CMKL: MULTIMODAL CONTINUAL LEARNING FOR EVOLVING BIOMEDICAL KNOWLEDGE GRAPHS

Anonymous authors
Paper under double-blind review

ABSTRACT

Biomedical knowledge graphs (KGs) are increasingly large, dynamic, and multimodal, driven by rapid advances in biotechnology such as high-throughput sequencing. Machine learning models can infer previously unobserved biomedical relationships and characterize biomedical entities in these graphs, but existing knowledge graph embedding (KGE) methods and their continual learning (CL) extensions either assume static graph structure or fail to exploit multimodal information under evolving data distributions. They also apply uniform regularization across all model parameters, ignoring that different modalities may exhibit distinct forgetting dynamics as the graph evolves. We propose the Continual Multimodal Knowledge Graph Learner (CMKL), a CL framework for biomedical KGs that natively encodes structure (R-GCN), text (frozen BiomedBERT), and molecules (Morgan fingerprints), fuses them through a Mixture-of-Experts (MoE) router, and protects previously learned knowledge with standard EWC regularization and a K-means-diverse multimodal replay buffer. We evaluate CMKL on a 129K-entity biomedical continual benchmark. On continual biomedical entity classification, CMKL reaches Macro-F1 0.591 and beats *feature-matched* continual baselines (0.414) by $+43\%$ ($p < 10^{-6}$)—an architectural gain isolated from feature availability—preserved across the sequence with near-zero forgetting (AF 0.008). On continual relationship prediction (real $t_0 \to t_1$ evolution tasks), CMKL reaches AP 0.103, on par with Naive Sequential and EWC (0.096) and above Joint Training (0.078); a residual-skip fusion variant reaches 0.117, recovering the structural-only performance (0.118) that embedding fusion otherwise sacrifices—fusion that does no harm. Finally, we correct a natural misreading of these graphs: a text-only ablation's high AP (0.226) is fully reproduced by a *zero-text* transductive node-embedding lookup, so the apparent frozen-text "ceiling" is a node-embedding artifact—frozen semantic modalities carry strong classification but negligible link-prediction signal, which the MoE router reflects by down-weighting text on link prediction. Our code is available at `https://anonymous.4open.science/r/cmkl-multimodal-ckg-FF56`.

1 INTRODUCTION

Biomedical knowledge graphs (KGs) combine relational structure, ontology descriptions, and molecular representations to support drug discovery, disease understanding, and target identification (Chandak et al., 2023; Huang et al., 2024). These signals are complementary: molecular fingerprints constrain binding targets, text captures clinical context, and graph neighborhoods reveal functional associations. As databases evolve, embeddings must incorporate updates without catastrophic forgetting (Kirkpatrick et al., 2017).

Structural continual knowledge graph embedding (CKGE) methods, including EWC (Kirkpatrick et al., 2017), replay (Rolnick et al., 2019), and LKGE (Cui et al., 2023), treat entities as featureless identifiers and regularize parameters uniformly, implicitly assuming equal forgetting rates. Yet text, molecules, and topology undergo different distribution shifts, and no existing CKGE method leverages or protects these modalities.

Continual Multimodal Knowledge Graph Learner (CMKL) addresses this unequal stability: one encoder is frozen, another is trainable with drifting topology, and a third covers few entities. MoE

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.

routing weights modalities per entity, down-weighting those that carry little signal for the task at hand; K-means-diverse multimodal replay revisits topologically distinct exemplars with all modality features. Standard per-encoder EWC is a configuration choice: its λ ablation matches uniform λ on the evaluated decoder (Section 4.5).

A text-only ablation attains the highest raw MRR on link prediction, but this is a transductive node-embedding artifact rather than frozen-text signal: a zero-text node-embedding lookup reproduces it while frozen-text retrieval stays at the floor (Section 4.6). Frozen semantic modalities thus carry strong entity-classification but negligible link-prediction signal, and the router down-weights text on link prediction accordingly. We compare MoE, gated cross-attention, concatenation, and score-level fusion with OGM-GE (Peng et al., 2022): embedding-level fusion (AP 0.100–0.104) outperforms score-level (best 0.081), and we select MoE for inspectable per-entity modality weights.

We evaluate CMKL on a 129K-entity, 8.1M-edge biomedical continual learning benchmark whose tasks derive from real $t_0 \to t_1$ database updates. On continual biomedical entity classification, CMKL reaches Macro-F1 = 0.591, a $+43\%$ gain over *feature-matched* multimodal baselines (0.414; $p < 10^{-6}$), with near-zero forgetting—an architectural gain isolated from feature access. On continual relationship prediction (real-evolution tasks), CMKL reaches AP = 0.103 filtered Mean Reciprocal Rank (MRR) with DistMult, on par with Naive/EWC (0.096) and above Joint Training (0.078); a do-no-harm residual-skip fusion reaches 0.117, recovering structural-only performance. The largest link-prediction headroom lies with rotational decoders, which CMKL's real-valued fusion does not yet exploit (Section 4.6).

Our contributions are: (i) CMKL, combining structural, textual, and molecular encoders with MoE routing and K-means-diverse multimodal replay; (ii) a multimodal *architecture* win on continual biomedical entity classification — $+43\%$ over feature-matched continual baselines ($p < 10^{-6}$), isolating the fusion architecture from mere feature access; (iii) a causal diagnosis correcting a natural misreading of these graphs: the apparent frozen-text "ceiling" is a transductive node-embedding artifact, so frozen semantic modalities carry strong entity-classification but negligible link-prediction signal; and (iv) an honest characterization of continual link prediction on real biomedical evolution, where CMKL matches or exceeds structural CL methods within its decoder without harming them. Per-encoder EWC is a configuration choice, not a methodological contribution.

2 Related Work

Continual Knowledge Graph Embedding. LKGE (Cui et al., 2023) provides a modular framework combining KGE models (TransE, DistMult, RotatE) with continual strategies (EWC, experience replay, distillation); IncDE (Liu et al., 2024) handles incremental KG embedding via incremental distillation; Daruna et al. (Daruna et al., 2021) study continual graph learning for robotics. Concurrent work by Li et al. (Li et al., 2026) proposes MRCKG, combining curriculum learning and contrastive replay for continual multimodal KGE on general-domain KGs ($\leq$15K entities) with synthetic temporal splits. In contrast, CMKL targets real biomedical KG evolution at 129K+ entity scale with domain-specific molecular modalities, a feature-matched multimodal architecture win on continual entity classification, and a causal analysis of how frozen semantic modalities behave under continual multimodal fusion. Prior to MRCKG, all CKGE methods operated exclusively on graph structure and applied uniform regularization, ignoring the multimodal attributes in modern biomedical KGs.

