baselines, a single EWC penalty $\lambda = 10$ is used. For CMKL, modality-specific lambdas are used: $\lambda_s = 10$ (structural), $\lambda_t = 5$ (textual), $\lambda_m = 1$ (molecular), $\lambda_f = 5$ (fusion), $\lambda_r = 50$ (main relation embeddings, set high because rotational decoders treat relations as rotation angles where any drift warps the full embedding geometry). Replay buffer: 1,000 triples total via K-means selection on structural embeddings. ComplEx and RotatE decoders follow the PyKEEN convention (entity dim $= 2D$, D complex pairs interleaved); RotatE relations are D phases initialized uniformly in $[-\pi, \pi]$, ComplEx relations are full-complex $2D$ real values initialized with Xavier. All results averaged over 5 seeds $\{42, 123, 456, 789, 1024\}$.

Scoring-function corrections (2026-04-12). During a mid-revision audit we identified and fixed four bugs in the PyKEEN-based scoring pipeline that had affected earlier drafts. (i) **RotatE complex-tensor handling**: raw PyKEEN entity embeddings are stored as $[N, 2D]$ real, which must be converted to $[N, D]$ complex via `torch.view_as_complex`; our prior implementation treated the raw tensor as already-complex, silently corrupting the head embedding. (ii) **RotatE evaluation norm**: eval used L_1 instead of L_2 on the complex difference, inconsistent with training. (iii) **ComplEx**: the same raw-to-complex conversion bug as RotatE. (iv) **TransE training norm**: training used L_2 where the PyKEEN convention for TransE is L_1; re-running confirmed the numerical effect is within seed noise for TransE, so reported TransE numbers are unchanged. All four corrections were verified against PyKEEN's `model.score_hrt` reference. The new RotatE and ComplEx numbers reported here post-date these fixes; we re-ran the full 4-decoder matrix under the corrected pipeline.

IncDE scalability note. We attempted to evaluate IncDE (Liu et al., 2024a), a representative incremental-distillation CKGE method, following the official repository. Over five attempts (including the bug fixes listed in the paper's GitHub issue tracker) we observed out-of-memory failures at 64 GB, 200 GB, and 350 GB RAM, and deadlocks in the training loop on the 5.67 M-triple base task. In our hands, the published implementation does not scale to the task-0 graph of PrimeKG-CL. We cannot rule out implementation differences relative to the authors' environment; we document our configurations alongside the released code. This negative result is informative: it highlights a gap between general-domain CKGE benchmarks ($\leq$15 K entities) and real biomedical graphs where methods that appear scalable on paper can become intractable in practice.

S2 Additional Results

Evaluation scope. PrimeKG-CL measures continual learning on genuine $t_0 \rightarrow t_1$ knowledge evolution, so its metrics are defined over the tasks that receive real additions from the freshly re-queried databases: the base task plus the five re-queried entity-type groups (Disease, Gene/Protein, Phenotype, Biological Process, Anatomy/Pathway). Entity-type groups whose source databases were not re-queried (the drug group; see §3.1) carry no genuine temporal signal and are therefore not part of the evolution workload; because such groups contribute no learnable evolution, their presence during fine-tuning does not advantage the reported numbers. AP, AF, and BWT are all reported over this real-evolution task set. Forward Transfer (FWT) equals zero for all continual methods in our sequential setting, as tasks are non-overlapping and models are not evaluated on unseen tasks before training. Hits@1, Hits@3, and Hits@10 are available in the released evaluation code.

Table S1: **Average Forgetting (AF)** ↓ for the full 6-method × 4-decoder matrix over the real-evolution tasks (mean ± std, 5 seeds). AF is peak-minus-final and is non-negative by construction. SI and Distillation achieve near-zero AF but at a large AP cost on rotational decoders (see Table 3); replay variants (ER, MIR) show moderate AF but devastating AP drops for ComplEx.

