Journal Title
XX(X):2–60
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

A Scoping Review of Neurosymbolic Reasoning over Ontologies and Knowledge Graphs: Datasets, Evaluation Practices, and Open Challenges

Julie Loesch[1] and Michel Dumontier[1] and Remzi Celebi[1]

Abstract

Despite growing interest in neurosymbolic AI, there is a lack of standardized benchmark datasets specifically designed to evaluate and compare neurosymbolic reasoning systems, particularly those leveraging knowledge graphs and ontologies. This study presents a scoping review of datasets used to evaluate neurosymbolic reasoning frameworks, with an emphasis on knowledge graph– and ontology-based approaches. We first review prominent neurosymbolic frameworks and techniques, then examine existing benchmark datasets, followed by an analysis of reasoning tasks and evaluation metrics they support. Based on these findings, we identify key limitations and outline directions for improving future dataset design. The review highlights substantial gaps in the current benchmarking landscape, including limited dataset diversity, inconsistent evaluation practices, and inadequate support for integrated neurosymbolic reasoning tasks. Addressing these issues is essential for enabling robust, interpretable, and comparable neurosymbolic AI systems.

Keywords

Scoping Review, Neurosymbolic Artificial Intelligence, Benchmark, Knowledge Graph, Ontology

Introduction

Neurosymbolic AI is an emerging subfield that combines two prominent AI paradigms: neural and symbolic methods (Hitzler and Sarker, 2022; Besold et al., 2017b; Garcez et al., 2019). Neural systems excel at pattern recognition, adaptability to new or noisy data, and capturing complex, non-linear relationships, but they often lack transparency and require large amounts of training data (Marcus, 2020). Symbolic systems, on the contrary, provide logical reasoning, interpretability capabilities, and explicit knowledge representation, yet face challenges with scalability and adaptability to novel data. Integrating these paradigms allows neurosymbolic AI to leverage the strengths of both approaches while mitigating their individual limitations.

Knowledge graphs and ontologies provide structured and explicit representations of domain knowledge that are accessible to both neural and symbolic reasoning modules (Gruber, 1993; Hitzler and Sarker, 2022). They encode domain concepts and relations to support logical inference, while neural models complement this by learning from unstructured or noisy inputs. As a result, these structured resources often serve as the symbolic backbone through which neurosymbolic systems connect statistical learning to formal reasoning. Within this paradigm, knowledge representation and reasoning (KRR) frameworks, particularly knowledge graphs (KGs) and ontologies, constitute the most mature symbolic foundations, offering formally grounded semantics and well-established inference mechanisms (Besold et al., 2017a; Hogan et al., 2020; Baader et al., 2007).

Several benchmark datasets have been developed for neural learning over KGs, including FB15k and its variants (Bordes et al., 2013; Toutanova and Chen, 2015), WN18/WN18RR (Bordes et al., 2013; Dettmers et al., 2018), and YAGO-based resources (Suchanek et al., 2007). These datasets are primarily assessed on embedding-based tasks, such as link prediction or knowledge graph completion, often implicitly encoding formal semantics and logical constraints. Conversely, the Semantic Web and Description Logic communities have developed benchmarks for ontology reasoning and deductive inference (Motik et al., 2008; Glimm et al., 2014), which focus on logical correctness rather than hybrid neural-symbolic integration.

Despite rapid progress in neurosymbolic AI, evaluation practices remain fragmented. There is no commonly accepted framework of benchmarks, reasoning

[1]Department of Advanced Computing Sciences, Maastricht University, Netherlands

Corresponding author:
Julie Loesch
Email: julie.loesch@maastrichtuniversity.nl

tasks, and metrics for systematically comparing neurosymbolic reasoning systems across integration paradigms and reasoning settings. Many systems rely either on heterogeneous public knowledge graphs designed for predictive tasks, such as link prediction/KG completion, or ontology-centric OWL/DL benchmarks that focus on deductive reasoning. These evaluation regimes are typically disjoint and often isolate symbolic inference from neural learning dynamics (Singh, 2023).

While prior work (Manhaeve et al., 2026; Delplanque et al., 2025; BOUGZIME et al., 2025) provides valuable insights into specific frameworks or task-oriented benchmarking efforts, these studies do not systematically map the landscape of available datasets, reasoning tasks, and evaluation metrics for KG- and ontology-based neurosymbolic systems. As a result, empirical findings remain difficult to compare and reproducibility is limited.

This gap motivates the present scoping review, which focuses on neurosymbolic techniques and benchmarks that utilize ontologies and knowledge graphs. The review aims to provide a systematic catalogue of datasets, analyze prevalent reasoning tasks and evaluation metrics, and identify limitations to guide the design of future benchmarks.

Specifically, this scoping review addresses the following objectives:

- Provide an overview of neurosymbolic reasoning systems and frameworks that integrate ontologies and knowledge graphs.
- Present a systematic catalogue of benchmark datasets based on ontologies and knowledge graphs for evaluating neurosymbolic AI.
- Discuss prevalent ontology- and knowledge graph–centric reasoning tasks and evaluation metrics used in the field.
- Identify key limitations in existing ontology- and knowledge graph–based datasets and evaluation practices, thereby improving future dataset development and benchmarking standards.

The remainder of this paper is structured as follows:

- Background defines neurosymbolic AI, its fundamental concepts, and key principles that underpin this interdisciplinary field.
- Methods details the systematic search strategy and criteria used to identify and select relevant neurosymbolic techniques and benchmark datasets with a focus on ontologies and knowledge graphs.
- Results presents an overview of existing neurosymbolic frameworks and techniques that integrate ontologies and knowledge graphs. In addition, it presents a comprehensive catalogue of benchmark datasets used to evaluate such neurosymbolic AI systems and reviews the reasoning tasks and evaluation metrics employed.
- Discussion critically examines the limitations associated with current ontology- and knowledge graph–based benchmarks and evaluation practices, and explores potential directions for future research and improvements.
- Finally, Conclusion summarizes the key findings and contributions of this scoping review.

Background

Recent efforts have focused on combining symbolic and neural reasoning, giving rise to neurosymbolic AI, which is a novel research area that seeks to integrate traditional rule-based AI approaches with modern deep learning techniques (Susskind et al., 2021; Besold et al., 2017b; Garcez et al., 2019). This hybrid approach aims to leverage the strengths of symbolic AI, such as explicit knowledge representation and logical inference, with the adaptability and pattern recognition capabilities of neural networks.

Symbolic reasoning systems often struggle with ambiguous and noisy data because they rely on rigid, monotonic rule sets that are difficult to revise once encoded (Russell and Norvig, 2016). Expert systems, for example, operate under monotonic logic, meaning that adding new rules can only increase the knowledge base without retracting previously encoded beliefs (Newell and Simon, 1980). This rigidity limits their ability to adapt to new or conflicting information. In contrast, deep learning models applied to knowledge graphs and reasoning tasks demonstrate robustness to noisy and incomplete data but suffer from a lack of interpretability and explicit reasoning processes (Ebrahimi et al., 2021a; Marcus, 2020). Neurosymbolic systems emerge to compensate for these limitations by combining symbolic rigor with neural flexibility (Yu et al., 2023; d'Avila Garcez et al., 2020).

Given the breadth of methods labeled neurosymbolic, it is useful to introduce an organizing taxonomy. Henry Kautz proposed a widely used framework, which distinguishes approaches by the mode of interaction between neural learning and symbolic reasoning (Kautz, 2022). Table 1 presents this taxonomy.

Table 1. Neurosymbolic integration paradigms from Henry Kautz (Kautz, 2022).

Name	**Name (Kautz)**	**Explanation**
Neural Processing in a Symbolic I/O Pipeline **(Type 1)**	Symbolic → Neuro → Symbolic	Symbolic inputs are transformed into neural representations, processed by a neural model, and decoded back into symbolic outputs. The neural component operates internally within a symbolic input–output framework.
Symbolic Reasoning with Neural Guidance **(Type 2)**	Symbolic[Neuro]	A predominantly symbolic system that leverages neural networks for specific subtasks within a logical reasoning framework. Neural components provide guidance or evaluation to assist symbolic processes.
Parallel Neuro–Symbolic Systems **(Type 3)**	Neuro | Symbolic	Neural and symbolic systems operate as distinct components that communicate and exchange intermediate results, but remain architecturally separate.

Name	**Name (Kautz)**	**Explanation**
Symbolic Constraints for Neural Learning **(Type 4)**	Neuro : Symbolic → Neuro	Symbolic knowledge is employed to constrain, supervise, or regularize the training and behavior of neural networks without necessarily producing explicit symbolic outputs.
Neural Architectures with Embedded Symbolic Reasoning **(Type 5)**	NeuroSymbolic	Logical rules or knowledge bases are directly integrated into neural architectures, shaping their internal representations and influencing generalization.
Neural Networks with Logical Reasoning Layers **(Type 6)**	Neuro[Symbolic]	Neural networks that invoke symbolic reasoning modules as subcomponents during execution, while remaining primarily neural architectures.

Neurosymbolic learning systems exhibit several key advantages: efficiency, generalization, and interpretability. Firstly, they can reason more efficiently than purely symbolic systems because neural components reduce the computational complexity associated with exhaustive search algorithms traditionally used in symbolic reasoning (Besold et al., 2017b). Secondly, these systems show improved generalization over standalone neural networks by leveraging symbolic knowledge as structured guidance or constraints during learning, thereby enhancing performance on unseen data (Garcez et al., 2019). Thirdly, neurosymbolic systems provide greater interpretability compared to black-box neural models. By integrating symbolic representations, they offer "gray-box" reasoning where explicit, traceable computational steps, such as chains of inference or rule applications, can be inspected and explained (d'Avila Garcez et al., 2020; Garcez et al., 2019).

Methods

Overview

We conducted this scoping review following the methodology outlined by Arksey et al. (Arksey and O'Malley, 2005) and further refined by Levac et al. (Levac et al., 2010), and reported it in line with the PRISMA-ScR guidelines (Tricco et al., 2018). The review adhered to five stages: (1) identifying the research questions, (2) identifying relevant studies, (3) selecting studies, (4) extracting and charting data, and (5) collating, summarizing, and reporting the results.

Stage 1: Identifying the research questions

This review aims to examine existing datasets used to evaluate neurosymbolic reasoning systems grounded in knowledge graphs and ontologies, and to systematically analyze their limitations. This objective gives rise to the following research questions (RQs):

RQ 1 What neurosymbolic reasoning systems have been developed for reasoning over ontologies and knowledge graphs?

RQ 2 How are these systems evaluated?

- **RQ 2.1** Which ontology- and knowledge graph-based benchmark datasets have been used to assess neurosymbolic reasoners?
- **RQ 2.2** What ontology- and knowledge graph-centric reasoning tasks and evaluation metrics are applied to assess neurosymbolic reasoning systems?

RQ 3 What limitations do existing ontology- and knowledge graph–based datasets present, and how can evaluation methodologies in neurosymbolic reasoning be improved?

Stage 2: Identifying relevant studies

A systematic search strategy was developed to identify studies addressing neurosymbolic reasoning systems and their evaluation. Searches were conducted across Google Scholar, IEEE Xplore, and the ACM Digital Library using predefined keyword combinations. Our search strategy included variations of "neurosymbolic" (e.g., neuro-symbolic, neurosymbolic, neural-symbolic) combined with terms related to reasoning systems and frameworks (e.g., reasoner, reasoning system, framework) and benchmarking concepts (e.g., benchmark, benchmark dataset, evaluation). We also incorporated keywords reflecting dataset limitations and future directions to ensure a broad coverage aligned with our research objectives. The search queries were constructed using Boolean operators to maximize retrieval of relevant literature, encompassing frameworks, benchmark datasets, reasoning tasks, evaluation metrics, and identified limitations in the field. The full search queries are provided in Full Search Queries (Appendix).

To capture both foundational work and recent advancements, we applied no restrictions on publication year. Additionally, no database-level language restrictions were imposed. Google Scholar queries were executed on 15 December 2025, while IEEE Xplore and the ACM Digital Library were searched on 25 February 2026. Query 6 was executed across all three databases on 25 February 2026.

The initial search retrieved 984 records from Google Scholar, 202 from the ACM Digital Library, and 81 from IEEE Xplore.

Stage 3: Study Selection

Due to the substantial number of records retrieved from Google Scholar, screening was restricted to the first 200 results per query, consistent with published recommendations and reported practice when using Google Scholar in evidence syntheses (Haddaway et al., 2015; Bramer et al., 2017). Query 6 was further limited to the first 100 results owing to significant duplication among retrieved records, which reduced the yield of unique contributions.

For IEEE Xplore and the ACM Digital Library, screening was limited to the first 100 results per query, reflecting their more focused and domain-specific indexing. All six queries were executed in Google Scholar to maximize coverage across systems, benchmarks, and evaluation practices. In contrast, only Query 1 and Query 6 were applied to IEEE Xplore and the ACM Digital Library to retrieve primary research contributions on neurosymbolic reasoning systems.

After duplicate removal, manuscripts were initially assessed for relevance based on their titles and abstracts. Title and abstract screening was conducted by a single reviewer according to predefined eligibility criteria.

To ensure a focused and relevant selection of studies, the following inclusion and exclusion criteria were applied during the screening process.

Inclusion criteria:

- Studies addressing neurosymbolic integration approaches involving ontologies, knowledge graphs, or knowledge bases.
- Studies proposing, analysing, or evaluating reasoning frameworks, inference mechanisms, benchmarks, datasets, or evaluation methodologies related to neurosymbolic systems over ontologies or knowledge graphs.
- Empirical studies, conceptual frameworks, reviews, or surveys relevant to the research questions.
- Publications written in English.
- Full-text journal articles and conference papers.

Exclusion criteria:

- Studies not focusing on neurosymbolic integration involving ontologies, knowledge graphs, or knowledge bases.
- Studies addressing neural or symbolic approaches in isolation, without integration or application to ontologies or knowledge graphs.
- Editorials, extended abstracts, posters, patents, or other non–peer-reviewed opinion-based publications.
- Publications not written in English.
- Studies without accessible full text.
- Studies lacking a clearly described methodology or empirical evaluation.
- Studies unrelated to ontologies or knowledge graphs (e.g., those focusing exclusively on image-based or text-only data).

Following title and abstract screening, the remaining studies underwent full-text screening. At this stage, studies were excluded for the following primary reasons: (1) ineligible publication type (e.g., non–peer-reviewed or out-of-scope formats); (2) absence of original empirical contribution; (3) omission of required components such as ontologies, knowledge graphs, or reasoning mechanisms; and (4) unavailability of the full text.

Stage 4: Charting the data

An iterative and collaborative process was employed to design the data charting forms. The selected studies were systematically reviewed, and relevant information was extracted into standardized data charting tables.

Two complementary charting tables were developed. The first table characterizes the included neurosymbolic reasoning systems by documenting the symbolic formalism, neural learning component, neural–symbolic integration paradigm, datasets, reasoning tasks, and evaluation metrics.