Multimodal Knowledge Graph Embedding. MoMoK (Zhang et al., 2025) introduces modality-aware KG completion using *relation-guided* MoE routing with expert disentanglement via mutual information minimization; our MoE design uses *entity-level* routing, better suited to biomedical KGs where modality availability is determined by entity type (e.g., only drugs have molecular fingerprints). NativE (Zhang et al., 2024a) and AdaMF-MAT (Zhang et al., 2024b) tackle modality imbalance via relation-guided dual adaptive fusion and adaptive modality weights respectively. MCLEA (Lin et al., 2022) employs contrastive learning for multimodal entity alignment. MoSE (Zhao et al., 2022) proposes score-level fusion with per-modality decoders. IMF (Li et al., 2023) uses interaction-based fusion, and OTKGE (Cao et al., 2022) applies optimal transport to align modality-specific embedding spaces. Liang et al. (Liang et al., 2024) survey fusion strategies (early, late, hybrid). All these methods assume a static graph setting and do not address the catastrophic forgetting issue.

Greedy Modality and Multimodal Optimization. The *greedy modality* problem (Wu et al., 2022) arises when one modality converges faster during joint training, suppressing slower modalities. OGM-GE (Peng et al., 2022) addresses this through gradient modulation; CGGM (Guo et al., 2024) extends it by considering gradient magnitude and direction. These gradient-level solutions operate during optimization; our MoE routing instead operates at the *representation level*, down-weighting a modality that is low-signal for a given task without gradient surgery. Mixing a frozen pretrained encoder (BiomedBERT) with jointly trained components makes such asymmetries salient; we further show that the apparent frozen-text dominance in our link-prediction setting is a transductive node-embedding artifact rather than a strong modality being suppressed, a distinction prior greedy-modality analyses did not draw.

Biomedical KG Reasoning and Continual Learning Mechanisms. PrimeKG (Chandak et al., 2023) integrates 20+ biomedical databases; TxGNN (Huang et al., 2024) leverages it for zero-shot drug repurposing. Both treat the KG as static. EWC (Kirkpatrick et al., 2017) prevents forgetting by penalizing changes to important parameters via the Fisher information; Synaptic Intelligence (Zenke et al., 2017) accumulates importance online; replay-based methods (Rolnick et al., 2019; Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019) revisit stored examples. Standard EWC computes a single Fisher matrix over all parameters; we apply it with per-encoder λ values as a configuration choice (without claiming this as a methodological contribution; an ablation against uniform λ is in Section 4.5).

To our knowledge, CMKL is the first combination of three established mechanisms in continual multimodal biomedical KGs: per-modality Fisher regularization, K-means-diverse multimodal replay, and entity-level MoE routing. We find MoE load-bearing for the frozen–learnable asymmetry, replay dominant for forgetting mitigation, and per-encoder EWC matching uniform EWC on DistMult (Section 4.5).

3 METHODOLOGY

CMKL is designed around an asymmetry between the three modalities of biomedical KGs: graph topology drifts whenever a new PrimeKG release adds or rewires edges, a drug's Morgan fingerprint is fixed once computed, and a frozen BiomedBERT description has no gradient at all. Formally, we observe an evolving biomedical knowledge graph as a sequence of snapshots $\{G_1, G_2, \ldots, G_T\}$, where each $G_t = (\mathcal{V}_t, \mathcal{E}_t, \mathcal{R}_t, \mathbf{X}_t)$ consists of entities $\mathcal{V}_t$, typed edges $\mathcal{E}_t \subseteq \mathcal{V}_t \times \mathcal{R}_t \times \mathcal{V}_t$, relation types $\mathcal{R}_t$, and multimodal entity attributes $\mathbf{X}_t$. The benchmark defines a task sequence $\mathcal{T} = \{T_1, \ldots, T_K\}$ via entity-type grouping, where each task T_k provides training triples for a specific entity-type group. The objective is to learn entity embeddings $\mathbf{h}_v \in \mathbb{R}^D$ that score link prediction across all tasks without catastrophic forgetting.

3.1 ARCHITECTURE OVERVIEW

CMKL consists of four stages: (i) three modality-specific encoders that extract structural, textual, and molecular representations; (ii) a Mixture-of-Experts (MoE) fusion module that adaptively combines per-modality representations via a learned router; (iii) a decoder (DistMult) that scores candidate triples using the fused entity embeddings; and (iv) continual learning mechanisms (standard EWC regularization and a multimodal memory replay buffer) that protect previously learned knowledge as the model trains on new tasks.

3.2 MODALITY-SPECIFIC ENCODERS

Structural ($\mathbf{h}_v^s$). A 2-layer R-GCN (Schlichtkrull et al., 2018) with basis decomposition,

$$\mathbf{h}_v^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{u\in\mathcal{N}_v^r}\tfrac{1}{|\mathcal{N}_v^r|}\mathbf{W}_r^{(l)}\mathbf{h}_u^{(l)} + \mathbf{W}_0^{(l)}\mathbf{h}_v^{(l)}\right), \tag{1}$$

with $\mathbf{W}_r^{(l)} = \sum_{b=1}^{B} a_{rb}^{(l)}\mathbf{V}_b^{(l)}$ and ReLU σ, producing $\mathbf{h}_v^s \in \mathbb{R}^D$.