CL Strategy	TransE	DistMult	ComplEx	RotatE
Naive Sequential	0.036 ± 0.000	0.043 ± 0.001	0.054 ± 0.001	0.065 ± 0.003
EWC	0.030 ± 0.001	0.043 ± 0.001	0.020 ± 0.001	0.066 ± 0.004
SI	≈ 0.000	≈ 0.000	≈ 0.000	0.008 ± 0.004
Distillation	≈ 0.000	≈ 0.000	0.001 ± 0.002	≈ 0.000
Experience Replay	0.037 ± 0.001	0.058 ± 0.002	0.063 ± 0.002	0.064 ± 0.003
MIR	0.039 ± 0.000	0.057 ± 0.001	0.062 ± 0.001	0.061 ± 0.003
Joint (oracle)	—	0.000	0.000	0.000

Table S2: **Backward Transfer (BWT)** ↑ for the full matrix over the real-evolution tasks. Positive values indicate earlier tasks improved after later training; negative values indicate forgetting.

CL Strategy	TransE	DistMult	ComplEx	RotatE
Naive Sequential	-0.036 ± 0.000	-0.007 ± 0.001	-0.054 ± 0.001	-0.048 ± 0.003
EWC	-0.028 ± 0.001	-0.011 ± 0.002	-0.017 ± 0.001	-0.040 ± 0.005
SI	$+0.001 \pm 0.000$	$+0.010 \pm 0.001$	$+0.008 \pm 0.001$	$+0.012 \pm 0.004$
Distillation	$+0.001 \pm 0.000$	$+0.010 \pm 0.001$	$+0.004 \pm 0.002$	$+0.010 \pm 0.000$
Experience Replay	-0.037 ± 0.001	-0.022 ± 0.003	-0.063 ± 0.002	-0.048 ± 0.003
MIR	-0.038 ± 0.000	-0.019 ± 0.001	-0.061 ± 0.001	-0.050 ± 0.003
Joint (oracle)	—	0.000	0.000	0.000

S3 Baseline Details

We evaluate the following ten methods, each with the indicated hyperparameter setting. All KGE-based baselines use embedding dimension $d = 256$, learning rate $\eta = 10^{-3}$, batch size 512, and 100 training epochs per task with the Adam optimizer.

- **Naive Sequential.** Fine-tunes on each task sequentially without any forgetting-mitigation mechanism. Serves as a lower bound on knowledge retention.
- **Joint Training.** Trains on the union of all tasks simultaneously. Serves as a reference oracle since no forgetting can occur by construction.
- **EWC** (Kirkpatrick et al., 2017). Regularization-based approach: penalizes parameter changes weighted by the diagonal Fisher information from previous tasks. Default $\lambda = 10$.
- **Experience Replay** (Rolnick et al., 2019). Memory-based approach: replays a 1,000-triple exemplar buffer selected by K-means diverse sampling.
- **SI (Synaptic Intelligence)** (Zenke et al., 2017). Regularization-based: tracks parameter importance online via accumulated gradients during training, without requiring a separate Fisher pass.
- **Distillation** (Hinton et al., 2015). Knowledge-distillation approach: minimizes divergence between current-task predictions and snapshot model outputs.
- **MIR (Maximally Interfered Retrieval)** (Aljundi et al., 2019). Replay variant: selects buffer exemplars most affected by the current gradient update.
- **LKGE** (Cui et al., 2023). Combined architecture and distillation framework, the state-of-the-art CGL approach for KGs. Constrained to TransE backbone.
- **RAG Agent (Qwen2.5-7B)** (Lewis et al., 2020). Non-parametric retrieval-augmented generation pipeline. Included for completeness; evaluated on the KGQA track.
- **CMKL (R-GCN + DistMult).** Multimodal CGL method with textual (BiomedBERT), molecular (Morgan FP), and structural (R-GCN) features fused via mixture-of-experts (Zhao et al., 2022); uses modality-aware EWC (λ_s=10, λ_t=5, λ_m=1, λ_f=5, λ_r=50) and a 1,000-triple multimodal replay buffer.

CMKL specification. For completeness we give CMKL's algorithm (evaluated here as a baseline). Per entity e, three modality encoders produce embeddings: structural h_e^s (2-layer R-GCN over the t_0 graph), textual h_e^t (frozen BiomedBERT [CLS] projected $768 \to 256$), and molecular h_e^m (Morgan fingerprint through a learned MLP; zero when absent). A mixture-of-experts router $g(\cdot)$ produces per-entity weights $(w_e^s, w_e^t, w_e^m) = \mathrm{softmax}(W_g[h_e^s; h_e^t; h_e^m])$, and the fused embedding is $h_e = w_e^s h_e^s + w_e^t h_e^t + w_e^m h_e^m$, scored with the DistMult decoder. Continual training minimizes the margin-ranking link-prediction loss plus (i) a modality-aware EWC penalty $\sum_{\text{mod}} \lambda_{\text{mod}} \sum_i F_i^{\text{mod}} (\theta_i - \theta_i^*)^2$ with per-modality Fisher F^{mod} and λ_s=10, λ_t=5, λ_m=1, λ_f=5, λ_r=50, and (ii) replay of a 1,000-triple buffer chosen by K-means-diverse selection over structural embeddings. A reference implementation is released with the benchmark.