The second table targeted benchmark datasets used for evaluating neurosymbolic reasoning over ontologies and knowledge graphs. For each dataset, we captured its application domain, the reasoning tasks it enables, and the systems evaluated on the dataset in relation to those tasks.

Stage 5: Collating, summarizing, and reporting the results

The extracted data were collated, synthesized, and reported in accordance with the review's research questions. Results were organized to provide a structured overview of existing neurosymbolic reasoning systems and benchmark datasets, enabling a coherent narrative that directly addresses the objectives of the scoping review and highlights prevailing trends, gaps, and limitations in the literature.

Results

To structure our review, we develop a multi-dimensional taxonomy (Figure 1) that also defines the scope of this work. We focus on hybrid neural–symbolic systems operating over ontologies and knowledge graphs, explicitly excluding purely neural or purely symbolic approaches.

We characterize each included system along four dimensions: (1) the underlying knowledge representation, (2) the neural component, (3) the symbolic component, and (4) the integration strategy linking neural and symbolic processes. However, the review is organized primarily along the integration dimension to enable a coherent comparison across neurosymbolic paradigms. For this purpose, we adopt Kautz's taxonomy (Kautz, 2022), which distinguishes approaches by the mode of interaction between neural learning and symbolic reasoning. The six integration paradigms are summarized in Table 1 (Background).

Literature Search

The database search yielded 984 records from Google Scholar, 202 from the ACM Digital Library, and 81 from IEEE Xplore. After duplicate removal, 917 unique papers remained for screening. Title and abstract screening identified 275 potentially relevant articles for further assessment.

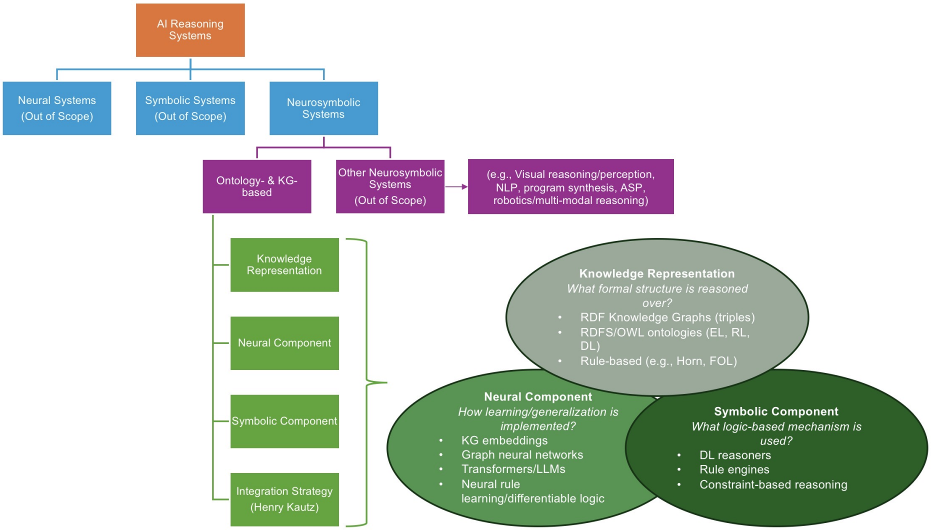
AI Reasoning Systems
Neural Systems (Out of Scope)
Symbolic Systems (Out of Scope)
Neurosymbolic Systems
Ontology- & KG-based
Other Neurosymbolic Systems (Out of Scope)
(e.g., Visual reasoning/perception, NLP, program synthesis, ASP, robotics/multi-modal reasoning)
Knowledge Representation
Neural Component
Symbolic Component
Integration Strategy (Henry Kautz)
Knowledge Representation
What formal structure is reasoned over?
• RDF Knowledge Graphs (triples)
• RDFS/OWL ontologies (EL, RL, DL)
• Rule-based (e.g., Horn, FOL)
Neural Component
How learning/generalization is implemented?
• KG embeddings
• Graph neural networks
• Transformers/LLMs
• Neural rule learning/differentiable logic
Symbolic Component
What logic-based mechanism is used?
• DL reasoners
• Rule engines
• Constraint-based reasoning

Figure 1. Taxonomy of neurosymbolic reasoning systems operating over ontologies and knowledge graphs.

Full-text screening was subsequently performed to evaluate eligibility with respect to the scope of this review. Articles were excluded if they did not focus on ontology- or knowledge graph–based neurosymbolic systems or if they did not address the defined research questions. Following this process, 63 articles were deemed eligible and included in the scoping review. The study selection process was conducted in accordance with the PRISMA-ScR guidelines (Tricco et al., 2018) and is summarized in the PRISMA-ScR flow diagram shown in Figure 2.

In parallel, the included articles were examined to compile an initial inventory of benchmark datasets. This process yielded 103 candidate datasets. Datasets were excluded if they primarily contained image-based or lexical data, or did not support ontology- or knowledge graph–centric reasoning tasks. Following this filtering process, a total of 83 datasets were retained for the data charting stage. Dataset variants (e.g., filtered subsets, alternative splits, and successive releases) were retained as separate entries to reflect how benchmarks are cited and used in the literature.

Notable benchmark families with multiple retained variants include FB15k/FB15k-237, WN18/WN18RR, LC-QuAD (1.0–2.1), QALD (8–9), and DBpedia-derived subsets/alignment benchmarks (DBpedia/DBpedia20k; DBP15K/DBP1M).

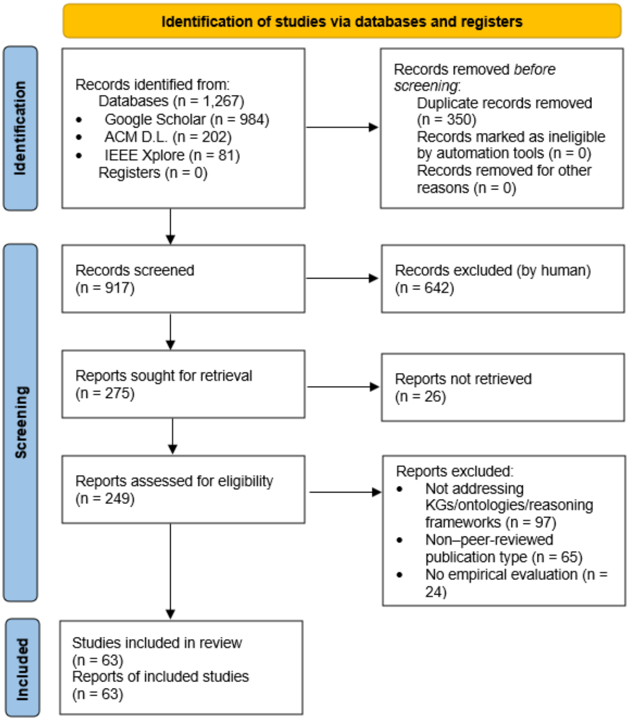
Identification of studies via databases and registers
Identification
Records identified from:
Databases (n = 1,267)
• Google Scholar (n = 984)
• ACM D.L. (n = 202)
• IEEE Xplore (n = 81)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 350)
Records marked as ineligible by automation tools (n = 0)
Records removed for other reasons (n = 0)
Screening
Records screened (n = 917)
Records excluded (by human) (n = 642)
Reports sought for retrieval (n = 275)
Reports not retrieved (n = 26)
Reports assessed for eligibility (n = 249)
Reports excluded:
• Not addressing KGs/ontologies/reasoning frameworks (n = 97)
• Non–peer-reviewed publication type (n = 65)
• No empirical evaluation (n = 24)
Included
Studies included in review (n = 63)
Reports of included studies (n = 63)

Figure 2. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram for the scoping review process.

Neurosymbolic systems (RQ 1)

To address RQ 1, we categorize the identified systems using Kautz's taxonomy, which distinguishes six paradigms for integrating neural learning with symbolic reasoning. The core distinction across these paradigms is *where* symbolic knowledge enters the processing pipeline and *how* neural computation interacts with symbolic representations. We next organize the reviewed systems by integration cluster and summarize, for each group, the datasets, reasoning tasks, and evaluation metrics.

Type 1: Neural Processing in a Symbolic I/O Pipeline. Type 1 systems keep symbolic representations at the input and output, while a neural model performs the central mapping: symbolic inputs (e.g., triples, axioms, or structured templates) mapped into neural representations, the model predicts or scores candidates, and the results are decoded back into symbolic outputs (e.g., inferred triples, predicted axioms, or ranked candidate relations) that are compared against ontology- or reasoner-derived ground truth.

In our corpus, one theme is *template-based symbolic output selection*, where HTL (Hereditary attentive Tree-LSTM) (Gomes et al., 2022) represents questions via a symbolic template inventory and trains a neural model to select the correct template class, yielding a symbolic template identifier as output. A second theme is *learning to approximate deductive closure*, where CFR (ChunfyReasoner) (Zhu et al., 2023) uses a symbolic reasoner to materialize entailments as training/evaluation targets and trains a neural model to reproduce these inferences by outputting symbolic inferred triples for fast approximate reasoning. A third theme is *predicting ontology relations, especially subsumption*, where the box-embedding approach (Memariani et al., 2025) learns geometric representations (e.g., SMILES→vectors; classes→boxes) and decodes them into symbolic ontology relations such as subsumption, disjointness, and overlap, and where BERTSubs (Chen et al., 2023) builds symbolic context from the ontology hierarchy and outputs ranked candidate subsumptions. Finally, a fourth theme is *embedding-based reasoning constrained by axioms*, where EmEL++ (Mondal et al., 2021) uses symbolic $\mathcal{EL}^{++}$axioms (and ELK inferences) to define training/evaluation constraints so that subsumption is preserved geometrically (e.g., via containment), and EBR (Kamdem Teyou et al., 2025) uses reasoner outputs as ground truth to learn embeddings that reconstruct complex concept semantics and yield approximate instance sets for instance retrieval and downstream concept learning.

Overall, Type 1 systems in our corpus preserve symbolic compatibility at the interfaces while delegating the core mapping between symbolic inputs and outputs to a neural model. As summarized in Table 2, they are benchmarked on tasks such as question answering/template selection, ontology-relation prediction (notably subsumption), and instance retrieval, and are evaluated using task-appropriate predictive, ranking, and similarity-based metrics, often complemented by runtime when efficiency is a key motivation.

Table 2. Type 1 neurosymbolic reasoners.

System	**Dataset(s)**	**Task(s)**	**Metric(s)**
HTL (Hereditary attentive Tree-LSTM) (Gomes et al., 2022)	LC-QuAD 1.0; LC-QuAD 2.0; LC-QuAD 2.1; WebQSP; ComplexWebQuestions	Knowledge Graph Question Answering; Semantic Parsing; Template Classification	Accuracy; Precision; Recall; F1
CFR (ChunfyReasoner) (Zhu et al., 2023)	Family; Time	Subsumption; Membership; Link Prediction	Precision; Recall; F1; Runtime
A box-embedding approach (Memariani et al., 2025)	ChEBI	Subsumption; Disjointness; Overlap; Hierarchical Multi-label Classification	Precision; Recall; F1

System	Dataset(s)	Task(s)	Metric(s)
BERTSubs (Subsumption prediction method) (Chen et al., 2023)	NCIT-DOID; HeLiS; FoodOn	Subsumption	MRR; Hits@K
EmEL++ (Embeddings for $\mathcal{EL}^{++}$) (Mondal et al., 2021)	SNOMED CT; Anatomy; Gene Ontology (GO); GALEN	Subsumption	Hits@K; AUC-ROC; Median Rank; 90th Percentile Rank; Accuracy
EBR (Embedding Based Reasoner) (Kamdem Teyou et al., 2025)	Carcinogenesis; Mutagenesis; Semantic Bible; Vicodi; Family; Father	Instance Retrieval; Robust Reasoning under Incompleteness/Inconsistency; Concept Learning Support	Jaccard Similarity; F1; Runtime

Type 2: Symbolic Reasoning with Neural Guidance. Type 2 systems are *symbolic-first*: an explicit symbolic procedure (e.g., query/program execution, rule-based deduction, or probabilistic logic inference) remains responsible for the overall reasoning and for producing the final output. Neural components are integrated as auxiliary modules that *guide* this symbolic process by proposing candidates, scoring alternatives, learning rule parameters, or prioritizing search decisions. In other words, neural models do not replace symbolic reasoning in Type 2; they support it by making symbolic inference more efficient, robust, or data-adaptive.

In our corpus, one theme is *neural guidance for symbolic query/program execution*. NS-KGQA (Agarwal and Bedathur, 2025) builds a structured symbolic program or question-specific subgraph and uses neural embeddings to guide grounding and scoring decisions while the symbolic resolver executes the reasoning steps. Similarly, TeBaQA (Vollmers et al., 2021) uses a learned component to predict an appropriate SPARQL pattern, after which a symbolic pipeline instantiates the query, applies modifiers and consistency checks, and executes SPARQL to obtain the answer. A second theme is *symbolic enrichment followed by neural prediction*. The neuro-symbolic link prediction system in (Rivas et al., 2024) applies symbolic deduction to augment the knowledge graph (reducing sparsity and making relations explicit) and then trains a neural embedding model to predict missing links on the enriched graph. A third theme is *neural generation paired with symbolic validation* for knowledge graph enrichment and ontology evolution: the context-aware hybrid neuro-symbolic KG enrichment framework (Boulakbech and Wannous, 2025) uses an LLM to propose candidate facts, while a symbolic layer validates and aligns them with ontological constraints to support ontology extension decisions (including human-in-the-loop oversight). Finally, a fourth theme is *neural parameterization and constraint-based filtering in symbolic inference*. DPLogic (Li et al., 2025) and DiffLogic (Shengyuan et al., 2023) retain symbolic probabilistic logic reasoning with weighted rules, while neural embeddings provide learned rule weights and/or differentiable truth estimates, typically via alternating optimization. Along related lines, KOSMOS (Purohit et al., 2025b) uses neural embeddings to generate candidate predictions that are subsequently filtered by symbolic constraints (e.g., SHACL) to enforce domain consistency for medical discovery.

Overall, Type 2 systems in our corpus keep symbolic reasoning "in the loop" and use neural components to guide, parameterize, or validate symbolic inference steps. As summarized in Table 3, they are benchmarked primarily on knowledge graph question answering, link prediction, and knowledge graph enrichment/ontology evolution, with evaluation combining standard predictive and ranking metrics with task-specific indicators of constraint satisfaction and ontology consistency where applicable.

Table 3. Type 2 neurosymbolic reasoners.