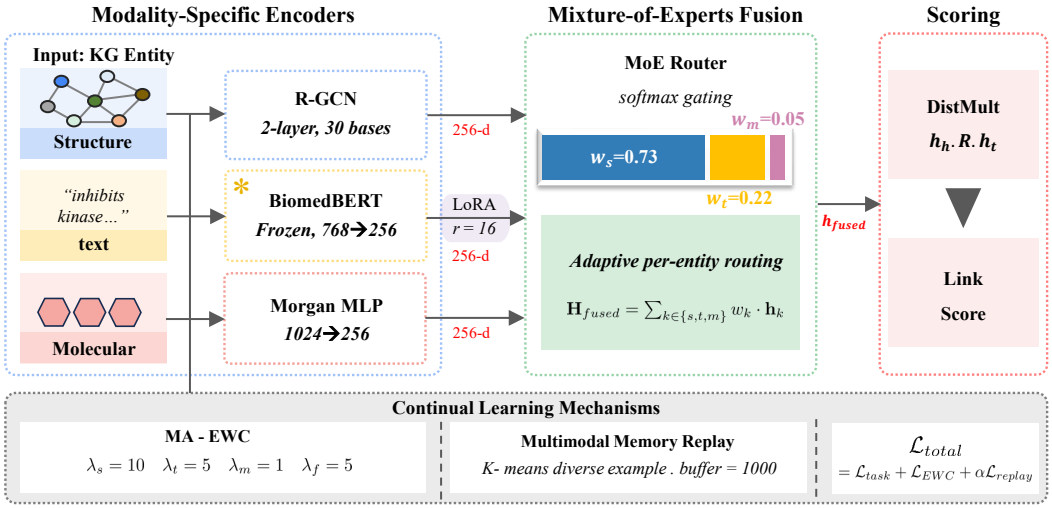

Figure 1: Architecture of CMKL. Three modality-specific encoders produce structural ($\mathbf{h}^s$) via a 2-layer R-GCN, textual ($\mathbf{h}^t$) via frozen BiomedBERT with an optional LoRA adapter, and molecular ($\mathbf{h}^m$) via a Morgan-fingerprint MLP. A Mixture-of-Experts router adaptively combines them per entity via learned softmax weights (w^s, w^t, w^m). The fused embedding $\mathbf{h}_{\text{fused}}$ is scored by a DistMult decoder. Continual-learning mechanisms (bottom): modality-aware EWC applies per-modality Fisher regularization ($\lambda_s, \lambda_t, \lambda_m, \lambda_f$) and K-means multimodal memory replay selects diverse exemplars from past tasks into a size-1000 buffer. The total training loss combines task, EWC, and replay terms.

Textual ($\mathbf{h}_v^t$). Frozen BiomedBERT (Gu et al., 2021) mean-pooled and linearly projected, $\mathbf{h}_v^t = \mathbf{W}_{\text{proj}} \cdot \text{MeanPool}(\text{BiomedBERT}(x_v^{\text{text}})) + \mathbf{b}_{\text{proj}}$ with $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{D\times 768}$. Only the projection trains; entities without text get a learned default.

Molecular ($\mathbf{h}_v^m$). For drug nodes, 1024-bit radius-2 Morgan fingerprints (Rogers & Hahn, 2010) are passed through a 2-layer MLP $\mathbf{h}_v^m = \mathbf{W}_2\,\sigma(\mathbf{W}_1 x_v^{\text{mol}} + \mathbf{b}_1) + \mathbf{b}_2$ with $\mathbf{W}_1 \in \mathbb{R}^{D\times 1024}$, $\mathbf{W}_2 \in \mathbb{R}^{D\times D}$. Non-drug entities get a learned default.

3.3 Multimodal Fusion

The fusion module must satisfy three constraints imposed by the encoder asymmetry from Section 1: it must (i) be able to place all weight on a single modality when the data justify it, (ii) not push frozen-encoder outputs through a learned bottleneck that destroys their pretrained geometry, and (iii) allow post-hoc inspection of which modality each entity actually used. Among the embedding-level and score-level alternatives we evaluate (Section 4.4), Mixture-of-Experts (MoE) routing is the only one that satisfies all three (constraint (i) is a representational capacity of the router, which soft routing does not always realize in practice, as we discuss below).

Mixture-of-Experts Fusion (selected). We treat each modality encoder as an expert and learn a per-entity router that produces modality weights:

$$\mathbf{w}_v = \text{softmax}\left(\mathbf{W}_r[\mathbf{h}_v^s; \mathbf{h}_v^t; \mathbf{h}_v^m] + \mathbf{b}_r\right) \in \mathbb{R}^3, \tag{2}$$

where $\mathbf{W}_r \in \mathbb{R}^{3\times 3D}$ is the router weight matrix. The fused representation is a weighted sum of the expert outputs:

$$\mathbf{h}_v = w_v^s \cdot \mathbf{h}_v^s + w_v^t \cdot \mathbf{h}_v^t + w_v^m \cdot \mathbf{h}_v^m. \tag{3}$$

In principle the softmax router can concentrate weight on a single expert, approaching the best single modality. In practice, however, the embedding-level MoE lands slightly *below* the structural-only ablation on link prediction (Table 1), so the "at least as good as the best modality" property is not guaranteed by soft routing alone. We therefore also evaluate a *do-no-harm residual-skip* fusion: $\mathbf{h}_v = \mathbf{h}_v^s + g_v \odot \text{MLP}([\mathbf{h}_v^t; \mathbf{h}_v^m])$ with a zero-initialized gate g_v (bias -8) and a detached structural term, so training *starts* exactly at the structural encoder and can only add a modality contribution

where it helps. This variant recovers the structural-only performance the MoE otherwise sacrifices while retaining the multimodal pathway (Table 1), delivering the do-no-harm property by construction rather than relying on the router to learn it (demonstrated for the DistMult decoder; extending it to rotational decoders is discussed in §4.6).

Alternative fusion variants. We compare MoE with gated cross-modal attention (4-head pairwise attention, MLP, and residual fuse), projected concatenation, and score-level fusion with per-modality decoders (Zhao et al., 2022). Results and equations appear in Table 3 and Section S4.

3.4 EWC Regularization (Per-Modality Configuration)

We apply standard EWC (Kirkpatrick et al., 2017) with per-encoder strengths λ_k as a configuration choice, not a methodological contribution. Partitioning $\theta = \{\theta_s, \theta_t, \theta_m, \theta_f\}$ (with θ_f the fusion/router parameters) gives

$$\mathcal{L}_{\text{EWC}} = \sum_{k \in \{s,t,m,f\}} \lambda_k \sum_i F_i^k \left(\theta_{k,i} - \theta_{k,i}^*\right)^2, \tag{4}$$

Validation selects λ_s=10, λ_t=5, λ_m=1, λ_f=5. Diagonal Fisher matrices F_i^k use squared post-task gradients. Per-encoder and uniform λ match on DistMult (Section 4.5); the former remains a modality-specific tuning option.