Per-cell tuning of the ComplEx column. To confirm the EWC×ComplEx effect (§4.2) is not an artifact of hyperparameters shared across CL strategies, we re-tuned the Naive-ComplEx baseline with a per-cell learning-rate sweep. Retuning does not lift Naive-ComplEx (it remains at ≈ 0.014–0.018 real-task AP), whereas EWC-ComplEx reaches 0.046–0.050; the interaction therefore holds at a matched, per-cell-tuned budget.

Methods that do not scale to our base task. FastKGE (Liu et al., 2024b) (`github.com/seukgcode`) and SAGE (Li et al., 2025) (`github.com/lyfxjtu`) have public implementations but assume a task-partition and negative-sampling protocol incompatible with our entity-type, all-entity-ranking setup without substantial reimplementation; ERPP (Yang et al., 2025) had no public code at submission time. We additionally attempted IncDE (Liu et al., 2024a), following the official repository: across five attempts that scaled RAM up to 350 GB, the published implementation did not complete training on our 5.67M-triple base task (out-of-memory failures and training-loop deadlocks). We document our exact configurations alongside the released code; the negative result itself documents the gap between general-domain CKGE benchmarks ($\leq$15K entities) and real biomedical-scale graphs.

S4 Per-Task Learning Matrices

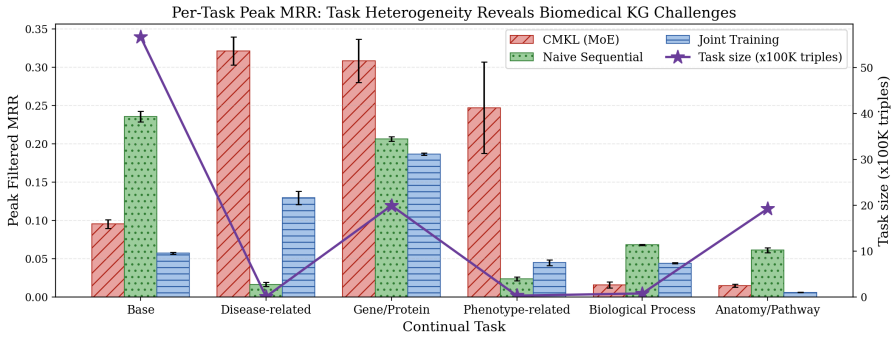

Figure S1: Per-task peak MRR for CMKL, Naive, and Joint; purple line = training-set size. CMKL peaks at 0.25–0.32 on Disease-related, Gene/Protein, Phenotype-related.

We complement the per-task and decoder-CL visualizations (Figures S1 to 3) with full per-task learning matrices in Figure S3. Each cell (i, j) reports filtered MRR on task j after training through task i (mean over 5 seeds, DistMult decoder), so the diagonal is each task's peak performance and below-diagonal columns trace catastrophic forgetting through the rest of the sequence. The matrix makes two patterns visible that the aggregate AP/AF metrics flatten: (i) Disease-related and Gene/Protein dominate the diagonal (peaks ≥ 0.30) but their column also shows the steepest decay, so the highest-performing tasks contribute disproportionately to forgetting; (ii) the bottom rows of the Naive matrix are nearly uniform horizontally, indicating that by the time the model has trained through later tasks, its representation no longer differentiates the easier task-types it once retained. The CMKL matrix shows the opposite pattern: peaks on Disease-related, Gene/Protein, and Phenotype-related are preserved through the sequence, with the dominant decay isolated to Disease-related alone.