System	**Dataset(s)**	**Task(s)**	**Metric(s)**
NS-KGQA (Agarwal and Bedathur, 2025)	KQA Pro; MetaQA; WebQSP	Knowledge Graph Question Answering	Accuracy; Hits@K; F1
A neuro-symbolic link prediction system (Rivas et al., 2024)	Lung cancer polypharmacy treatment KG (clinical records + DrugBank DDIs)	Link Prediction	Precision; Recall; F1
A context-aware hybrid neuro-symbolic KG enrichment framework (Boulakbech and Wannous, 2025)	Datatourisme KG	Knowledge Graph Enrichment; Ontology Evolution; Constraint Validation	Precision; Recall; F1; Ontology Match Rate; Constraint Satisfaction Rate; Ontology Extension Accuracy
DPLogic (Differentiable Probabilistic Logic) (Li et al., 2025)	WN18RR; FB15k-237; Kinship; UMLS; Family	Link Prediction	MRR; Hits@K
DiffLogic (Differentiable Logic) (Shengyuan et al., 2023)	YAGO3-10; WN18; WN18RR; CodeX; Kinship	Link Prediction; Probabilistic Logic Reasoning	MRR; Hits@K

System	Dataset(s)	Task(s)	Metric(s)
KOSMOS (Knowledge Oriented Symbolic learning for Medical Ontology-based decision System) (Purohit et al., 2025b)	Lung cancer KG	Link Prediction	MRR; Hits@K
TeBaQA (Vollmers et al., 2021)	QALD-8; QALD-9; LC-QuAD 1.0; LC-QuAD 2.0	Knowledge Graph Question Answering; Semantic Parsing; Template Classification	Precision; Recall; F1 (QALD); Runtime

Type 3: Parallel Neuro-Symbolic Systems. Type 3 systems keep neural and symbolic components as distinct modules that operate side-by-side and exchange intermediate representations. Typically, a neural component proposes candidates or produces a structured intermediate output, while a symbolic component applies explicit rules, constraints, or logic-based operators to refine, validate, or augment these intermediates; the resulting signals can then be fed back to the neural component, yielding iterative improvement.

In our corpus, one theme is *parallel neural scoring and symbolic rule reasoning for knowledge graph completion.* FaSt-FLiP (Khojasteh et al., 2023) runs a neural ("fast") predictor alongside a symbolic ("slow") rule component, filters candidate rules using a neural verifier, and combines neural scores with rule-derived candidates to produce final link predictions and explanations. IterE (Zhang et al., 2019) similarly couples neural embeddings and symbolic rules, but in an explicitly iterative loop: embeddings are learned, rules are extracted and pruned based on the learned embedding structure, and rule-inferred triples are injected back to improve subsequent embedding learning rounds. Poderoso (Martinez Lorenzo et al., 2025) follows a modular design in which a neural embedding model produces candidate predictions and a symbolic module reasons over them to filter and/or augment the outputs.

A second theme is *alignment as modular candidate generation, neural refinement, and symbolic consolidation.* NeSyMatch (Sharma and Jain, 2025) uses symbolic cues for candidate generation and structural constraints, applies neural scoring to refine candidate matches, and then uses symbolic post-processing to finalize alignments. NeuSymEA (Chen et al., 2025) adopts a variational EM-style procedure in which a neural model proposes entity alignments and a symbolic rule-based inference component updates constraints or weights, iterating until convergence. Finally, ENeSy (Xu et al., 2022) illustrates *neural-symbolic collaboration for complex query answering*: a neural projection generates intermediate candidates, symbolic operators revise or complete these results, and final answers are obtained by combining the neural and symbolic outputs.

Overall, Type 3 systems exhibit bidirectional interaction between neural prediction modules and symbolic refinement components. As reflected in Table 4, the dominant applications are link prediction and complex query answering in knowledge graphs, as well as entity and ontology alignment, with evaluation primarily based on ranking metrics (e.g., MRR and Hits@K) and, where applicable, alignment quality measures (Precision/Recall/F1) and efficiency or rule-quality indicators.

Table 4. Type 3 neurosymbolic reasoners.

System	**Dataset(s)**	**Task(s)**	**Metric(s)**
FaSt-FLiP (Fast and Slow Thinking with Filtered rules for Link Prediction task) (Khojasteh et al., 2023)	FB15k-237	Link Prediction	Hits@K; MRR
IterE (Zhang et al., 2019)	FB15k; FB15k-237; WN18; WN18RR	Link Prediction; Rule Learning	MRR; Hits@K; Runtime; Head Coverage; High Quality Rules (count); High Quality Rules (%)
Poderoso (Martinez Lorenzo et al., 2025)	LUBM; DBpedia20k	Link Prediction	MRR; Runtime
NeSyMatch (Sharma and Jain, 2025)	OAEI Bio-ML 2023	Ontology Alignment	Precision; Recall; F1; MRR; Hits@K
ENeSy (Xu et al., 2022)	FB15k-237; NELL-995	Complex Query Answering	MRR
NeuSymEA (Chen et al., 2025)	DBP15K; OpenEA; DBP1M	Entity Alignment	Hits@K; MRR

Type 4: Symbolic Constraints for Neural Learning. Type 4 systems use symbolic knowledge primarily to *shape neural learning*. Symbolic rules, ontological constraints, or deductive closure are used during training (and sometimes evaluation) to generate supervision, regularize model parameters, constrain predictions, or structure the learning problem so that the resulting neural model better respects known logical or structural properties.

In our corpus, one theme is *constraint-driven grounding and revision under description logic*. EmALC (Wu and Zhao, 2025) starts from neural groundings and subsequently revises or optimizes them to satisfy $\mathcal{ALC}$constraints, using the ontology as a regularizer of the final grounding. A second theme is *using deductive closure and constraints to construct training signals and evaluation protocols*. CoPCA (Purohit et al., 2025a) uses symbolic constraints to validate or generate training examples (e.g., valid/invalid triples) that guide neural embedding learning. Along similar lines, DELE (Mashkova et al., 2026) uses symbolic deduction to filter negative samples and to structure evaluation in $\mathcal{EL}^{++}$completion, while ELEmbeddings with negative sampling and deductive closure filtering (Mashkova et al., 2024) uses closure-aware negative sampling and filtered evaluation to prevent logical leakage. The walking-rdf-and-owl method (Alshahrani et al., 2017) likewise leverages symbolic closure so that the learned embeddings encode both asserted and inferred knowledge.

A third theme is *logic-regularized neural representations for complex query answering*. DAGE (He et al., 2025) regularizes neural query embeddings with logical constraints so learned representations better preserve logic properties when answering complex DAG queries. A fourth theme uses symbolic structure as *training guidance for multi-hop reasoning agents*. RKLE (Liu et al., 2024) and PoLo (Liu et al., 2021b) incorporate symbolic logical structure into the learning objective, for instance, via rewards, constraints, or rule guidance to steer neural path reasoning toward rule-consistent multi-hop explanations and improved link prediction. RRN (Hohenecker and Lukasiewicz, 2020) follows a related idea by using symbolic rules to generate training targets, enabling the neural model to approximate entailment without invoking a symbolic reasoner at inference.

Finally, Type 4 also includes approaches where symbolic artifacts guide embedding learning for *ontology-level prediction and alignment*, and where constraints regularize uncertainty-aware models. OWL2Vec* (Chen et al., 2021a) uses an OWL ontology (and optionally its entailments) to generate an embedding training corpus that supports missing-axiom prediction via embedding similarity or ranking; OWL2Vec4OA (Teymurova et al., 2024) uses alignment seeds to steer embedding learning and improve ontology alignment ranking. BEUrRE (Chen et al., 2021b) combines neural uncertainty modeling with symbolic constraints that encourage global consistency. In addition, some systems leverage symbolic guidance for *recommendation*: KGTORe (Mancino et al., 2023) uses structured decision paths and knowledge graph semantics to regularize neural recommendation learning, while the neuro-symbolic KGE-based recommender framework (Spillo et al., 2024) injects first-order rules into embedding learning to obtain rule-aware representations that improve recommendation quality beyond accuracy.

Overall, Type 4 systems in our corpus operationalize symbolic knowledge mainly as supervision and regularization for neural models, spanning settings from link prediction and complex query answering to ontology completion/alignment, uncertainty-aware prediction, and recommendation. This breadth is reflected in the tasks and evaluation measures reported in Table 5, which combine standard predictive and ranking metrics with task-specific indicators (e.g., robustness, constraint satisfaction, or beyond-accuracy recommendation measures) where applicable.

Table 5. Type 4 neurosymbolic reasoners.

System	Dataset(s)	Task(s)	Metric(s)
EmALC (Wu and Zhao, 2025)	Ontodm; Nifdys; Nihss; GlycoRDF; Sso; Family; Family2	Masked ABox Revision; Complex Query Answering	Success Rate; KL Divergence; Precision; Recall
CoPCA (Purohit et al., 2025a)	DB100K; YAGO3-10; French Royalty	Link Prediction	MRR; Hits@K
DAGE (He et al., 2025)	FB15k-237; FB15k; NELL	Complex Query Answering	MRR; Runtime
DELE (Mashkova et al., 2026)	GO+STRING; FoodOn; GALEN	Link Prediction (PPI); Subsumption	Hits@K; Mean Rank; AUC-ROC
ELEmbeddings with Negative Sampling and Deductive Closure Filtering (Mashkova et al., 2024)	GO+STRING	Link Prediction (PPI); Link Prediction	Hits@K; Mean Rank; AUC-ROC
KGTORe (Mancino et al., 2023)	MovieLens 1M; Yahoo! Movies; Facebook Books	Recommendation	nDCG; Precision; Hits@K; Recall
RKLE (Reinforcement Learning-Based Knowledge Reasoning Model with Logical Embedding) (Liu et al., 2024)	FB15k-237; WN18RR; NELL-995	Multi-hop Reasoning; Path Reasoning; Link Prediction	MRR; Hits@K; MAP
PoLo (Liu et al., 2021b)	Hetionet	Link Prediction; Multi-hop Reasoning	Hits@K; MRR
walking-rdf-and-owl (neuro-symbolic representation learning method) (Alshahrani et al., 2017)	Gene Ontology (GO); HPO; Disease Ontology (DO); SwissProt GO annotations; HPO; STRING; STITCH; DisGeNET; SIDER	Link Prediction; Drug Repurposing	F1; AUC-ROC

System	Dataset(s)	Task(s)	Metric(s)
RRN (Recursive Reasoning Network) (Hohenecker and Lukasiewicz, 2020)	Family Trees; Countries; DBpedia; Claros; UMLS	Membership	Accuracy; F1
OWL2Vec* (Chen et al., 2021a)	HeLiS; FoodOn; Gene Ontology (GO)	Ontology Completion; Membership; Subsumption	Hits@K; MRR
OWL2Vec4OA (Teymurova et al., 2024)	OAEI Bio-ML 2023	Ontology Alignment	MRR; Hits@K
BEUrRE (Chen et al., 2021b)	CN15k; NL27k	Confidence Prediction; Link Prediction	Mean Squared Error; Mean Absolute Error; nDCG
Neuro-symbolic KGE + recommender framework (Spillo et al., 2024)	Last.fm; DBbook; MovieLens 1M	Recommendation	Precision; Recall; F1; MAP; Mean Average Recall (MAR); nDCG; Novelty; Diversity

Type 5: Neural Architectures with Embedded Symbolic Reasoning. Type 5 systems embed symbolic structure directly into the neural architecture, so prior knowledge influences the model through its design rather than only through data. In this paradigm, logical operators, query structure, or ontology semantics are "compiled" into neural computation (e.g., via differentiable constraints, neuralized logical operators, or structured message passing), allowing the model to perform compositional reasoning within the forward pass.

In our corpus, one prominent theme is *neural execution of symbolic queries via learned operators*. ULTRAQUERY (Galkin et al., 2024) embeds logic inside the architecture via fuzzy logical operators to support complex query answering across multiple knowledge graphs. GQEs (Hamilton et al., 2018), CLMPT (Zhang et al., 2024), NewLook (Liu et al., 2021a), and Query2Box (Ren et al., 2020) similarly treat symbolic query structures as computation graphs that are executed through learned neural operators in embedding space, with Query2Box additionally handling union through symbolic rewriting. CaQR (Kim et al., 2024) and LinE (Huang et al., 2022) follow a related approach in which the symbolic query structure conditions the learned query representation and logical operators are encoded as neural transformations, supporting multi-hop reasoning and richer query fragments (e.g., negation).

A second theme is *ontology semantics embedded as differentiable constraints for completion and approximate deduction*. CatE (Zhapa-Camacho and Hoehndorf, 2023) projects symbolic $\mathcal{ALC}$axioms into a graph-derived neural representation and uses the resulting embeddings for deductive and inductive completion. ELEm (Kulmanov et al., 2019), ELBE (Peng et al., 2022), Box2EL (Jackermeier et al., 2024), EmELvar (Mohapatra et al., 2021), and TransBox (Yang et al., 2025) encode (variants of) $\mathcal{EL}/\mathcal{EL}^{++}$semantics as geometric or differentiable constraints, training embeddings that support ranking or predicting axioms while approximating deductive behavior. LNN-MP and LNN-CM (Sen et al., 2021) instantiate a related idea by coupling neural representations with logic-inspired mixture/chain constructions and training them to satisfy closure-style constraints.

A third theme is *learning rules and symbolic functions as internal neural structure*. DegreEmbed (Li et al., 2023) uses embeddings to guide the discovery and selection of symbolic rules that can subsequently support knowledge graph completion. LERP (Han et al., 2023) learns interpretable logical functions as part of the entity representation, enabling differentiable rule learning and optionally regularizing neural embeddings. NeSyKHG (Bhuyan et al., 2024) combines hypergraph representation learning with higher-order symbolic reasoning to improve prediction and interpretability.

Finally, some systems embed logic in architectures to support *explainability and transfer*. KG Deductive Reasoner (Ebrahimi et al., 2021b) treats deductive entailment as the task definition and trains a neural model to approximate deductive reasoning in a way that can transfer across knowledge graphs. KG-LRR (Wang et al., 2025) decodes recommendations via neuralized logic operators constrained by symbolic logic laws, yielding explainable logic expressions. The two-system architecture in (Hua and Zhang, 2022) augments neural representations with an explicit neural logic layer regularized by symbolic constraints and evaluates under logic-aware negatives. BoxE (Abboud et al., 2020) is also included in this cluster as it supports rule-aware completion via architectural or training-time integration of symbolic rules, depending on the configuration.

Overall, Type 5 systems in our corpus operationalize symbolic knowledge by embedding it into neural computation itself, most commonly via neuralized logical operators, query-execution architectures, and ontology-constrained embedding models. This is reflected in the variety of benchmarked tasks summarized in Table 6, spanning complex query answering, link prediction and rule learning, ontology completion/subsumption, knowledge graph entailment, and explainable recommendation.

Table 6. Type 5 neurosymbolic reasoners.