3.5 Multimodal Memory Replay

CMKL maintains a fixed-size buffer $\mathcal{M}$ of $|\mathcal{M}|$ representative triples balanced across all completed tasks. After each task T_k, we re-cluster: K-means is run on the structural embeddings of head entities from all triples currently in $\mathcal{M} \cup T_k$ with $K = |\mathcal{M}|/k$ clusters *per task*, and the closest-to-centroid triple from each cluster is retained. This rebalancing strategy preserves diverse exemplars from *every* prior task (at least $|\mathcal{M}|/k$ per task after task k) rather than purging early tasks via a strict FIFO queue. We cluster on structural embeddings because they cover 100% of entities, whereas text and molecular features cover only 36% and 6%, respectively; concatenating partial-coverage features would produce heterogeneous, high-dimensional cluster inputs. Selected triples are *replayed with all available modalities* during subsequent tasks, which is what makes the buffer multimodal. Replay batches $\mathcal{B}_r$ are sampled uniformly from $\mathcal{M}$ and mixed with the current task, with loss

$$\mathcal{L}_{\text{replay}} = \frac{1}{|\mathcal{B}_r|} \sum_{(h,r,t) \in \mathcal{B}_r} \ell(f_r(\mathbf{h}_h, \mathbf{h}_t), y). \tag{5}$$

3.6 Training Procedure

For each task T_k, all entities are encoded by the three modality-specific encoders, fused by the MoE router, and scored by DistMult. The total objective combines the task loss with EWC regularization and replay loss, the latter two active only for $k > 1$:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{EWC}} + \alpha \cdot \mathcal{L}_{\text{replay}}, \quad (\alpha\text{=1.0 replay weight}). \tag{6}$$

After each task, we compute per-modality Fisher matrices and add K-means-selected diverse exemplars to $\mathcal{M}$. Full pseudocode appears in Section S5.

4 Experiments

4.1 Experimental Setup

Benchmark. We evaluate on a continual learning benchmark derived from PrimeKG (Chandak et al., 2023) with two temporal snapshots (t_0: June 2021; t_1: July 2023) reconstructed from nine biomedical databases. The benchmark contains 129K+ entities across 10 node types and 8.1M+ edges spanning 30 relation types, organized into continual learning tasks via entity-type grouping; we report on the six tasks that receive real $t_0 \to t_1$ additions (a base task plus five re-queried entity-type groups). Within each task, triples are split into train, validation, and test at a 70/10/20 ratio. We follow standard CL protocol (Lopez-Paz & Ranzato, 2017; Chaudhry et al., 2019) and report Average

Table 1: CMKL on continual biomedical relationship prediction, reported on the six real $t_0 \to t_1$ evolution tasks (5 seeds, DistMult; do-no-harm row: 3 seeds). AP: filtered MRR over 129K+ entities; baselines share the DistMult decoder for fair comparison. The text-only row is a transductive node-embedding artifact (see §4.6), not a text signal.

Variant	AP ↑	AF ↓	BWT ↑	REM ↑
CMKL (MoE fusion)	0.103 ± 0.016	0.081 ± 0.011	-0.077 ± 0.011	0.919 ± 0.011
CMKL (do-no-harm fusion)	0.117 ± 0.011	0.084 ± 0.002	-0.081 ± 0.004	0.916 ± 0.002
w/o text & mol (struct only)	0.118 ± 0.008	0.083 ± 0.009	-0.076 ± 0.009	0.917 ± 0.009
w/o struct & mol (text only)[†]	0.226 ± 0.004	0.019 ± 0.002	-0.011 ± 0.002	0.981 ± 0.002
Naive Sequential	0.096 ± 0.002	0.043 ± 0.001	-0.007 ± 0.001	0.957 ± 0.001
EWC (Kirkpatrick et al., 2017)	0.096 ± 0.001	0.043 ± 0.001	-0.011 ± 0.002	0.957 ± 0.001
Joint Training	0.078 ± 0.002	—	—	—

[†]The text-only "ceiling" is reproduced by a zero-text transductive node-embedding lookup (§4.6); it reflects that lookup, not frozen-text signal. Baselines share the DistMult decoder for a fair same-decoder comparison; across decoders the strongest baseline is EWC+RotatE (AP 0.146), which exceeds all DistMult-based methods including CMKL— CMKL is scoped to bilinear decoders (§4.6). LKGE (Cui et al., 2023) (TransE) runs under a different 9-task protocol (no base task) and reaches AP 0.039 ± 0.001, not directly comparable to the aligned rows.

Performance (AP, the mean filtered MRR across tasks after the final training step), Average Forgetting (AF), Backward Transfer (BWT), and Remembering (REM). Formal definitions are released with the benchmark.

Baselines and metrics. We compare against Naive Sequential, Joint Training, EWC (Kirkpatrick et al., 2017), and LKGE (Cui et al., 2023). All baselines share the DistMult decoder (dim 256) with CMKL for a fair comparison, except LKGE, which requires TransE. AP for relationship prediction uses filtered MRR over 129K+ candidate entities; AP for entity classification uses Macro-F1.

Implementation details. CMKL uses embedding dimension D=256, a 2-layer R-GCN with B=30 basis matrices, frozen BiomedBERT (microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract) with a 768→256 linear projection, a Morgan-fingerprint MLP (1024→256→256), MoE fusion with a 3-way softmax router, and a DistMult decoder (Yang et al., 2015). Training uses Adam at learning rate 10^{-3} for 100 epochs per task, batch size 512, and a memory buffer of size 1000. EWC strengths are λ_s=10, λ_t=5, λ_m=1, λ_f=5, with replay weight α=1.0. All experiments use 5 random seeds $\{42, 123, 456, 789, 1024\}$ on NVIDIA V100 GPUs (32 GB) and report mean ± std.

4.2 Biomedical Relationship Prediction (Link Prediction) Results

Table 1 reports CMKL's performance on continual biomedical relationship prediction (real-evolution tasks), alongside modality ablations. CMKL (MoE fusion) reaches AP = 0.103 ± 0.016, above Joint Training (0.078, 1.32×) and on par with Naive Sequential and EWC (0.096): the nominal 7% gap is not significant (p=0.49 / 0.41; 95% bootstrap CIs, 10K resamples, $[0.089, 0.117]$ for CMKL vs. $[0.095, 0.098]$ / $[0.095, 0.097]$, overlapping). The embedding MoE fusion sits slightly *below* the structural-only ablation (0.118), i.e., naive embedding fusion trades a little raw LP for the stable multimodal representations that drive the entity-classification gain (§4.3). Our *do-no-harm* residual-skip fusion closes this gap: it reaches AP = 0.117 ± 0.011 ($[0.102, 0.127]$), matching structural-only within noise while retaining the multimodal pathway—fusion that does not sacrifice link prediction. We therefore read CMKL's LP contribution as matching or recovering the strongest structural CL methods within its decoder, with markedly different per-task and per-stratum behavior (Figure 2, Figure S1); the materially significant multimodal gain appears on entity classification (+43% over feature-matched baselines, §4.3).