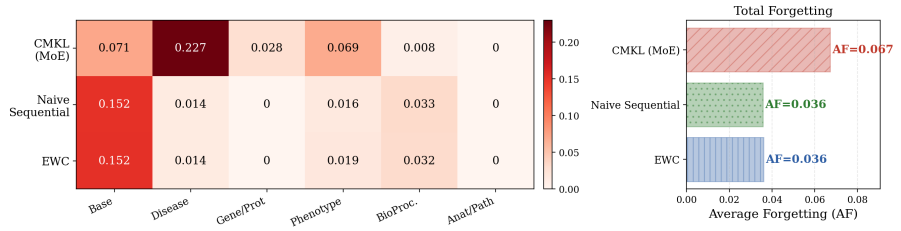

Figure S2: Per-task forgetting (peak − final MRR). CMKL concentrates forgetting on Disease-related (56%, 0.227); Naive/EWC concentrate it on Base (t_0) (70%).

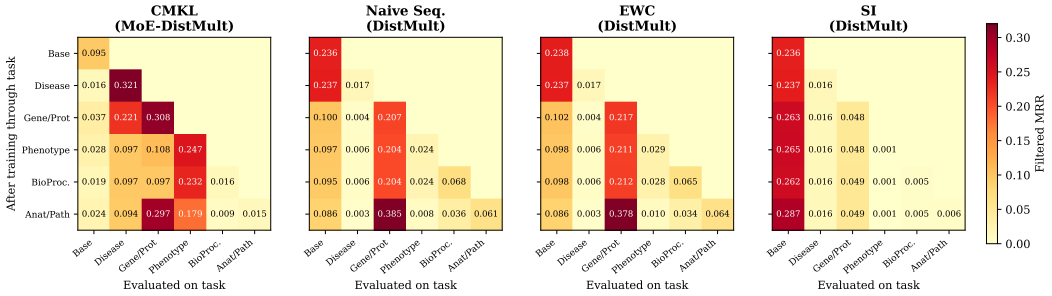

Figure S3: Per-task learning matrices over the six real tasks (DistMult, 5-seed mean; panels: CMKL, Naive, EWC, SI). Entry (i, j) = filtered MRR on task j after training through task i; diagonal = peak per-task performance, off-diagonal decay = forgetting trajectory. CMKL preserves the diagonal more uniformly than the sequential baselines.

S5 KGQA Details

We evaluate continual KGQA using three configurations of Qwen2.5-7B-Instruct to ablate the contribution of retrieval:

- **Full RAG:** LLM with ChromaDB subgraph retrieval (200 questions/task).
- **Retrieval-only:** Extracts answers from retrieved triples via majority voting, without LLM generation.
- **Zero-shot LLM:** LLM answers directly from parametric knowledge, without any retrieval.

The zero-shot LLM achieves token F1 = 0.000 across all tasks, failing to generate correct biomedical entity names from parametric knowledge alone. Retrieval-only achieves token F1 = 0.003 ± 0.001, demonstrating that retrieved triples contain relevant information but simple entity extraction is insufficient. Full RAG achieves the highest performance (token F1 ≤ 0.015), highlighting the challenge of generating exact biomedical entity names from free-text LLM output. All configurations exhibit near-zero forgetting (AF ≤ 0.001), as retrieval-augmented approaches store knowledge in the index rather than model weights.

S6 Analysis Details

Per-task performance variation. Among the five real added tasks, Disease-related, Gene/Protein, and Phenotype-related associations achieve substantially higher peak MRR (> 0.1), driven by larger training sets and denser graph neighborhoods, whereas the Biological-Process and Anatomy/Pathway groups are harder (peak MRR below 0.05 for all methods), reflecting sparser, more fan-out–structured annotation edges. This heterogeneity is a realistic characteristic of biomedical KGs that synthetic benchmarks with uniform random splits do not capture.

Task-order sensitivity. Because entity-type grouping does not impose a unique chronological order, we test sensitivity to the task ordering on Naive-DistMult (real-task AP, mean over seeds): the reported canonical order gives 0.096, a reversed order gives 0.072, and a randomly shuffled order gives 0.027. Ordering therefore has a substantial effect (up to $\approx 3.5\times$), and the canonical order is a favorable case; absolute AP/AF values should be read as order-conditioned rather than order-invariant. This is a property of the domain-incremental setup and motivates reporting the ordering explicitly (which we do) and treating cross-method comparisons at a fixed order.

Gene/protein dominance effect. Gene/protein entities account for approximately 73% of triples in PrimeKG, creating a substantial type imbalance. This affects all methods: performance on gene/protein-centric tasks is consistently higher than on tasks involving rarer entity types.