System	Dataset(s)	Task(s)	Metric(s)
ULTRAQUERY (Galkin et al., 2024)	FB15k-237; NELL-995; FB15k; WikiTopics-QA	Complex Query Answering	MRR; Hits@K; AUC-ROC; Mean Absolute Percentage Error

System	Dataset(s)	Task(s)	Metric(s)
BoxE (Abboud et al., 2020)	JF17K; FB-AUTO; SportsNELL; YAGO3-10	Link Prediction; Rule Injection	Mean Rank; MRR; Hits@K
CatE (Zhapa-Camacho and Hoehndorf, 2023)	Food-Biomarker Ontology (FOBI); Neural Reprogramming Ontology (NRO); Gene Ontology (GO); FoodOn; Yeast PPI dataset	Consistency Checking; Subsumption; Ontology Completion; Link Prediction (PPI)	Mean Rank; AUC-ROC; Hits@K
LNN-MP (Logic Neural Networks-Mixture of Paths); LNN-CM (Logic Neural Networks-Chain of Mixtures) (Sen et al., 2021)	WN18RR; FB15k-237; Kinship; UMLS	Link Prediction; Rule Learning	MRR; Hits@K
CLMPT (Conditional Logical Message Passing Transformer) (Zhang et al., 2024)	FB15k; FB15k-237; NELL-995	Complex Query Answering	MRR; Hits@K
DegreEmbed (Li et al., 2023)	FB15k-237; WN18; UMLS; Kinship; Family	Link Prediction; Rule Learning	Hits@K
ELBE ($\mathcal{EL}$Box Embedding) (Peng et al., 2022)	Gene Ontology (GO)	Link Prediction (PPI); Equivalence	Hits@K; Mean Rank; AUC-ROC
Box2EL (Jackermeier et al., 2024)	GALEN; Gene Ontology (GO); Anatomy; GO+STRING	Subsumption; Link Prediction; Approximating Deductive Reasoning	Hits@K; Median Rank; MRR; Mean Rank; AUC-ROC
ELEm ($\mathcal{EL}$Embeddings) (Kulmanov et al., 2019)	GO+STRING	Link Prediction (PPI)	Hits@K; Mean Rank; AUC-ROC
GQEs (Graph Query Embeddings) (Hamilton et al., 2018)	Biological Interaction Network (Bio); Reddit Interaction Network (Reddit)	Complex Query Answering	AUC-ROC; Average Percentile Rank (APR)

System	Dataset(s)	Task(s)	Metric(s)
EmELvar (Mohapatra et al., 2021)	ORE; OWL2Bench	Subsumption	Hits@K; Median Rank; Percentile Rank
FuzzyVis (Zhurov et al., 2025)	HPO; Pizza ontology	Semantic Search; Approximate Query Answering	MRR; Hits@K; Overlap@k; Subsumption violation rate; Similarity vs. distance curve
CaQR (Context-aware Query Representation learning) (Kim et al., 2024)	FB15k-237; NELL	Multi-hop Reasoning; Complex Query Answering	MRR; Runtime
LinE (Line Embedding) (Huang et al., 2022)	FB15k-237; NELL-995; WN18RR	Complex Query Answering (Negation); Multi-hop Reasoning	MRR; Hits@K
LERP (Logical Entity RePresentation) (Han et al., 2023)	UMLS; Kinship; Family; WN18; WN18RR	Link Prediction	MRR; Hits@K
NeSyKHG (Bhuyan et al., 2024)	CMHR (Chinese Medical High-order Relational dataset)	Higher-order Relational Reasoning; Hypergraph Link Prediction	F1; Accuracy; AUC-ROC
NewLook (Liu et al., 2021a)	FB15k; FB15k-237; NELL	Complex Query Answering	Hits@K; MRR
KG Deductive Reasoner (Ebrahimi et al., 2021b)	"OWL-Centric Dataset" (based on Linked Data Cloud and Data Hub)	Knowledge Graph Entailment; Deductive Reasoning	Accuracy; Precision; Recall; F1
Query2Box (Ren et al., 2020)	FB15k; FB15k-237; NELL-995	Complex Query Answering	MRR; Hits@K
KG-LRR (Knowledge Graphs-based Logic Reasoning Recommendation) (Wang et al., 2025)	Yelp2018; Amazon-book; Amazon-electronics	Recommendation (Explainable)	Precision; Recall; nDCG

System	Dataset(s)	Task(s)	Metric(s)
Two-system architecture (System 1 representation learning + System 2 neural logic layer) (Hua and Zhang, 2022)	ConceptNet-100K; WebChild-comparative	Link Prediction (Commonsense)	MRR; Mean Rank; Hits@K
TransBox (Yang et al., 2025)	GALEN; Gene Ontology (GO); Anatomy	Link Prediction; Subsumption	Hits@K; Median Rank; MRR; Mean Rank; AUC-ROC

Type 6: Neural Networks with Logical Reasoning Layers. Type 6 systems are *predominantly neural* at the system level but explicitly invoke a symbolic reasoning module during execution. In this paradigm, the neural component remains responsible for the main representation learning and prediction, while symbolic reasoning is called as a subroutine for steps that benefit from exact symbolic manipulation, such as enforcing constraints, constructing multi-hop explanations, or validating candidate solutions.

In our corpus, one theme is *neural proposal with symbolic path/rule construction for interpretable reasoning.* CAFE (Xian et al., 2020) uses a neural model to propose and filter promising candidates and guiding signals, while a symbolic path-reasoning stage constructs and selects explanatory knowledge-graph paths that support the final recommendation. Similarly, the commonsense reasoner in (Moghimifar et al., 2021) uses a neural model to propose relations and a symbolic component to stitch these predictions into multi-hop rule chains, yielding interpretable reasoning paths for knowledge graph completion. A second theme is *neural prediction coupled with symbolic constraint checking and explanation.* I-SBR and I-DCR (Ontiveros et al., 2026) use a neural component to select or weight reasoning signals, while a symbolic forward-chaining procedure executes rules to produce both the final link prediction and an explicit proof trace; NS-KAG (Rajalakshmi et al., 2025) likewise pairs neural predictions with symbolic constraints/rules that provide a consistency and explanation layer; and NSQA (Kapanipathi et al., 2020) generates candidate symbolic queries with neural modules and uses a logic-based reasoner to evaluate and filter them against knowledge-base evidence before producing answers. A third theme is *structured query answering with symbolic query materialization.* GNNQ (Pflueger et al., 2022) materializes query structure into the knowledge graph and then applies a GNN to reason over this enriched structure under incompleteness, while InductiveQE (Galkin et al., 2022) learns neural representations intended to generalize to unseen entities while relying on explicit symbolic query structures to define the reasoning objective and evaluation protocol. Finally, FD-PORT (Shen et al., 2025) illustrates *symbolic search used to derive supervision for neural policies*: a symbolic procedure generates step-level "good vs. bad" preferences for structured traversal, which are then used to train the neural policy/value model to guide future multi-hop reasoning.

Overall, Type 6 systems in our corpus use symbolic reasoning layers primarily for validation, multi-step compositional structure, and interpretability, while maintaining a neural backbone for representation learning and prediction. Accordingly, the reviewed benchmarks emphasize complex query answering and multi-hop reasoning, explainable link prediction and recommendation, and classification with constraint- or explanation-oriented measures in addition to standard predictive metrics (Table 7).

Table 7. Type 6 neurosymbolic reasoners.

System	Dataset(s)	Task(s)	Metric(s)
CAFE (CoArse-to-FinE neural symbolic reasoning approach) (Xian et al., 2020)	Amazon-CDs&Vinyl; Amazon-Clothing; Amazon-CellPhones; Amazon-Beauty	Recommendation (Explainable); Path Reasoning	nDCG; Recall; Hits@K; Precision
FD-PORT (Flow-guided Direct Preference Optimization for knowledge graph reasoning with trees) (Shen et al., 2025)	WebQSP; ComplexWebQuestions; MetaQA	Knowledge Graph Question Answering; Multi-hop Reasoning	Hits@K
GNNQ (Pflueger et al., 2022)	WatDiv; FB15k-237	Inductive Complex Query Answering; Node Classification	Precision; Recall; Average Precision
InductiveQE (Galkin et al., 2022)	FB15k-237; FB15k; NELL-995; OGBL-WIKIKG2	Inductive Complex Query Answering	Hits@K; AUC-ROC
I-SBR (Interpretable Semantic Based Regularization); I-DCR (Interpretable Deep Concept Reasoners) (Ontiveros et al., 2026)	Countries; Family; WN18RR	Link Prediction (Explainable)	MRR; Hits@K; Coherence
Commonsense reasoner (relation prediction) (Moghimifar et al., 2021)	ATOMIC; ConceptNet-100K	Link Prediction; Multi-hop Reasoning; Rule-Path Reasoning	MRR; Hits@K

System	Dataset(s)	Task(s)	Metric(s)
NS-KAG (Rajalakshmi et al., 2025)	ADNI; OASIS	Classification	Accuracy; Precision; Recall; F1; AUC-ROC; Explainability Index; Symbolic Consistency Rate
NSQA (Kapanipathi et al., 2020)	QALD-9; LC-QuAD 1.0; DBpedia	Knowledge Graph Question Answering; Complex Query Answering	Precision; Recall; F1 (QALD)

Evaluation Methodology (RQ 2)

Datasets (RQ 2.1). Across the included studies, we identified 83 ontology- and knowledge graph–based benchmark datasets used to assess neurosymbolic reasoners (Table 8 and Table 11 in the Appendix). The dataset landscape is heterogeneous, spanning (1) general-purpose KGs widely used in neural link prediction research (e.g., Freebase- and WordNet-derived benchmarks), (2) domain KGs and ontologies, particularly in biomedicine (e.g., GO, HPO, SNOMED CT, and UMLS), (3) Semantic Web/Description Logic benchmarks and ontology-focused resources (e.g., LUBM, ORE, and OWL2Bench), and (4) KGQA benchmarks, where evaluation is mediated through question answering over an underlying KG. Overall, these findings suggest that neurosymbolic evaluation currently draws on a mix of historically embedding-oriented KG benchmarks and ontology-centric resources with stronger formal semantics, rather than a single standardized benchmark suite.

We split the dataset catalogue into two tables: (1) datasets used by at least two neurosymbolic systems reported in the included literature (Table 8), and (2) datasets used by exactly one system (Table 11). This separation distinguishes datasets that serve as community reference points from those reflecting bespoke, domain- or system-specific evaluation choices. Datasets appearing in multiple studies are more likely to support cross-paper comparability, whereas single-use datasets often reflect specialized task formulations or the need to validate capabilities in narrow domains.

The most recurrent benchmarks cluster around a small set of well-established dataset families, including Freebase-derived KG completion resources (e.g., FB15k and FB15k-237), WordNet-derived resources (e.g., WN18 and WN18RR), and widely reused biomedical KGs/ontologies (e.g., GO, HPO, UMLS, protein–protein interaction resources such as STRING, and integrated resources such as GO+STRING). These datasets are widely adopted due to their availability, strong citation footprint, and established experimental protocols, making them convenient for evaluating neurosymbolic pipelines that combine representation learning with relational inference. At the same time, many of these benchmarks were originally developed for link prediction/KG completion and may only partially capture ontology-oriented reasoning phenomena governed by explicit semantic constraints (e.g., disjointness, subsumption, and consistency).

At the same time, the reviewed studies also rely on ontology-centric evaluation resources, including OWL benchmark suites and reasoning datasets (e.g., LUBM, ORE, and OWL2Bench). These resources more directly test deductive reasoning behaviors such as subsumption and consistency. The co-existence of KG-style benchmarks and OWL-centric resources suggests that current neurosymbolic evaluation spans two partially distinct traditions: one emphasizes performance on predictive KG tasks, whereas the other emphasizes deductive reasoning.

Mapping datasets to tasks shows that neurosymbolic systems are evaluated across a broad spectrum of reasoning problems. For KG benchmarks, evaluations most frequently target link prediction/KG completion, often complemented by multi-hop or path reasoning, rule learning/injection, and complex query answering. Ontology-centric datasets support tasks that more directly reflect symbolic inference, including subsumption, consistency checking, ontology completion, and entailment-style evaluations. KGQA datasets (e.g., the LC-QuAD and QALD families, WebQSP, ComplexWebQuestions, MetaQA, and KQA Pro) assess the ability to answer questions grounded in a KG, typically through semantic parsing or end-to-end question answering pipelines.

Importantly, these task families differ in what they measure: predictive KG tasks often assess statistical generalization over graph patterns, whereas ontology reasoning tasks emphasize semantic faithfulness to axioms and logical constraints. The way datasets are combined with tasks therefore shapes what is considered “reasoning” in neurosymbolic evaluation.

Table 8. Ontology and knowledge graphs benchmark datasets used by at least two reported neurosymbolic systems.

Dataset	**Domain**	**Task(s)**
FB15k-237	General-purpose knowledge graph benchmark (Freebase subset, filtered)	Complex Query Answering (Negation) (Huang et al., 2022), Complex Query Answering (Galkin et al., 2024; Zhang et al., 2024; He et al., 2025; Kim et al., 2024; Liu et al., 2021a; Xu et al., 2022; Ren et al., 2020), Inductive Complex Query Answering (Pflueger et al., 2022; Galkin et al., 2022), Link Prediction (Sen et al., 2021; Li et al., 2023, 2025; Khojasteh et al., 2023; Zhang et al., 2019; Liu et al., 2024), Multi-hop Reasoning (Kim et al., 2024; Liu et al., 2024; Huang et al., 2022), Node Classification (Pflueger et al., 2022), Path Reasoning (Liu et al., 2024), Rule Learning (Sen et al., 2021; Li et al., 2023; Zhang et al., 2019)

Dataset	**Domain**	**Task(s)**
Family	Family ontology/relations KG	Complex Query Answering (Wu and Zhao, 2025), Concept Learning Support (Kamdem Teyou et al., 2025), Instance Retrieval (Kamdem Teyou et al., 2025), Link Prediction (Explainable) (Ontiveros et al., 2026), Link Prediction (Zhu et al., 2023; Li et al., 2023, 2025; Han et al., 2023), Masked ABox Revision (Wu and Zhao, 2025), Membership (Zhu et al., 2023), Robust Reasoning under Incompleteness/Inconsistency (Kamdem Teyou et al., 2025), Rule Learning (Li et al., 2023), Subsumption (Zhu et al., 2023)
FB15k	General-purpose knowledge graph benchmark (Freebase subset)	Complex Query Answering (Galkin et al., 2024; Zhang et al., 2024; He et al., 2025; Liu et al., 2021a; Ren et al., 2020), Inductive Complex Query Answering (Galkin et al., 2022), Link Prediction (Zhang et al., 2019), Rule Learning (Zhang et al., 2019)
Gene Ontology (GO)	Gene function ontology	Approximating Deductive Reasoning (Jackermeier et al., 2024), Consistency Checking (Zhapa-Camacho and Hoehndorf, 2023), Drug Repurposing (Alshahrani et al., 2017), Equivalence (Peng et al., 2022), Link Prediction (PPI) (Zhapa-Camacho and Hoehndorf, 2023; Peng et al., 2022), Link Prediction (Jackermeier et al., 2024; Alshahrani et al., 2017; Yang et al., 2025), Membership (Chen et al., 2021a), Ontology Completion (Zhapa-Camacho and Hoehndorf, 2023; Chen et al., 2021a), Subsumption (Zhapa-Camacho and Hoehndorf, 2023; Jackermeier et al., 2024; Mondal et al., 2021; Chen et al., 2021a; Yang et al., 2025)