The frozen-text "ceiling" is a node-embedding artifact, not a text signal. The text-only ablation reaches the highest AP in Table 1 (0.226), which might suggest frozen BiomedBERT carries strong link-prediction signal. A controlled causal decomposition shows it does not. (i) A learnable transductive node-embedding table with *zero* text input reproduces the text-only ablation's performance, so the

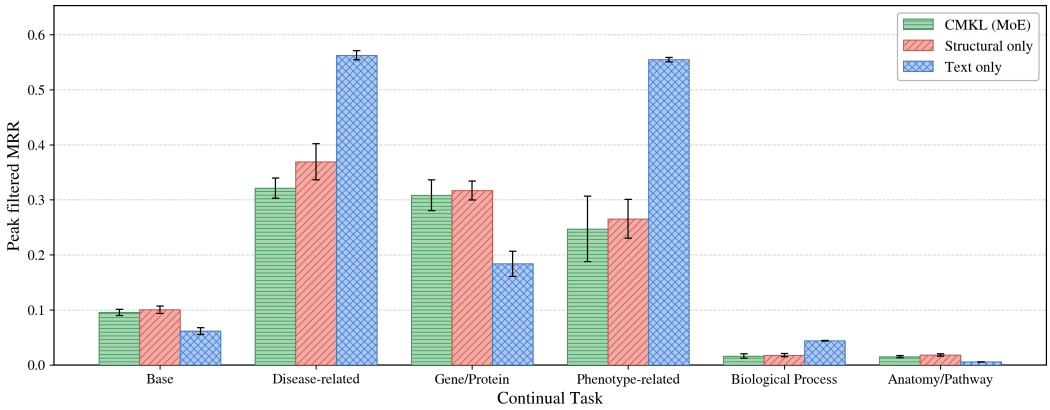

Figure 2: Per-task peak MRR for CMKL vs. single-modality ablations. The text-only ablation's high per-task MRR is a transductive node-embedding artifact (§4.6), not frozen-text signal; structural features drive the trained model.

Table 2: CMKL on continual biomedical entity classification (5 seeds). AP: mean Macro-F1. The *feature-matched* controls give Naive/EWC the same multimodal features as CMKL, isolating the fusion architecture from feature availability; the frozen-text-only row rules out label leakage.

Variant	**AP** ↑	**AF** ↓	**BWT** ↑
CMKL (MoE, all modalities)	$\mathbf{0.591 \pm 0.006}$	$\mathbf{0.008 \pm 0.009}$	$\mathbf{+0.003 \pm 0.007}$
Feature-matched controls (same multimodal features as CMKL)			
Naive Sequential (multimodal)	0.414 ± 0.001	—	—
EWC (multimodal)	0.414 ± 0.003	—	—
Structural-only / leakage controls			
Joint Training (structural)	0.370 ± 0.002	—	—
CMKL w/o text & mol (struct only)	0.345 ± 0.004	0.011 ± 0.002	$+0.003 \pm 0.005$
Frozen-text only	0.246 ± 0.002	—	—

CMKL's $+0.177$ Macro-F1 over the feature-matched CL baselines ($+43\%$; paired $t = 77$, $p = 1.7 \times 10^{-7}$, 5 seeds) is an architectural gain, not merely feature access.

value measures what a plain transductive lookup achieves, not frozen-text semantics. (ii) Frozen-text nearest-neighbor retrieval scores at the floor, and a trainable text projection or LoRA never leaves the floor, so frozen text carries essentially no rankable link-prediction signal on its own. (iii) Adding frozen text to the node-embedding lookup *lowers* its AP rather than raising it. We therefore do not claim a frozen-encoder ceiling "unreachable by gradients"; the honest finding is that frozen semantic modalities carry strong *entity-classification* signal (§4.3) but little *link-prediction* signal, and the MoE router accordingly down-weights text on LP because it contributes little there.

4.3 Biomedical Entity Classification (Node Classification) Results

Table 2 reports results on continual biomedical entity classification, where the model must assign the correct biomedical type (gene/protein, disease, drug, etc.) to each entity as the graph evolves. CMKL reaches AP = 0.591 ± 0.006 with near-zero forgetting (AF = 0.008) and positive backward transfer (BWT = $+0.003$). Crucially, it beats the *feature-matched* continual baselines — Naive and EWC given the *same* multimodal features (0.414) — by $+0.177$ Macro-F1 ($+43\%$; paired $t = 77$, $p = 1.7 \times 10^{-7}$), so the gain is an architectural effect, not merely access to features.

We attribute this gap carefully. The feature-availability control that earlier work deferred is now run: giving Naive and EWC the same textual and molecular features raises them only to 0.414 (Table 2), well below CMKL's 0.591. The $+0.177$ margin is therefore contributed by CMKL's fusion architecture, not by mere feature access—a simpler multimodal classifier with identical features

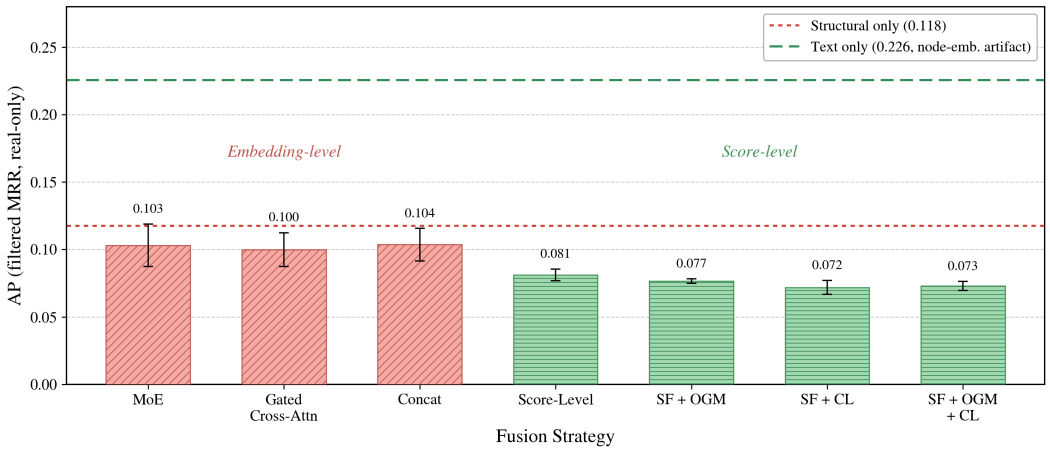

Figure 3: Fusion-strategy comparison. Embedding-level (MoE, gated, concat) outperforms score-level. Dashed lines: single-modality references.