Entity-type-specific forgetting patterns. For CMKL (MoE), Disease-related accounts for 56% of total forgetting (0.227), followed by Base (t_0) at 18% and Phenotype-related at 17%. In contrast, Naive Sequential and EWC concentrate forgetting on Base (t_0) at 70% (0.152), with substantially less Disease forgetting (0.014). This divergence reveals that the dominant source of forgetting depends on model architecture: multimodal representations shift the vulnerability from the largest task (Base, which dominates structural embeddings) to disease-related tasks where textual and structural signals interact. The Biological-Process and Anatomy/Pathway tasks show near-zero forgetting across all methods, suggesting these entity types have more stable embedding neighborhoods.

Domain-specific insights. The biomedical domain introduces unique CGL challenges absent from generic benchmarks. First, ontology updates (e.g., MONDO disease hierarchy changes) cause systematic triple removals that look like "forgetting" but actually reflect corrected knowledge. Second, highly connected hub nodes (e.g., TP53, BRCA1) participate in many relation types, creating inter-task dependencies that complicate task-isolated continual learning. Third, the multimodal nature of biomedical entities means that forgetting can be modality-specific: a model may retain a drug's molecular properties while forgetting its textual description-based associations.

S7 Limitations and Future Work

- **Two snapshots.** Only t_0 (2021) and t_1 (2023) are currently included; the construction pipeline supports extension to additional snapshots as upstream databases update.
- **Seven databases not re-queried.** DrugBank, UMLS, DrugCentral, SIDER, Reactome, PPI sources, and DisGeNET (restrictive licensing or API changes) were not re-queried for t_1; incorporating them would enrich drug-related dynamics.
- **Gene/protein dominance.** 73% of triples involve gene/protein entities, biasing evaluation toward this type. Type-balanced splits are a future direction.
- **Evaluation paradigm gap.** KGE methods use filtered MRR with all-entity ranking; the RAG agent uses entity-name matching against retrieved subgraphs. Direct cross-paradigm comparison requires caution.

Future work. Promising directions include integrating licensed databases for richer drug-disease dynamics; co-designed decoder-CL pairs (e.g., replay variants for Hermitian bilinear decoders, regularization schemes that preserve rotational geometry); multi-hop reasoning tasks; inductive continual learning where new entity types appear across snapshots; and foundation-model approaches for continual biomedical KG learning building on KG-LLM complementarity (Pan et al., 2024).

S8 Data Availability, License, and Maintenance

License. Code components are released under the MIT license. Data derived from permissively-licensed upstream sources (GO, HPO, HPOA, MONDO, Uberon, Bgee, CTD, Gene2GO, HGNC) are released under Creative Commons Attribution 4.0 International (CC BY 4.0), inheriting PrimeKG's stance for those sources. Content from restrictively-licensed databases (DrugBank, UMLS, DrugCentral, SIDER, DisGeNET) is *not* redistributed: these sources were not re-queried for t_1 (see Limitations), and we instead ship a rebuild script that re-queries them under each user's own

accepted license terms. A per-source license table is provided in `DATASHEET.md`; the full LICENSE file ships in the dataset archive root. **Hosting.** Long-term hosting via HuggingFace Datasets (DOI assigned upon acceptance) and a GitHub release tag at the publication SHA. Croissant 1.0 metadata (`croissant.json`) accompanies the release. **Maintenance plan.** We commit to (i) corrective releases for downstream-database-driven errata for at least 24 months after publication, (ii) a tagged release for each future t_i snapshot we add, (iii) GitHub Issues triage with a 72h initial response target during the year following publication. **Contents.** Temporal snapshots (`kg_t0.csv`, `kg_t1.csv`), task directories with 70/10/20 train/valid/test splits, multimodal features (`text_embeddings.pt`, `mol_features.pt`, R-GCN tensors), persistent/added/removed test stratification (`test_stratification.json`), per-task statistics, and reference baseline + CMKL implementations. **Datasheet for Datasets.** A datasheet following Gebru et al. (2021) is included as `DATASHEET.md` in the release. **Stratified evaluation scope.** The stratified-MRR table (Table 5) reports 6 cells (3 methods $\times$ 2 decoders, 5 seeds each), the configurations our compute budget supported within the submission window. Extending the matrix to the full 6 CL methods $\times$ 4 decoders is one of the first follow-on experiments the released benchmark and stratification metadata enable.