Dataset	**Domain**	**Task(s)**
NELL-995	NELL subset knowledge graph benchmark	Complex Query Answering (Negation) (Huang et al., 2022), Complex Query Answering (Galkin et al., 2024; Zhang et al., 2024; Xu et al., 2022; Ren et al., 2020), Inductive Complex Query Answering (Galkin et al., 2022), Link Prediction (Liu et al., 2024), Multi-hop Reasoning (Liu et al., 2024; Huang et al., 2022), Path Reasoning (Liu et al., 2024)
WN18RR	WordNet knowledge graph benchmark (lexical-semantic, revised)	Complex Query Answering (Negation) (Huang et al., 2022), Link Prediction (Explainable) (Ontiveros et al., 2026), Link Prediction (Sen et al., 2021; Li et al., 2025; Shengyuan et al., 2023; Zhang et al., 2019; Liu et al., 2024), Multi-hop Reasoning (Liu et al., 2024; Huang et al., 2022), Path Reasoning (Liu et al., 2024), Probabilistic Logic Reasoning (Shengyuan et al., 2023), Rule Learning (Sen et al., 2021; Zhang et al., 2019)
FoodOn	Food ontology	Consistency Checking (Zhapa-Camacho and Hoehndorf, 2023), Link Prediction (PPI) (Zhapa-Camacho and Hoehndorf, 2023; Mashkova et al., 2026), Membership (Chen et al., 2021a), Ontology Completion (Zhapa-Camacho and Hoehndorf, 2023; Chen et al., 2021a), Subsumption (Zhapa-Camacho and Hoehndorf, 2023; Chen et al., 2023; Mashkova et al., 2026; Chen et al., 2021a)
GALEN	Clinical ontology (anatomy/medicine)	Approximating Deductive Reasoning (Jackermeier et al., 2024), Link Prediction (PPI) (Mashkova et al., 2026), Link Prediction (Jackermeier et al., 2024; Yang et al., 2025), Subsumption (Mashkova et al., 2026; Jackermeier et al., 2024; Mondal et al., 2021; Yang et al., 2025)

Dataset	**Domain**	**Task(s)**
GO+STRING	Integrated biomedical KG (GO with protein-interaction context)	Approximating Deductive Reasoning (Jackermeier et al., 2024), Link Prediction (PPI) (Mashkova et al., 2026; Kulmanov et al., 2019; Mashkova et al., 2024), Link Prediction (Jackermeier et al., 2024; Mashkova et al., 2024), Subsumption (Mashkova et al., 2026; Jackermeier et al., 2024)
Kinship	Kinship relations knowledge graph benchmark	Link Prediction (Sen et al., 2021; Li et al., 2023, 2025; Shengyuan et al., 2023), Probabilistic Logic Reasoning (Shengyuan et al., 2023), Rule Learning (Sen et al., 2021; Li et al., 2023)
UMLS	Biomedical metathesaurus (ontology/KG resource)	Link Prediction (Sen et al., 2021; Li et al., 2023, 2025), Membership (Hohenecker and Lukasiewicz, 2020), Rule Learning (Sen et al., 2021; Li et al., 2023)
Anatomy	Biomedical anatomy ontology	Approximating Deductive Reasoning (Jackermeier et al., 2024), Link Prediction (Jackermeier et al., 2024; Yang et al., 2025), Subsumption (Jackermeier et al., 2024; Mondal et al., 2021; Yang et al., 2025)
HPO	Human phenotype ontology	Approximate Query Answering (Zhurov et al., 2025), Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017), Semantic Search (Zhurov et al., 2025)
LC-QuAD 1.0	Knowledge-graph question answering (DBpedia)	Complex Query Answering (Kapanipathi et al., 2020), Knowledge Graph Question Answering (Gomes et al., 2022; Vollmers et al., 2021; Kapanipathi et al., 2020), Semantic Parsing (Gomes et al., 2022; Vollmers et al., 2021), Template Classification (Gomes et al., 2022; Vollmers et al., 2021)
NELL	General-purpose knowledge graph (NELL)	Complex Query Answering (He et al., 2025; Kim et al., 2024; Liu et al., 2021a), Multi-hop Reasoning (Kim et al., 2024)

Dataset	**Domain**	**Task(s)**
WebQSP	Knowledge-graph question answering (Freebase)	Knowledge Graph Question Answering (Gomes et al., 2022; Agarwal and Bedathur, 2025; Shen et al., 2025), Multi-hop Reasoning (Shen et al., 2025), Semantic Parsing (Gomes et al., 2022), Template Classification (Gomes et al., 2022)
WN18	WordNet knowledge graph benchmark (lexical-semantic)	Link Prediction (Li et al., 2023; Shengyuan et al., 2023; Zhang et al., 2019), Probabilistic Logic Reasoning (Shengyuan et al., 2023), Rule Learning (Li et al., 2023; Zhang et al., 2019)
YAGO3-10	General-purpose knowledge graph benchmark (YAGO subset)	Link Prediction (Abboud et al., 2020; Purohit et al., 2025a; Shengyuan et al., 2023), Probabilistic Logic Reasoning (Shengyuan et al., 2023), Rule Injection (Abboud et al., 2020)
ComplexWebQuestions	Knowledge-graph question answering (Freebase)	Knowledge Graph Question Answering (Gomes et al., 2022; Shen et al., 2025), Multi-hop Reasoning (Shen et al., 2025), Semantic Parsing (Gomes et al., 2022), Template Classification (Gomes et al., 2022)
Countries	Geopolitical knowledge graph	Link Prediction (Explainable) (Ontiveros et al., 2026), Membership (Hohenecker and Lukasiewicz, 2020)
DBpedia	General-purpose knowledge graph (Wikipedia-derived)	Complex Query Answering (Kapanipathi et al., 2020), Knowledge Graph Question Answering (Kapanipathi et al., 2020), Membership (Hohenecker and Lukasiewicz, 2020)
HeLiS	Health and lifestyle ontology/knowledge graph	Membership (Chen et al., 2021a), Ontology Completion (Chen et al., 2021a), Subsumption (Chen et al., 2023, 2021a)
LC-QuAD 2.0	Knowledge-graph question answering (DBpedia/Wikidata-based)	Knowledge Graph Question Answering (Gomes et al., 2022; Vollmers et al., 2021), Semantic Parsing (Gomes et al., 2022; Vollmers et al., 2021), Template Classification (Gomes et al., 2022; Vollmers et al., 2021)
MetaQA	Knowledge-graph question answering (movie domain)	Knowledge Graph Question Answering (Agarwal and Bedathur, 2025; Shen et al., 2025), Multi-hop Reasoning (Shen et al., 2025)

Dataset	Domain	Task(s)
OAEI Bio-ML 2023	Biomedical ontology matching benchmark (OAEI)	Ontology Alignment (Sharma and Jain, 2025; Teymurova et al., 2024)
QALD-9	Knowledge-graph question answering/semantic parsing benchmark	Complex Query Answering (Kapanipathi et al., 2020), Knowledge Graph Question Answering (Vollmers et al., 2021; Kapanipathi et al., 2020), Semantic Parsing (Vollmers et al., 2021), Template Classification (Vollmers et al., 2021)

Tasks and Metrics (RQ 2.2). To characterize evaluation practice in the reviewed literature, we extracted all reported evaluation metrics and computed their frequencies across studies. We then report the ten most frequently used metrics in ontology- and knowledge-graph-centric evaluations of neurosymbolic reasoning systems:

1. **Hits@K:** the proportion of queries for which at least one correct answer is ranked within the top K predictions.
2. **Mean Reciprocal Rank (MRR):** the mean of the reciprocal rank of the first correct prediction for each query, i.e., $\frac{1}{|Q|}\sum_{q\in Q}\frac{1}{\text{rank}_q}$.
3. **Recall:** the proportion of relevant items that are retrieved, $\frac{TP}{TP+FN}$.
4. **Precision:** the proportion of retrieved items that are relevant, $\frac{TP}{TP+FP}$.
5. **AUC-ROC:** the area under the receiver operating characteristic curve, summarizing discrimination performance across all classification thresholds.
6. **F1 score:** the harmonic mean of precision and recall, $2\cdot\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$.
7. **Mean Rank (MR):** the average rank of the correct answer(s) in the predicted ranking across queries (lower is better).
8. **Accuracy:** the proportion of correctly predicted instances, $\frac{TP+TN}{TP+TN+FP+FN}$.
9. **Runtime:** time required to perform inference/reasoning on a given benchmark.
10. **Normalized Discounted Cumulative Gain (nDCG):** a ranking-quality measure that discounts lower-ranked correct items and normalizes by the ideal ranking, commonly used in recommendation-style evaluations.

Beyond the most commonly reported metrics, the reviewed studies showed a clear “long tail” of largely study-specific measures: many appeared in only a single paper. These included alternative ranking and retrieval metrics (e.g., Median Rank, MAP, Average Precision, Percentile Rank, Overlap@K, and MAR) as well as benchmark-specific variants such as F1 for QALD-style question answering. Several papers proposed rule- and ontology-focused quality indicators such as High Quality Rules (count/%), Head Coverage, Ontology Match Rate, Ontology Extension Accuracy, Coherence, and Subsumption Violation Rate, aiming to capture logical validity. Fewer studies highlighted symbolic constraints and faithfulness (Constraint Satisfaction Rate, Symbolic Consistency Rate) or interpretability (Explainability Index). Others quantified distributional similarity and error using KL Divergence, Jaccard Similarity, similarity–distance curves, and regression losses (MSE, MAE, MAPE), alongside broader diversity/novelty measures and coarse end-to-end outcomes like Success Rate.

Across the included studies (Table 9), evaluation concentrated on a small set of ontology- and KG-centric reasoning tasks with a well-defined neural–symbolic division of responsibilities. Link prediction (KG completion) was the most common setting, typically implemented as neural triple scoring (via embeddings or GNNs) complemented by symbolic schema or rules used as constraints—either to regularize training or to filter invalid predictions (G2 Representation learning with logic). A second cluster addressed structured reasoning over learned representations (G3 Structured reasoning over learned representations), including complex query answering and multi-hop reasoning: neural components learned query operators or traversal policies, while symbolic components provided query semantics and constraint checking. Ontology-centric structure induction was also prominent. In subsumption and rule learning (G4 Symbolic structure induction), neural models ranked candidate axioms or rules, and symbolic semantics determined correctness through coherence and violation checks. Natural-language KGQA appeared as a grounding setting (G1 Grounding & parsing), combining neural entity/relation linking and logical-form prediction with symbolic execution (e.g., SPARQL or logic) to ensure grounded, constraint-consistent answers. Overall, ranking metrics (Hits@K, MRR, MAP, rank-based statistics) dominated KG completion and multi-hop reasoning, whereas ontology-centric tasks additionally employed classification metrics (accuracy/F1) and, less frequently, explicit constraint/coherence measures.

To complement the dominant tasks, Table 12 (Appendix) summarizes infrequently used evaluation settings observed only once or twice across the included studies. These capture the "long tail" of neurosymbolic evaluation in ontology/KG-centric reasoning and span two additional meta-groups: search and planning with neural guidance (G5) and explainable inference (G6). We include this mapping in the appendix because these task types were least commonly adopted in the reviewed neurosymbolic reasoners, but they remain informative for characterizing emerging or niche evaluation practices.

Table 9. Task-to-metric mapping.

Task	**Supported metric(s)**	**Neuro role**	**Symbolic role**
(G1) Grounding & parsing			

Task	**Supported metric(s)**	**Neuro role**	**Symbolic role**
KG Question Answering (KGQA): answer NL questions by grounding to KG entities/relations and executing queries	Accuracy; F1; F1 (QALD); Hits@K; Precision; Recall; Runtime	Map NL to entities/relations/logical form (semantic parsing + entity/relation linking); rerank candidates	Execute SPARQL/logic; enforce grounded semantics and constraints (types, ontology consistency); sometimes repair/validate generated queries
(G2) Representation learning with logic			
Link prediction (KG completion): predict missing or plausible triples	AUC-ROC; F1; Head Coverage; High Quality Rules (%); High Quality Rules (count); Hits@K; MAP; MRR; Mean Absolute Error; Mean Rank; Mean Squared Error; Median Rank; Precision; Recall; Runtime; nDCG	Learn a scoring/ranking function over candidate triples (KG embeddings, GNN encoders, neural scorers)	Inject ontology schema and rules as constraints (type/domain/range, disjointness), filter invalid predictions, or regularize training toward logical consistency
(G3) Structured reasoning over learned representations			
Complex Query Answering: answer structured KG queries (often multi-hop; sometimes with operators like conjunction; occasionally negation)	AUC-ROC; Average Percentile Rank (APR); F1 (QALD); Hits@K; KL Divergence; MRR; Mean Absolute Percentage Error; Precision; Recall; Runtime; Success Rate	Learn query embeddings/differentiable query operators; retrieve and rank answers efficiently	Provide query semantics (operators), enforce type constraints, and validate outputs against ontology/schema (sometimes also used to decompose/execute queries)
Multi-hop Reasoning: infer answers through relational paths across the KG	Hits@K; MAP; MRR; Runtime	Learn traversal/attention over neighbors and composition over hops; score paths/targets	Constrain allowed hops using schema/rules; validate path consistency (typed paths, rule-compliant paths)

Task	Supported metric(s)	Neuro role	Symbolic role
Membership (realization): decide if an individual belongs to a class given ontology definitions and constraints	Accuracy; F1; Hits@K; MRR; Precision; Recall; Runtime	Score instance-of/membership predictions (often embedding-based)	Membership is defined by ontology axioms/rules; check entailment/consistency and reject logically invalid assignments
(G4) Symbolic structure induction			
Subsumption (classification): decide whether one class is a subclass of another given ontology definitions and constraints	90th Percentile Rank; AUC-ROC; Accuracy; F1; Hits@K; MRR; Mean Rank; Median Rank; Percentile Rank; Precision; Recall; Runtime	Predict subsumption likelihood via embeddings or neural entailment scoring	DL semantics defines correctness; coherence/violation checks (e.g., penalize subsumption violations, maintain hierarchy consistency)
Rule learning: induce explicit symbolic rules from KG data (often Horn rules)	Head Coverage; High Quality Rules (%); High Quality Rules (count); Hits@K; MRR; Runtime	Propose candidate rules or score rule bodies (neural-guided search, embedding support)	Output explicit rules; evaluate support/confidence and enforce logical form; optionally validate against ontology constraints

Discussion

Limitations and future improvements (RQ 3)

Existing ontology- and KG-based datasets used to evaluate neurosymbolic reasoners exhibit several limitations. First, the benchmark landscape is fragmented across two partially distinct traditions: widely adopted KG completion datasets (e.g., Freebase- and WordNet-derived benchmarks) and ontology-centric OWL/DL benchmarks (e.g., LUBM, ORE, and OWL2Bench). This fragmentation limits cross-study comparability and weakens claims of general progress when evaluations are restricted to a single benchmark family. Second, many commonly used KG benchmarks were not designed to test formal logical constraints and axioms, which underrepresent ontology-oriented reasoning phenomena such as subsumption, disjointness, and consistency. Third, KGQA benchmarks operationalize reasoning through end-to-end question answering pipelines, which makes it difficult to isolate the contribution of the reasoning component, because errors in semantic parsing, entity/relation linking, or evidence retrieval can dominate end-to-end accuracy. Table 10 summarizes these limitations and their consequences.