Table 3: Fusion strategy comparison on continual biomedical relationship prediction (5 seeds, DistMult+EWC). MoE, gated cross-attention, and concatenation operate at the embedding level; score-level (SF) variants use per-modality decoders, optionally combined with OGM-GE gradient modulation (Peng et al., 2022) or contrastive alignment.

Fusion Strategy	**AP** $\uparrow$	**AF** $\downarrow$
MoE (selected)	0.103 ± 0.016	0.081 ± 0.011
Gated Cross-Attention	0.100 ± 0.013	0.074 ± 0.010
Concatenation	$\mathbf{0.104 \pm 0.012}$	$\mathbf{0.062 \pm 0.009}$
Score-Level (SF-only)	0.081 ± 0.004	0.071 ± 0.004
SF + OGM-GE	0.077 ± 0.002	0.068 ± 0.005
SF + Contrastive	0.072 ± 0.005	0.070 ± 0.007
SF + OGM + Contrastive	0.073 ± 0.003	0.072 ± 0.004

does not close it. A frozen-text-only classifier reaches 0.246, ruling out label leakage from the text descriptions. CMKL both extracts the multimodal advantage *and* preserves it across the continual sequence (AF near zero, BWT positive); its structural-only ablation (0.345) confirms that structure alone is insufficient.

4.4 Fusion Strategy Ablation

Figure 3 and Table 3 compare seven fusion strategies (three embedding-level and four score-level) on the same DistMult + EWC backbone. Among the embedding-level strategies, concatenation and MoE reach comparable AP (0.104 and 0.103), with concatenation achieving the lowest forgetting (AF = 0.062). We select MoE as the primary fusion because its learned router provides *inspectable* per-entity modality weights; concatenation hides the per-entity modality preference inside its $3D \to D$ projection. All score-level variants underperform the embedding-level approaches, with the best score-level result (SF-only, AP = 0.081) falling 21% below MoE.

Why score-level fusion underperforms. Score-level fusion (Zhao et al., 2022) independently trains per-modality decoders. The structural decoder learns ranking patterns; the frozen text decoder carries little link-prediction signal (§4.6). OGM-GE offers no benefit (AP 0.077): modulation weights remain ≈ 1, and both decoders converge similarly because text contributes little on this task. Lower text-decoder weight helps (α=0.3: 0.097 versus α=2.0: 0.057; single seed). Embedding-level MoE simply down-weights the low-signal text modality per entity, which score-level reweighting of a frozen text decoder cannot do as cleanly.

4.5 Continual Learning Component Ablation

An ablation of EWC configuration and replay-buffer size (Section S8) shows that per-encoder and uniform EWC match within seed noise and that removing EWC preserves AP but raises forgetting: replay is the primary forgetting-mitigation mechanism, with EWC a secondary configuration choice. Performance is robust across a $50\times$ λ range and buffers of 500–5,000 (only $\approx 0.06\%$ of 8.1M edges), requiring little tuning.

4.6 Analysis

AF must be read against per-task peak MRR. CMKL's AF $= 0.081$ exceeds EWC's 0.043, but the comparison is misleading without context. Naive Sequential and EWC peak lower on most tasks, leaving less signal to lose, while CMKL peaks at 0.32 on Disease-related, so its drift is measured against a much taller curve. CMKL's Remembering score $= 0.919$ retains 91.9% of acquired knowledge in absolute terms, and Naive/EWC's lower AF reflects flatter learning rather than stronger retention; an aggressive regularization strategy could trivially achieve AF ≈ 0 by capping AP low. CMKL sits on the efficient corner of the AP–AF frontier.

Modality-specific forgetting and per-modality routing. Figure S1 shows per-task forgetting by modality. The structural-only ablation forgets sharply on Disease-related (0.29) and Base (t_0) (0.08); the text-only ablation is uniformly stable (≤ 0.05, frozen by construction); the full CMKL model shifts the dominant source of forgetting to Disease-related (0.23). The MoE router assigns $w^s \approx 0.70$–1.00 and down-weights text ($w^t \approx 0.00$–0.25) on link prediction, consistent with the finding above that frozen text carries little link-prediction signal (whereas it carries strong entity-classification signal, where the router weights it heavily). The router thus reflects each modality's task-specific usefulness at the *representation level*, rather than being a fusion-stage fix; a frozen-encoder modality that is low-signal for a task is down-weighted per entity without a bottleneck.

Scope: bilinear-decoder family. CMKL's real-valued additive MoE composes with bilinear DistMult scoring, but not rotational or Hermitian geometry. Across four design iterations, CMKL-RotatE plateaus at AP 0.042 versus EWC-RotatE 0.146, and CMKL-ComplEx at 0.034 versus EWC-ComplEx 0.048 (real-only). The failure is geometric rather than statistical: linear fusion does not respect RotatE's rotational metric. The gap appears at task-0 training (MRR ≈ 0.12 versus baseline ≈ 0.26), not through forgetting. Decoder-aware fusion, such as complex-valued routing or near-identity initialization in the rotational metric, would extend the recipe to these families.

5 Conclusion

CMKL fuses R-GCN, frozen BiomedBERT, and Morgan-fingerprint encoders through MoE routing with EWC and multimodal replay. Its central result is a multimodal *architecture* win on continual entity classification: Macro-F1 0.591 beats feature-matched continual baselines (0.414) by $+43\%$ ($p < 10^{-6}$), isolating architecture from feature access. On link prediction it matches the strongest same-decoder baselines (0.103 vs. 0.096), and a do-no-harm residual-skip fusion (0.117) recovers the structural-only performance that embedding fusion otherwise sacrifices. Finally, the apparent frozen-text "ceiling" (text-only AP 0.226) is a transductive node-embedding artifact—reproduced with zero text input—so frozen semantic modalities carry strong classification but negligible link-prediction signal.

Limitations and future work. CMKL's decisive win is on entity classification; on link prediction its contribution is competitiveness and do-no-harm within its decoder (the MoE gap over Naive/EWC is not significant, p=$0.49, 0.41$), not a new state of the art. It is scoped to bilinear (DistMult) decoders—a real-valued additive MoE does not respect rotational/Hermitian geometry and our extensions plateaued below the RotatE/ComplEx baselines (Section 4.6), so decoder-aware fusion is the natural follow-up—and is evaluated on a single benchmark over one $t_0 \to t_1$ transition; cross-benchmark and multi-snapshot validation are important extensions.