Table 10. Benchmark limitations and their consequences.

Limitation	**Consequence**
Fragmentation across "predictive KG" vs. "ontology reasoning" benchmark traditions (e.g., FB15k/WN18RR/OGBL-WIKIKG2 vs. LUBM/ORE/OWL2Bench)	Limited cross-study comparability; gains on one benchmark family may not translate to the other
Dominance of established KG completion benchmarks (e.g., FB15k/FB15k-237, WN18/WN18RR, NELL/NELL-995)	Evaluation may primarily reflect graph-pattern generalization rather than ontological reasoning
Limited coverage of explicit axioms/constraints in common KGs (e.g., Freebase/WordNet subsets)	Difficult to assess semantic faithfulness (e.g., soundness with respect to subsumption, disjointness, or consistency constraints)
Ontology-centric benchmarks often emphasize a narrow set of DL services (commonly subsumption/consistency) (e.g., LUBM, SNOMED CT, GO, HPO, GALEN)	Systems may appear strong on a limited set of services while lacking broader neurosymbolic capabilities
KGQA datasets conflate reasoning with upstream components (e.g., LC-QuAD, QALD, WebQSP, ComplexWebQuestions, MetaQA, KQA Pro)	End-to-end scores obscure whether failures arise from reasoning, parsing, linking, or evidence retrieval
Uneven operator coverage in complex query benchmarks (e.g., negation appears more often than recursion, quantification, or temporal operators)	Reasoners may be under-tested on logical phenomena central to symbolic reasoning
Synthetic benchmarks and generators (e.g., OWL2Bench) may not reflect real-world noise and heterogeneity	Risk of overestimating performance on clean, templated data
Non-standardized evaluation protocols (e.g., splits, negative sampling, filtered metrics, timeouts)	Reduced reproducibility and comparability even when the same dataset is used

These evaluation settings rarely assess the integrated behavior of neural and symbolic components, nor do they systematically measure reasoning depth, the ability to compose learned reasoning patterns to handle novel combinations of entities and relations (compositionality), or the ability to generalize to data that differs systematically from the training distribution, including zero-shot and out-of-distribution scenarios.

To address these limitations, we propose a five-dimensional evaluation framework, as illustrated in Figure 3, encompassing:

1. **Dataset Characteristics:** Evaluates whether datasets are suitable for reasoning, including logical depth (multi-hop inference chains), ontology expressivity (RDF/OWL), open- vs closed-world assumptions, presence

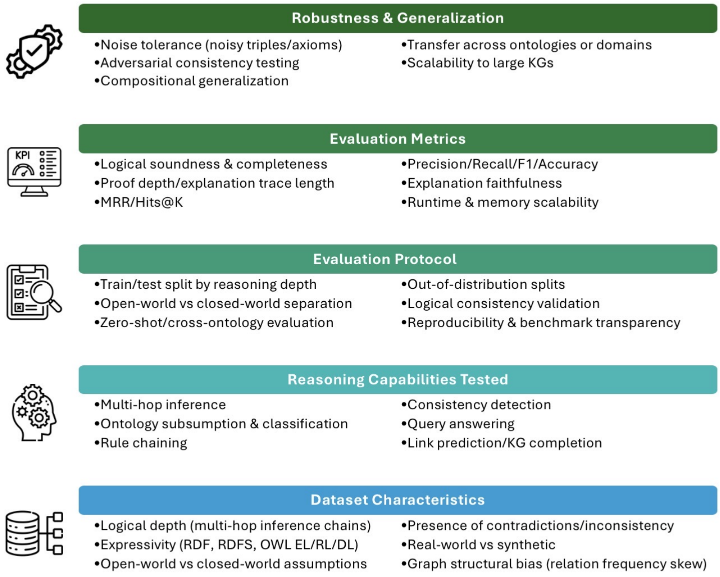
Robustness & Generalization
•Noise tolerance (noisy triples/axioms)
•Adversarial consistency testing
•Compositional generalization
•Transfer across ontologies or domains
•Scalability to large KGs
Evaluation Metrics
KPI
•Logical soundness & completeness
•Proof depth/explanation trace length
•MRR/Hits@K
•Precision/Recall/F1/Accuracy
•Explanation faithfulness
•Runtime & memory scalability
Evaluation Protocol
•Train/test split by reasoning depth
•Open-world vs closed-world separation
•Zero-shot/cross-ontology evaluation
•Out-of-distribution splits
•Logical consistency validation
•Reproducibility & benchmark transparency
Reasoning Capabilities Tested
•Multi-hop inference
•Ontology subsumption & classification
•Rule chaining
•Consistency detection
•Query answering
•Link prediction/KG completion
Dataset Characteristics
•Logical depth (multi-hop inference chains)
•Expressivity (RDF, RDFS, OWL EL/RL/DL)
•Open-world vs closed-world assumptions
•Presence of contradictions/inconsistency
•Real-world vs synthetic
•Graph structural bias (relation frequency skew)

Figure 3. A five-dimensional framework for evaluating ontology- and knowledge graph–based neurosymbolic reasoners.

of contradictions, and real-world vs synthetic datasets. Graph structural biases, such as relation frequency skew, hub-node degree imbalance, and shortcut pattern regularities, are also assessed to ensure that evaluation is not dominated by memorization or shortcuts.

2. **Reasoning Capabilities Tested:** Specifies the types of reasoning evaluated, including multi-hop inference, ontology subsumption and classification, rule chaining, logical entailment generation, consistency detection, query answering, and link prediction/KG completion. This ensures that hybrid systems are assessed across the full spectrum of symbolic and neural reasoning capabilities.
3. **Evaluation Protocol:** Describes how experiments are structured to enable fair and reproducible evaluation. This includes depth-aware train/test splits, open- vs closed-world evaluation separation, zero-shot or cross-ontology evaluation, out-of-distribution splits, logical consistency validation during evaluation, and reproducibility and benchmark transparency.
4. **Evaluation Metrics:** Measures both symbolic correctness and statistical performance, including logical soundness and completeness, proof depth or explanation trace length, MRR/Hits@K, precision/recall/F1/accuracy, explanation faithfulness, and runtime/memory scalability.

5. **Robustness & Generalization:** Assesses stability under challenging conditions, including noise tolerance (noisy triples or axioms), adversarial consistency testing, compositional generalization stress tests, transfer across ontologies or domains, and scalability to large knowledge graphs.

By structuring evaluation across these five dimensions, the framework enables systematic comparison of ontology- and knowledge graph–based neurosymbolic systems and provides guidance for future benchmark development.

Conclusion

We conclude by revisiting the research questions that guided this scoping review. The following subsections synthesize the key findings for each question and outline the main takeaways.

RQ 1: What neurosymbolic reasoning systems have been developed for reasoning over ontologies and knowledge graphs?

In addressing RQ 1, we identified 63 neurosymbolic reasoning systems and frameworks developed for reasoning over ontologies and knowledge graphs. Mapped to Kautz's taxonomy, these systems span all six neuro-symbolic integration paradigms and primarily target link prediction/knowledge graph completion and complex query answering, with a smaller subset focusing on ontology-centric reasoning such as subsumption, ontology completion, and alignment. While this corpus demonstrates a broad design space of approaches, evaluation remains fragmented across datasets, tasks, and metrics: most studies report predictive and ranking measures (e.g., Accuracy, F1, MRR, Hits@K), whereas fewer operationalize logical soundness (e.g., constraint satisfaction, consistency/coherence, proof/path quality) or efficiency beyond runtime, and dataset usage ranges from standard KGC benchmarks (FB15k/FB15k-237, WN18/WN18RR) and biomedical resources (GO, HPO, UMLS, STRING/GO+STRING) to ontology reasoning (LUBM, ORE, OWL2Bench) and KGQA benchmarks (LC-QuAD/QALD, WebQSP, ComplexWebQuestions, MetaQA, KQA Pro), limiting direct comparability across integration types.

RQ 2: How are these systems evaluated?

RQ 2.1: Which ontology- and knowledge graph-based benchmark datasets have been used to assess neurosymbolic reasoners? From the 63 included neurosymbolic reasoners, we identified 83 ontology- and knowledge graph-based benchmark datasets used for evaluation. The most frequently used are general-purpose knowledge graph completion benchmarks derived from Freebase and WordNet (e.g., FB15k/FB15k-237 and WN18/WN18RR), alongside biomedical ontologies and interaction graphs (e.g., GO, HPO, UMLS, STRING, and integrated resources such as GO+STRING). Additional studies evaluate on ontology reasoning benchmarks (e.g., LUBM, ORE, OWL2Bench) or knowledge graph question answering benchmarks (e.g., LC-QuAD and QALD families, WebQSP, ComplexWebQuestions, MetaQA, KQA Pro) for end-to-end QA grounded in KGs.

Despite this breadth, benchmark usage is fragmented across tasks and communities, and only a small subset of datasets is reused across systems, limiting reproducibility and direct comparison. As a result, current evaluations rarely jointly capture (1) predictive performance on KG benchmarks, (2) semantic faithfulness to ontological constraints, and (3) reasoning behavior expressed through complex queries or QA. By cataloguing datasets and linking them to the tasks used in neurosymbolic evaluation, this review provides an empirical basis for developing more standardized benchmarks that integrate formal semantics, logical constraints, and learning-based generalization.

RQ 2.2: What ontology- and knowledge graph-centric reasoning tasks and evaluation metrics are applied to assess neurosymbolic reasoning systems? The task–to-metric mapping indicates that evaluation concentrates on a small core of ontology/KG-centric tasks. KG completion/link prediction (G2) and multi-hop/complex query reasoning (G3) dominate, and are assessed primarily with ranking metrics (e.g., Hits@K, MRR, MAP, and rank statistics). Ontology induction tasks such as subsumption and rule learning (G4) more often incorporate classification metrics (e.g., accuracy, precision/recall, F1) and, less frequently, semantic validity checks (e.g., coherence, constraint-violation rates). Across tasks, measures that directly capture logical properties (e.g., soundness/consistency, proof or explanation quality, and robustness to constraint satisfaction) remain comparatively rare.

RQ 3: What limitations do existing ontology- and knowledge graph–based datasets present, and how can evaluation methodologies in neurosymbolic reasoning be improved?

Our synthesis highlights three main limitations of current ontology- and KG-based evaluation. First, evaluation remains fragmented across two partially distinct traditions: predictive KG benchmarks (e.g., link prediction/KG completion on large KGs) and ontology-centric OWL/DL benchmarks targeting deductive reasoning (e.g., subsumption and consistency). This split constrains cross-study comparability and weakens claims about overall progress when systems are assessed within only one benchmark family. Second, many widely used KG benchmarks were not designed to probe explicit axioms and logical constraints, and thus underrepresent ontology-oriented reasoning (e.g., subsumption, disjointness, and consistency). Third, KGQA benchmarks operationalize reasoning via end-to-end pipelines and can conflate reasoning performance with upstream components (e.g., semantic parsing, entity/relation linking, and evidence retrieval), complicating attribution of gains to the reasoner itself.

To address these limitations, we advocate for a systematic, multi-dimensional evaluation approach that aligns dataset design, experimental protocols, and metrics. Concretely, the five-dimensional framework proposed in this review (e.g., covering dataset characteristics, reasoning capabilities, evaluation protocols, evaluation metrics, and robustness/generalization) provides a structured basis for designing and comparing benchmarks in a way that jointly captures predictive performance, semantic faithfulness to ontological constraints, and reasoning behavior under distribution shift and noise. Adopting such standardized evaluation practices would improve reproducibility, enable more meaningful comparisons across neuro-symbolic integration paradigms, and support the development of benchmarks that better reflect the goals of neurosymbolic reasoning over ontologies and knowledge graphs.

References

Abboud R, Ceylan I, Lukasiewicz T and Salvatori T (2020) Boxe: A box embedding model for knowledge base completion. In: Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds.) *Advances in Neural Information Processing Systems*, volume 33. Curran Associates, Inc., pp. 9649–9661. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6dbbe6abe5f14af882ff977fc3f35501-Paper.pdf.

Agarwal P and Bedathur S (2025) A zero-shot neuro-symbolic approach for complex knowledge graph question answering. In: Christodoulopoulos C, Chakraborty T, Rose C and Peng V (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7, pp. 11514–11527. DOI:10.18653/v1/2025.findings-emnlp.617. URL https://aclanthology.org/2025.findings-emnlp.617/.

Alshahrani M, Khan MA, Maddouri O, Kinjo AR, Queralt-Rosinach N and Hoehndorf R (2017) Neuro-symbolic representation learning on biological knowledge graphs. *Bioinformatics* 33(17): 2723–2730. DOI:10.1093/bioinformatics/btx275. URL https://doi.org/10.1093/bioinformatics/btx275.

Arksey H and O'Malley L (2005) Scoping studies: towards a methodological framework. *International Journal of Social Research Methodology* 8(1): 19–32.

Baader F, Calvanese D, Mcguinness D, Nardi D and Patel-Schneider P (2007) *The Description Logic Handbook: Theory, Implementation, and Applications.*

Besold TR, d'Avila Garcez AS, Bader S, Bowman H, Domingos PM, Hitzler P, Kühnberger K, Lamb LC, Lowd D, Lima PMV, de Penning L, Pinkas G, Poon H and Zaverucha G (2017a) Neural-symbolic learning and reasoning: A survey and interpretation. *CoRR* abs/1711.03902. URL http://arxiv.org/abs/1711.03902.

Besold TR, Garcez Ad, Bader S, Bowman H, Domingos P, Hitzler P, Kühnberger KU, Lamb LC, de Penning L, Pinkas G et al. (2017b) Neural-symbolic learning and reasoning: A survey and interpretation. *Frontiers in Artificial Intelligence and Applications* 304: 1–54.

Bhuyan BP, Singh TP, Tomar R and Ramdane-Cherif A (2024) Nesykhg: Neuro-symbolic knowledge hypergraphs. *Procedia Computer Science* 235: 1278–1288. DOI:https://doi.org/10.1016/j.procs.2024.04.121. URL https://www.sciencedirect.com/science/article/pii/S187705092400797X. International Conference on Machine Learning and Data Engineering (ICMLDE 2023).

Bordes A, Usunier N, Garcia-Duran A, Weston J and Yakhnenko O (2013) Translating embeddings for modeling multi-relational data. In: Burges C, Bottou L, Welling M, Ghahramani Z and Weinberger K (eds.) *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca8133fa24680a88d2f9-Paper.pdf.

BOUGZIME O, Jabbar S, Cruz C and Demoly F (2025) Evaluating neuro-symbolic ai architectures: Design principles, qualitative benchmark, comparative analysis and results. In: H Gilpin L, Giunchiglia E, Hitzler P and van Krieken E (eds.) *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, *Proceedings of Machine Learning Research*, volume 284. PMLR, pp. 1119–1143. URL https://proceedings.mlr.press/v284/bougzime25a.html.

Boulakbech M and Wannous R (2025) Context-Aware Hybrid Neuro-Symbolic Approach for Knowledge Graph Enrichment. In: *WISE 2025*. Marrakesh, Morocco. URL https://univ-rochelle.hal.science/hal-05343251.

Bramer WM, Rethlefsen ML, Kleijnen J and Franco OH (2017) Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Systematic Reviews* 6(1): 245. DOI:10.1186/s13643-017-0644-y. URL https://pubmed.ncbi.nlm.nih.gov/29208034/.

Chen J, He Y, Geng Y, Jimenez-Ruiz E, Dong H and Horrocks I (2023) Contextual semantic embeddings for ontology subsumption prediction. URL https://arxiv.org/abs/2202.09791.

Chen J, Hu P, Jimenez-Ruiz E, Holter OM, Antonyrajah D and Horrocks I (2021a) Owl2vec*: Embedding of owl ontologies. URL https://arxiv.org/abs/2009.14654.

Chen S, Yuan Z, Zhang Q, Hua W, Cao J and Huang X (2025) Neuro-symbolic entity alignment via variational inference. URL https://arxiv.org/abs/2410.04153.

Chen X, Boratko M, Chen M, Dasgupta SS, Li XL and McCallum A (2021b) Probabilistic box embeddings for uncertain knowledge graph reasoning. In: Toutanova K, Rumshisky A, Zettlemoyer L, Hakkani-Tur D, Beltagy I, Bethard S, Cotterell R, Chakraborty T and Zhou Y (eds.) *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 882–893. DOI: 10.18653/v1/2021.naacl-main.68. URL https://aclanthology.org/2021.naacl-main.68/.

d'Avila Garcez A, Lamb L and Gabbay D (2020) Neuro-symbolic ai: The 3rd wave. *Communications of the ACM* 63(11): 58–66.

Delplanque G, Werner L, Layaïda N and Geneves P (2025) A comparative analysis of neurosymbolic methods for link prediction. In: H Gilpin L, Giunchiglia E, Hitzler P and van Krieken E (eds.) *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, *Proceedings of Machine Learning Research*, volume 284. PMLR, pp. 674–696. URL https://proceedings.mlr.press/v284/delplanque25a.html.

Dettmers T, Minervini P, Stenetorp P and Riedel S (2018) Convolutional 2d knowledge graph embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence* 32(1). DOI:10.1609/aaai.v32i1.11573. URL https://ojs.aaai.org/index.php/AAAI/article/view/11573.

Ebrahimi M, Eberhart A, Bianchi F and Hitzler P (2021a) Towards bridging the neuro-symbolic gap: deep deductive reasoners. *Applied Intelligence* 51(9): 6326–6348. DOI:10.1007/s10489-020-02165-6. URL https://doi.org/10.1007/s10489-020-02165-6.

Ebrahimi M, Sarker MK, Bianchi F, Xie N, Eberhart A, Doran D, Kim H and Hitzler P (2021b) Neuro-symbolic deductive reasoning for cross-knowledge graph entailment. In: *AAAI Spring Symposium Combining Machine Learning with Knowledge Engineering*. URL https://api.semanticscholar.org/CorpusID:231853605.

Galkin M, Zhou J, Ribeiro B, Tang J and Zhu Z (2024) A foundation model for zero-shot logical query reasoning. In: Globerson A, Mackey L, Belgrave D, Fan A, Paquet U, Tomczak J and Zhang C (eds.) *Advances in Neural Information Processing Systems*, volume 37. Curran Associates, Inc., pp. 54137–54160. DOI:10.52202/079017-1715. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/616521c3cf15f9f7018565c427d40e3b-Paper-Conference.pdf.

Galkin M, Zhu Z, Ren H and Tang J (2022) Inductive logical query answering in knowledge graphs. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds.) *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., pp. 15230–15243. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/6246e04dcf42baf7c71e3a65d3d93b55-Paper-Conference.pdf.

Garcez Ad, Lamb LC and Gabbay DM (2019) *Neural-Symbolic Cognitive Reasoning*. Springer.

Glimm B, Horrocks I, Motik B, Stoilos G and Wang Z (2014) Hermit: An owl 2 reasoner. *Journal of Automated Reasoning* 53. DOI:10.1007/s10817-014-9305-1.

Gomes J, de Mello RC, Ströele V and de Souza JF (2022) A hereditary attentive template-based approach for complex knowledge base question answering systems. *Expert Systems with Applications* 205: 117725. DOI:https://doi.org/10.1016/j.eswa.2022.117725. URL https://www.sciencedirect.com/science/article/pii/S0957417422010089.

Gruber TR (1993) A translation approach to portable ontology specifications. *Knowledge acquisition* 5(2): 199–220.

Haddaway NR, Collins AM, Coughlin D and Kirk S (2015) The role of google scholar in evidence reviews and its applicability to grey literature searching. *PLOS ONE* 10(9): e0138237. DOI:10.1371/journal.pone.0138237. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0138237.

Hamilton W, Bajaj P, Zitnik M, Jurafsky D and Leskovec J (2018) Embedding logical queries on knowledge graphs. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N and Garnett R (eds.) *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/ef50c335cca9f340bde656363ebd02fd-Paper.pdf.

Han C, He Q, Yu C, Du X, Tong H and Ji H (2023) Logical entity representation in knowledge-graphs for differentiable rule learning. URL https://arxiv.org/abs/2305.12738.

He Y, Xiong B, Hernández D, Zhu Y, Kharlamov E and Staab S (2025) Dage: Dag query answering via relational combinator with logical constraints. In: *Proceedings of the ACM on Web Conference 2025*, WWW '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712746, p. 2514–2529. DOI:10.1145/3696410.3714677. URL https://doi.org/10.1145/3696410.3714677.

Hitzler P and Sarker M (2022) *Neuro-symbolic Artificial Intelligence: The State of the Art*. Frontiers in artificial intelligence and applications. IOS Press. ISBN 9781643682440. URL https://books.google.nl/books?id=jnL0zgEACAAJ.

Hogan A, Blomqvist E, Cochez M, d'Amato C, de Melo G, Gutierrez C, Gayo JEL, Kirrane S, Neumaier S, Polleres A, Navigli R, Ngomo AN, Rashid SM, Rula A, Schmelzeisen L, Sequeda JF, Staab S and Zimmermann A (2020) Knowledge graphs. *CoRR* abs/2003.02320. URL https://arxiv.org/abs/2003.02320.

Hohenecker P and Lukasiewicz T (2020) Ontology reasoning with deep neural networks. *Journal of Artificial Intelligence Research* 68. DOI:10.1613/jair.1.11661. URL http://dx.doi.org/10.1613/jair.1.11661.

Hua W and Zhang Y (2022) System 1 + system 2 = better world: Neural-symbolic chain of logic reasoning. In: Goldberg Y, Kozareva Z and Zhang Y (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 601–612. DOI:10.18653/v1/2022.findings-emnlp.42. URL https://aclanthology.org/2022.findings-emnlp.42/.

Huang Z, Chiang MF and Lee WC (2022) Line: Logical query reasoning over hierarchical knowledge graphs. In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850, p. 615–625. DOI:10.1145/3534678.3539338. URL https://doi.org/10.1145/3534678.3539338.

Jackermeier M, Chen J and Horrocks I (2024) Dual box embeddings for the description logic el++. In: *Proceedings of the ACM Web Conference 2024*, WWW '24. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701719, p. 2250–2258. DOI:10.1145/3589334.3645648. URL https://doi.org/10.1145/3589334.3645648.

Kamdem Teyou LM, Friedrichs L, Kouagou NJ, Demir C, Mahmood Y, Heindorf S and Ngonga Ngomo AC (2025) Neural reasoning for robust instance retrieval in shoiq. In: *Proceedings of the 13th Knowledge Capture Conference 2025*, K-CAP '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400718670, p. 61–68. DOI:10.1145/3731443.3771348. URL https://doi.org/10.1145/3731443.3771348.

Kapanipathi P, Abdelaziz I, Ravishankar S, Roukos S, Gray AG, Astudillo RF, Chang M, Cornelio C, Dana S, Fokoue A, Garg D, Gliozzo A, Gurajada S, Karanam HP, Khan N, Khandelwal D, suk Lee Y, Li Y, Luus FPS, Makondo N, Mihindukulasooriya N, Naseem T, Neelam S, Popa L, Reddy RG, Riegel R, Rossiello G, Sharma U, Bhargav GPS and Yu M (2020) Question answering over knowledge bases by leveraging semantic parsing and neuro-symbolic reasoning.

ArXiv abs/2012.01707. URL https://api.semanticscholar.org/CorpusID:227253707.

Kautz H (2022) The third ai summer: Aaai robert s. engelmore memorial lecture. *AI Magazine* 43(1): 105–125. DOI:10.1002/aaai.12036. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/19122.

Khojasteh MH, Torabian N, Farjami A, Hosseini S and Minaei-Bidgoli B (2023) Emulating the human mind: A neural-symbolic link prediction model with fast and slow reasoning and filtered rules. URL https://arxiv.org/abs/2310.13996.

Kim J, Jung H, Jang H and Park H (2024) Improving multi-hop logical reasoning in knowledge graphs with context-aware query representation learning. In: Ku LW, Martins A and Srikumar V (eds.) *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 15978–15991. DOI:10.18653/v1/2024.findings-acl.946. URL https://aclanthology.org/2024.findings-acl.946/.

Kulmanov M, Liu-Wei W, Yan Y and Hoehndorf R (2019) El embeddings: Geometric construction of models for the description logic el ++. URL https://arxiv.org/abs/1902.10499.

Levac D, Colquhoun H and O'Brien KK (2010) Scoping studies: advancing the methodology. *Implementation Science* 5(1): 69.

Li H, Liu H, Wang Y, Xin G and Wei Y (2023) Degreembed: Incorporating entity embedding into logic rule learning for knowledge graph reasoning. *Semantic Web* 14(6): 1099–1119. DOI:10.3233/SW-233413. URL https://journals.sagepub.com/doi/abs/10.3233/SW-233413.

Li Z, Yu L, Yue K and Wu X (2025) Differentiable probabilistic logic reasoning for knowledge graph completion. In: *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, CIKM '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720406, p. 1758–1767. DOI:10.1145/3746252.3761081. URL https://doi.org/10.1145/3746252.3761081.

Liu L, Du B, Ji H, Zhai C and Tong H (2021a) Neural-answering logical queries on knowledge graphs. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325, p. 1087–1097. DOI:10.1145/3447548.3467375. URL https://doi.org/10.1145/3447548.3467375.

Liu R, Yin G and Liu Z (2024) Learning to walk with logical embedding for knowledge reasoning. *Information Sciences* 667: 120471. DOI:https://doi.org/10.1016/j.ins.2024.120471. URL https://www.sciencedirect.com/science/article/pii/S0020025524003840.

Liu Y, Hildebrandt M, Joblin M, Ringsquandl M, Raissouni R and Tresp V (2021b) Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. URL https://arxiv.org/abs/2103.10367.

Mancino ACM, Ferrara A, Bufi S, Malitesta D, Di Noia T and Di Sciascio E (2023) Kgtore: Tailored recommendations through knowledge-aware gnn models. In: *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702419, p. 576–587. DOI:10.1145/3604915.3608804. URL https://doi.org/10.1145/3604915.3608804.

Manhaeve R, Giannini F, Ali M, Azzolini D, Bizzarri A, Borghesi A, Bortolotti S, De Raedt L, Dhami D, Diligenti M, Dumančić S, Faltings B, Gentili E, Gerevini A, Gori M, Guns T, Homola M, Kersting K, Lehmann J, Lombardi M, Lorello L, Marconato E, Melacci S, Passerini A, Paul D, Riguzzi F, Teso S, Yorke-Smith N and Lippi M (2026) Benchmarking in neuro-symbolic ai. In: Dai WZ (ed.) *Learning and Reasoning*. Cham: Springer Nature Switzerland. ISBN 978-3-032-09087-4, pp. 238–249.

Marcus G (2020) The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177* .

Martinez Lorenzo AC, Perfilyev A, Markl V, Clokie M, Sicheritz-Pontén T and Kaoudi Z (2025) Modular neuro-symbolic knowledge graph completion. In: *VLDB 2025 Workshops*. VLDB Endowment, pp. 1–4. URL https://karmaresearch.github.io/NILS2025/. Workshop on New Ideas for Large-Scale Neurosymbolic Learning Systems, NILS ; Conference date: 05-09-2025.

Mashkova O, Zhapa-Camacho F and Hoehndorf R (2024) Enhancing geometric ontology embeddings for $\mathcal{EL}^{++}$ with negative sampling and deductive closure filtering. URL https://arxiv.org/abs/2405.04868.

Mashkova O, Zhapa-Camacho F and Hoehndorf R (2026) Dele: Deductive el++ embeddings for knowledge base completion. *Neurosymbolic Artificial Intelligence* 2: 29498732261420011. DOI:10.1177/29498732261420011. URL https://doi.org/10.1177/29498732261420011.

Memariani A, Glauer M, Flügel S, Neuhaus F, Hastings J and Mossakowski T (2025) Box embeddings for extending ontologies: A data-driven and interpretable approach. DOI:10.21203/rs.3.rs-6546788/v1.

Moghimifar F, Qu L, Zhuo TY, Haffari G and Baktashmotlagh M (2021) Neural-symbolic commonsense reasoner with relation predictors. In: Zong C, Xia F, Li W and Navigli R (eds.) *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 797–802. DOI:10.18653/v1/2021.acl-short.100. URL https://aclanthology.org/2021.acl-short.100/.

Mohapatra B, Bhatia S, Mutharaju R and Srinivasaraghavan G (2021) Emelvar: A neurosymbolic reasoner for the el++ description logic. In: *Proceedings of the Semantic Reasoning Evaluation Challenge (SemREC 2021), co-located with the 20th International Semantic Web Conference (ISWC 2021)*, *CEUR Workshop Proceedings*, volume 3123. Virtual Event, pp. 44–51. URL https://ceur-ws.org/Vol-3123/paper6.pdf.