AI USE STATEMENT

We used generative AI tools, including Claude Code, for assistance with literature review, research-idea suggestions used by the authors as starting points, implementation of routine experiments, and interpretation of their results. More complex experimental implementation and synthesis across experiments involved direct human input. Experiments used existing biomedical knowledge-graph data; generative AI was not used to synthesize the benchmark data. AI also assisted with manuscript drafting and figure generation using 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

CMKL is trained and evaluated only on publicly available biomedical knowledge-graph data derived from PrimeKG; it uses no private, patient-identifiable, or human-subjects data. Model outputs (predicted drug–disease, gene–phenotype, and related associations) are research artifacts and must not be used for clinical or therapeutic decisions without expert validation. We respect the licenses of the underlying data sources and do not redistribute restrictively-licensed content. The frozen-text encoder is a general-purpose biomedical language model used only for entity-feature extraction, and we make no claims about its safety for clinical deployment.

REPRODUCIBILITY STATEMENT

An anonymized code repository is available for review (linked in the abstract). All reported results are mean $\pm$ standard deviation over five seeds $\{42, 123, 456, 789, 1024\}$ on a single NVIDIA V100 (32 GB); the model architecture, fusion variants, and all hyperparameters are specified in Section 4.1 and Section S1. Link prediction uses filtered MRR over the 129K+ entity set on the six real evolution tasks. The released code and configuration files, together with the benchmark, allow the reported experiments to be reproduced.

REFERENCES

Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. OTKGE: Multi-modal knowledge graph embeddings via optimal transport. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

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

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with A-GEM. In *International Conference on Learning Representations*, 2019.

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.

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.

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.

Zirun Guo, Tao Jin, Jingyuan Chen, and Zhou Zhao. Classifier-guided gradient modulation for enhanced multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30:3601–3613, 2024.

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.

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.

Xinhang Li, Xiangyu Zhao, Jiaxing Xu, Yong Zhang, and Chunxiao Xing. IMF: Interactive multimodal fusion model for link prediction. In *Proceedings of the ACM Web Conference 2023*, pp. 2572–2580, 2023.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42, 2024.

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, 2024.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

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 Lillicrap, and Gregory Wayne. Experience replay for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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

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

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

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

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. NativE: Multi-modal knowledge graph completion in the wild. In *Proceedings of the 47th International ACM SIGIR Conference*, pp. 91–101, 2024a.

Yichi Zhang, Zhuo Chen, Lei Liang, Huajun Chen, and Wen Zhang. Unleashing the power of imbalanced modality information for multi-modal knowledge graph completion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pp. 17120–17130, 2024b.

Yichi Zhang, Zhuo Chen, Lingbing Guo, Yajing Xu, Binbin Hu, Ziqi Liu, Wen Zhang, and Huajun Chen. Multiple heads are better than one: Mixture of modality knowledge experts for entity representation learning. In *International Conference on Learning Representations*, 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.

CMKL: Multimodal Continual Learning for Evolving Biomedical Knowledge Graphs

Supplementary Material

S1 Experimental Details

All experiments use NVIDIA V100 GPUs (32 GB) with PyTorch 2.0 and PyTorch Geometric. Training uses Adam optimizer with learning rate 0.001, batch size 512, and 50,000 sampled triples per epoch for 100 epochs per task. Embedding dimension: 256. R-GCN: 2 layers, 30 basis matrices. MoE router: hidden dimension 128, load-balance weight 0.01. Per-encoder EWC lambdas: $\lambda_s = 10$, $\lambda_t = 5$, $\lambda_m = 1$, $\lambda_f = 5$. Replay buffer: 1,000 total triples, K-means selection on structural embeddings. Score-level fusion: $\alpha_{\text{text}} = 0.5$, $\alpha_{\text{mol}} = 0.3$. Contrastive alignment: temperature 0.1, weight 0.1. OGM-GE: $\alpha = 1.0$. All results averaged over 5 seeds $\{42, 123, 456, 789, 1024\}$.

S2 Additional Results

Alpha sensitivity. The score-fusion text weight (α_{text}) sweep (single seed, seed=42; real-only AP) shows lower text weight is consistently better: $\alpha = 0.3$ gives 0.097 (best) versus $\alpha = 2.0$ gives 0.057 (worst), confirming that structural features dominate biomedical relationship prediction and that the frozen text decoder adds little there.

Forward Transfer. FWT equals zero for all methods in our sequential setting, as tasks are non-overlapping and models are not evaluated on future tasks before training.

S3 Limitations and Future Work

- **The text-only ablation is a node-embedding artifact.** The text-only ablation's high AP is *not* frozen-text signal: a zero-text transductive node-embedding lookup reproduces it, frozen-text retrieval scores at the floor, and a trainable text projection/LoRA never leaves the floor (§4.6). We therefore do not treat it as a recoverable ceiling; frozen text contributes strongly to entity classification but little to link prediction.
- **MoE router collapse.** The router occasionally collapses to a single expert (structural weight w^s=1.0 on one seed), suggesting training instability that load-balancing losses could mitigate.
- **Per-encoder EWC vs uniform EWC.** On DistMult, per-encoder EWC shows no advantage over uniform EWC; per-modality variants may matter more on decoders where EWC's effect is larger.
- **Bilinear-decoder scope.** CMKL is designed around DistMult. Extending the multimodal fusion recipe to complex-valued decoders (ComplEx, RotatE) requires a fusion architecture that respects their geometric constraints (our residual-skip do-no-harm variant preserves the floor on DistMult but does not yet recover rotational geometry).
- **Modality coverage.** Text covers 36% of entities and molecular features 6%, bounding the potential contribution of those modalities.
- **Other future directions** include extending evaluation to additional temporal snapshots, modality-specific learning-rate schedules, and direct per-encoder Fisher analysis.

S4 Fusion-Strategy Equations

The two embedding-level variants summarized in the main paper are:

Gated cross-modal attention. For each entity v we form a 3-token modality sequence $\mathbf{S}_v = [\mathbf{h}_v^s; \mathbf{h}_v^t; \mathbf{h}_v^m] \in \mathbb{R}^{3\times D}$ and apply H=4-head pairwise multi-head attention across modality pairs (i, j):

$$\text{CrossAttn}(\mathbf{S}_v^i, \mathbf{S}_v^j) = \text{softmax}\big(\mathbf{Q}_i \mathbf{K}_j^\top / \sqrt{D/H}\big)\mathbf{V}_j,$$

where $\mathbf{Q}_i = \mathbf{S}_v^i \mathbf{W}_Q^i$ and $\mathbf{K}_j, \mathbf{V}_j$ are computed analogously over the j-th modality token; the softmax is therefore over the per-head key dimension, not over a single scalar. Stacking the four cross-attended outputs together with the structural anchor $\mathbf{h}_v^s$ gives $\mathbf{c}_v = [\mathbf{h}_v^s; \hat{\mathbf{h}}_v^{s\leftarrow t}; \hat{\mathbf{h}}_v^{s\leftarrow m}; \hat{\mathbf{h}}_v^{t\leftarrow s}; \hat{\mathbf{h}}_v^{m\leftarrow s}] \in \mathbb{R}^{5D}$ and a final MLP-with-residual fuse $\mathbf{h}_v = \text{LayerNorm}(\text{MLP}_{\text{fuse}}(\mathbf{c}_v) + \mathbf{h}_v^s)$. Structural anchoring (using $\mathbf{h}_v^s$ as the residual base) reflects that structural features cover all entities while text and molecular features cover only 36% and 6%.

Concatenation. $\mathbf{h}_v = \mathbf{W}_c[\mathbf{h}_v^s; \mathbf{h}_v^t; \mathbf{h}_v^m] + \mathbf{b}_c$, $\mathbf{W}_c \in \mathbb{R}^{D\times 3D}$.

Why score-level fusion underperforms. Score-level fusion (Zhao et al., 2022) trains per-modality decoders to combine at evaluation time. The frozen text decoder carries negligible link-prediction signal (§4.6), so a separate text decoder adds little and the structural decoder dominates. OGM-GE gradient modulation (Peng et al., 2022) provided no benefit (modulation weights stayed ≈ 1); an α sweep over text-decoder weight confirmed less text is better ($\alpha = 0.3$: AP = 0.097 vs. $\alpha = 2.0$: 0.057; real-only, single seed).

S5 Training Algorithm

Algorithm 1 CMKL Training Procedure

Require: Task sequence $\mathcal{T} = \{T_1, \ldots, T_K\}$, modality features $\mathbf{X}$, buffer size $|\mathcal{M}|$
Ensure: Trained parameters $\theta = \{\theta_s, \theta_t, \theta_m, \theta_f\}$

```
Initialize encoders, MoE router, decoder; M ← ∅
for k = 1 to K do
    for each epoch do
        H^s ← R-GCN(V_k, E_k; θ_s); H^t ← BiomedBERT-Proj(X^text; θ_t); H^m ← MLP(X^mol; θ_m)
        w ← softmax(W_r[H^s; H^t; H^m] + b_r); H ← w^s H^s + w^t H^t + w^m H^m
        L_task ← link-prediction loss on T_k via decoder(H)
        if k > 1 then
            L_EWC ← Equation (4); L_replay ← replay loss on B_r ⊂ M
        else
            L_EWC ← 0; L_replay ← 0
        end if
        L ← L_task + L_EWC + α L_replay; Adam update on θ
    end for
    Compute F^s, F^t, F^m, F^f; θ* ← θ; M ← K-means-Rebalance(M ∪ T_k, H^s, |M|)
end for
```

S6 Contextualizing Absolute MRR Values

Filtered MRR over 129K+ candidate entities is inherently challenging. KG-FIT (Jiang et al., 2024) reports TransE MRR = 0.048 on a 10K-entity PrimeKG subset; scaling to 129K entities makes ranking proportionally harder. The meaningful comparisons in Table 1 are therefore relative differences between methods under the same decoder: CMKL outperforms DistMult Joint Training by $1.3\times$ and LKGE (TransE-constrained) by $1.6\times$.

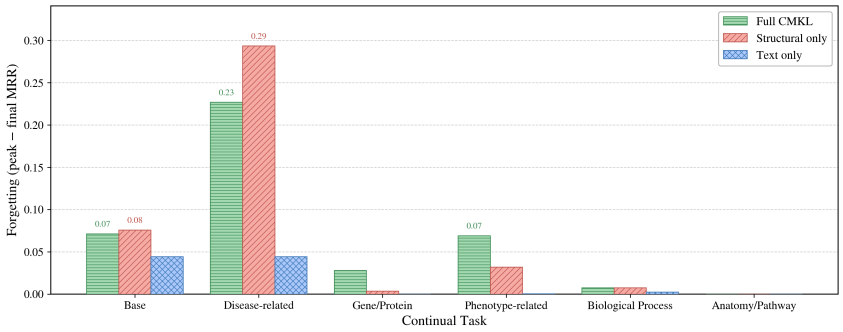

Figure S1: Per-task forgetting (peak − final MRR) by modality across the six real tasks. Grouped bars: full CMKL (red), structural-only (blue), text-only (green). The structural encoder dominates forgetting; the frozen-text branch is inherently stable (and low-signal for link prediction).

S7 Modality-Specific Forgetting

S8 Continual Learning Component Ablation

Table S1 compares EWC configurations and replay-buffer sizes. Per-encoder and uniform λ=10 match within seed noise; removing EWC preserves AP but raises AF, so replay is the primary forgetting-mitigation mechanism with EWC secondary. Performance is robust across a 50× λ range and buffers of 500–5,000. Numbers here are reported at all-task scale (including the four synthetic tasks) purely for this internal configuration comparison, where the relative ordering across configurations is scale-invariant; all main-text LP numbers (Table 1) are on the six real tasks.

Table S1: Continual learning component ablation (5 seeds, MoE-DistMult, all-task scale; internal configuration comparison only). (a) EWC strategy. (b) λ sensitivity (per-encoder, scaled by a common factor relative to default). (c) Replay buffer size.

Configuration	**AP** ↑	**AF** ↓
(a) EWC strategy		
Per-encoder EWC (default)	0.062 ± 0.010	0.043 ± 0.007
Uniform EWC (λ=10)	0.063 ± 0.003	0.043 ± 0.002
No EWC (λ=0)	0.064 ± 0.006	0.050 ± 0.009
(b) λ sensitivity (relative to default)		
0.1×	0.059 ± 0.007	$\mathbf{0.039 \pm 0.001}$
0.5×	0.064 ± 0.003	0.040 ± 0.003
1× (default)	0.062 ± 0.010	0.043 ± 0.007
2×	0.064 ± 0.011	0.044 ± 0.007
5×	$\mathbf{0.065 \pm 0.005}$	0.042 ± 0.006
(c) Replay buffer size		
500	0.060 ± 0.005	$\mathbf{0.041 \pm 0.009}$
1,000 (default)	0.062 ± 0.010	0.043 ± 0.007
2,000	$\mathbf{0.066 \pm 0.007}$	0.049 ± 0.001
5,000	0.062 ± 0.004	0.048 ± 0.007

S9 Data Availability

An anonymized code repository is available for review (linked in the abstract); the full code and pretrained CMKL models will be released publicly on GitHub upon publication.