Mondal S, Bhatia SK and Mutharaju R (2021) Emel++: Embeddings for el++ description logic. In: *AAAI Spring Symposium Combining Machine Learning with Knowledge Engineering*. URL https://api.semanticscholar.org/CorpusID:235416241.

Motik B, Patel-Schneider P, Bock C, Fokoue A, Haase P, Hoekstra R, Horrocks I, Ruttenberg A, Sattler U and Smith M (2008) Owl 2 web ontology language: Structural specification and functional-style. *Journal of Pragmatics - J PRAGMATICS* 27.

Newell A and Simon HA (1980) *Human Problem Solving*. Prentice-Hall.

Ontiveros RC, Bonabi Mobaraki E, Giannini F, Barbiero P, Gori M and Diligenti M (2026) Interpretable link prediction via neural-symbolic reasoning. In: Guidotti R, Schmid U and Longo L (eds.) *Explainable Artificial Intelligence*. Cham: Springer Nature Switzerland. ISBN 978-3-032-08324-1, pp. 319–331.

Peng X, Tang Z, Kulmanov M, Niu K and Hoehndorf R (2022) Description logic el++ embeddings with intersectional closure. URL https://arxiv.org/abs/2202.14018.

Pflueger M, Tena Cucala DJ and Kostylev EV (2022) Gnnq: A neuro-symbolic approach to query answering over incomplete knowledge graphs. In: Sattler U, Hogan A, Keet M, Presutti V, Almeida JPA, Takeda H, Monnin P, Pirrò G and d'Amato C (eds.) *The Semantic Web – ISWC 2022*. Cham: Springer International Publishing. ISBN 978-3-031-19433-7, pp. 481–497.

Purohit D, Chudasama Y and Vidal ME (2025a) Capturing symbolic knowledge of constraints and incompleteness to guide inductive learning in neuro-symbolic knowledge graph completion. pp. 111–118. DOI:10.1145/3731443.3771355.

Purohit D, Chudasama Y and Vidal ME (2025b) Enhancing medical knowledge discovery: A neuro-symbolic system for inductive learning over medical kgs. In: *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713293, p. 1108–1109. DOI:10.1145/3701551.3708814. URL https://doi.org/10.1145/3701551.3708814.

Rajalakshmi R, Unhelkar B and Shankar S (2025) Neuro-symbolic integration using knowledge attention graphs with advanced deep learning techniques for detecting brain disorders. *International Insurance Law Review* 33(S5): 420–436. DOI:10.65677/iilr.33.S5.27. URL https://lumarpub.com/iilr/article/view/33.S5.27.

Ren H, Hu W and Leskovec J (2020) Query2box: Reasoning over knowledge graphs in vector space using box embeddings. URL https://arxiv.org/abs/2002.05969.

Rivas A, Collarana D, Torrente M and Vidal ME (2024) A neuro-symbolic system over knowledge graphs for link prediction. *Semantic Web* 15(4): 1307–1331. DOI:10.3233/SW-233324. URL https://journals.sagepub.com/doi/abs/10.3233/SW-233324.

Russell SJ and Norvig P (2016) *Artificial Intelligence: A Modern Approach*. 3rd edition. Pearson.

Sen P, de Carvalho BWSR, Abdelaziz I, Kapanipathi P, Luus FPS, Roukos S and Gray AG (2021) Combining rules and embeddings via neuro-symbolic AI for knowledge base completion. *CoRR* abs/2109.09566. URL https://arxiv.org/abs/2109.09566.

Sharma A and Jain S (2025) Nesymatch: A neuro-symbolic approach for knowledge alignment. DOI:10.21203/rs.3.rs-7921039/v1.

Shen T, Mao R, Wang J, Zhang X and Cambria E (2025) Flow-guided direct preference optimization for knowledge graph reasoning with trees. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400715921, p. 1165–1175. DOI:10.1145/3726302.3729980. URL https://doi.org/10.1145/3726302.3729980.

Shengyuan C, Cai Y, Fang H, Huang X and Sun M (2023) Differentiable neuro-symbolic reasoning on large-scale knowledge graphs. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M and Levine S (eds.) *Advances in Neural Information Processing Systems*, volume 36. Curran Associates, Inc., pp. 28139–28154. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/5965f3a748a8d41415db2bfa44635cc3-Paper-Conference.pdf.

Singh G (2023) Benchmarking symbolic and neuro-symbolic description logic reasoners. Doctoral Consortium at International Semantic Web Conference.

Spillo G, Musto C, Degemmis M, Lops P and Semeraro G (2024) Recommender systems based on neuro-symbolic knowledge graph embeddings encoding first-order logic rules. *User Modeling and User-Adapted Interaction* 34: 2039 – 2083. URL https://api.semanticscholar.org/CorpusID:272940957.

Suchanek F, Kasneci G and Weikum G (2007) Yago: a core of semantic knowledge. pp. 697–706. DOI:10.1145/1242572.1242667.

Susskind Z, Arden B, John LK, Stockton P and John EB (2021) Neuro-symbolic ai: An emerging class of ai workloads and their characterization.

Teymurova S, Jiménez-Ruiz E, Weyde T and Chen J (2024) Owl2vec4oa: Tailoring knowledge graph embeddings for ontology alignment. URL https://arxiv.org/abs/2408.06310.

Toutanova K and Chen D (2015) Observed versus latent features for knowledge base and text inference. In: Allauzen A, Grefenstette E, Hermann KM, Larochelle H and Yih SWt (eds.) *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*. Beijing, China: Association for Computational Linguistics, pp. 57–66. DOI:10.18653/v1/W15-4007. URL https://aclanthology.org/W15-4007/.

Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, Moher D, Peters MDJ, Horsley T, Weeks L, Hempel S, Akl EA, Pronovost PJ, Westert GJ, Tutungi R and Straus SE (2018) PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. *Annals of Internal Medicine* 169(7): 467–473. DOI:10.7326/M18-0850.

Vollmers D, Jalota R, Moussallem D, Topiwala H, Ngonga Ngomo AC and Usbeck R (2021) *Knowledge Graph Question Answering Using Graph-Pattern Isomorphism*. IOS Press. DOI:10.3233/ssw210038. URL http://dx.doi.org/10.3233/SSW210038.

Wang S, Xie B, Ding L, Chen J and Xiang Y (2025) Reinforced logical reasoning over kgs for interpretable recommendation system. *Mach. Learn.* 114(4). DOI:10.1007/s10994-024-06646-4. URL https://doi.org/10.1007/s10994-024-06646-4.

Wu X and Zhao Y (2025) A neuro-symbolic approach to symbol grounding for alc-ontologies. In: *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714542, p. 3240–3249. DOI:10.1145/3711896.3736926. URL https://doi.org/10.1145/3711896.3736926.

Xian Y, Fu Z, Zhao H, Ge Y, Chen X, Huang Q, Geng S, Qin Z, de Melo G, Muthukrishnan S and Zhang Y (2020) Cafe: Coarse-to-fine neural symbolic reasoning for explainable recommendation. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20. ACM, p. 1645–1654. DOI:10.1145/3340531.3412038. URL http://dx.doi.org/10.1145/3340531.3412038.

Xu Z, Zhang W, Ye P, Chen H and Chen H (2022) Neural-symbolic entangled framework for complex query answering. In: Koyejo S, Mohamed S, Agarwal A, Belgrave D, Cho K and Oh A (eds.) *Advances in Neural Information Processing Systems*, volume 35. Curran Associates, Inc., pp. 1806–1819. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/0bcfb525c8f8f07ae10a93d0b2a40e00-Paper-Conference.pdf.

Yang H, Chen J and Sattler U (2025) Transbox: El++-closed ontology embedding. In: *Proceedings of the ACM on Web Conference 2025*, WWW '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400712746, p. 22–34. DOI:10.1145/3696410.3714672. URL https://doi.org/10.1145/3696410.3714672.

Yu D, Yang B, Liu D, Wang H and Pan S (2023) A survey on neural-symbolic learning systems. *Neural Networks* 166: 105–126. DOI:https://doi.org/10.1016/j.neunet.2023.06.028. URL https://www.sciencedirect.com/science/article/pii/S0893608023003398.

Zhang C, Peng Z, Zheng J and Ma Q (2024) Conditional logical message passing transformer for complex query answering. URL https://arxiv.org/abs/2402.12954.

Zhang W, Paudel B, Wang L, Chen J, Zhu H, Zhang W, Bernstein A and Chen H (2019) Iteratively learning embeddings and rules for knowledge graph reasoning. In: *The World Wide Web Conference*, WWW '19. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748, p. 2366–2377. DOI:10.1145/3308558.3313612. URL https://doi.org/10.1145/3308558.3313612.

Zhapa-Camacho F and Hoehndorf R (2023) Cate: Embedding $\mathcal{ALC}$ ontologies using category-theoretical semantics. DOI:10.48550/arXiv.2305.07163.

Zhu X, Liu B, Zhu C, Ding Z and Yao L (2023) Approximate reasoning for large-scale abox in owl dl based on neural-symbolic learning. *Mathematics* 11(3). DOI:10.3390/math11030495. URL https://www.mdpi.com/2227-7390/11/3/495.

Zhurov V, Kausch J, Sedig K and Milani M (2025) Fuzzy ontology embeddings and visual query building for ontology exploration. *Informatics* 12(4). DOI:10.3390/informatics12040133. URL https://www.mdpi.com/2227-9709/12/4/133.

Supporting Material

Full Search Queries

Query 1: Broad coverage of neurosymbolic systems over ontologies and knowledge graphs. This query aims to identify neurosymbolic reasoning systems and frameworks that operate on ontologies and knowledge graphs.

```
("neuro-symbolic" OR "neurosymbolic" OR "neuro symbolic"
OR "neural-symbolic" OR "neuralsymbolic" OR "neural symbolic"
OR "hybrid neural symbolic" OR "neural-symbolic integration")
AND
("ontology" OR "ontologies"
OR "knowledge graph" OR "knowledge graphs"
OR "knowledge base" OR "knowledge bases"
OR "RDF" OR "OWL" OR "description logic")
AND
("reasoner" OR "framework"
OR "reasoning system"
OR "reasoning framework"
OR "inference system")
```

Query 2: Ontology- and knowledge graph–based benchmarks and datasets. This query retrieves datasets and benchmarks used to evaluate neurosymbolic reasoning systems.

```
("neuro-symbolic" OR "neurosymbolic" OR "neuro symbolic"
OR "neural-symbolic" OR "neuralsymbolic" OR "neural symbolic"
OR "hybrid neural symbolic" OR "neural-symbolic integration")
AND
("ontology" OR "ontologies"
OR "knowledge graph" OR "knowledge graphs"
OR "knowledge base" OR "knowledge bases"
OR "RDF" OR "OWL" OR "description logic")
AND
("benchmark" OR "dataset" OR "datasets"
OR "benchmark dataset" OR "benchmark datasets"
OR "evaluation")
```

Query 3: Reasoning tasks and evaluation approaches over ontologies and knowledge graphs. Designed to capture reasoning tasks and evaluation methods applied to neurosymbolic systems.

```
("neuro-symbolic" OR "neurosymbolic" OR "neuro symbolic"
OR "neural-symbolic" OR "neuralsymbolic" OR "neural symbolic"
OR "hybrid neural symbolic" OR "neural-symbolic integration")
```

```
AND
("ontology" OR "knowledge graph" OR "knowledge base")
AND
("reasoning task" OR "reasoning"
OR "logical inference"
OR "ontology reasoning"
OR "knowledge graph reasoning"
OR "evaluation")
```

Query 4: Dataset limitations, benchmarking challenges, and standards. Aimed at identifying limitations in existing datasets, benchmarking challenges, and proposed standards.

```
("neuro-symbolic" OR "neurosymbolic" OR "neuro symbolic"
OR "neural-symbolic" OR "neuralsymbolic" OR "neural symbolic"
OR "hybrid neural symbolic" OR "neural-symbolic integration")
AND
("ontology" OR "knowledge graph" OR "knowledge base")
AND
("dataset limitation" OR "dataset limitations"
OR "benchmarking challenge"
OR "benchmarking standard"
OR "dataset creation")
```

Query 5: Reviews and surveys on neurosymbolic reasoning over ontologies and knowledge graphs. This query collects existing reviews or survey articles to provide context and identify gaps in prior syntheses.

```
("neuro-symbolic" OR "neurosymbolic" OR "neuro symbolic"
OR "neural-symbolic" OR "neuralsymbolic" OR "neural symbolic"
OR "hybrid neural symbolic" OR "neural-symbolic integration")
AND
("ontology" OR "knowledge graph" OR "knowledge base")
AND
("review" OR "survey")
```

Query 6: Extended coverage. To ensure coverage of systems not explicitly labeled as neurosymbolic, we included an additional query targeting embedding-based and differentiable reasoning approaches.

This additional query mitigates terminology bias in the neurosymbolic literature, where hybrid approaches are often described using embedding- or constraint-oriented terminology rather than explicitly labeled as neurosymbolic.

```
("ontology embedding"
OR "description logic embedding"
OR "logical embedding"
OR "axiom embedding"
OR "box embedding"
OR "neural theorem proving"
OR "differentiable reasoning"
OR "logic regularization"
OR "logic constraint"
OR "constraint-based learning"
OR "logic-guided learning")
AND
("ontology" OR "ontologies"
OR "knowledge graph" OR "knowledge graphs"
OR "knowledge base" OR "knowledge bases"
OR "RDF" OR "OWL" OR "description logic")
AND
("reasoning" OR "inference" OR "entailment"
OR "classification" OR "subsumption"
OR "consistency" OR "query answering")
```

Supplementary Tables

Table 11. Ontology and knowledge graphs benchmark datasets used by exactly one reported neurosymbolic system.

Dataset	**Domain**	**Task(s)**
Biological Interaction Network (Bio)	Biological interaction knowledge graph	Complex Query Answering (Hamilton et al., 2018)
Carcinogenesis	Chemo-/bioinformatics relational benchmark (carcinogenesis)	Concept Learning Support (Kamdem Teyou et al., 2025), Instance Retrieval (Kamdem Teyou et al., 2025), Robust Reasoning under Incompleteness/Inconsistency (Kamdem Teyou et al., 2025)
ChEBI	Chemical ontology	Disjointness (Memariani et al., 2025), Hierarchical Multi-label Classification (Memariani et al., 2025), Overlap (Memariani et al., 2025), Subsumption (Memariani et al., 2025)

Dataset	Domain	Task(s)
Claros	Cultural heritage linked-data knowledge graph	Membership (Hohenecker and Lukasiewicz, 2020)
CMHR (Chinese Medical High-order Relational dataset)	Clinical/biomedical knowledge graph	Higher-order Relational Reasoning (Bhuyan et al., 2024), Hypergraph Link Prediction (Bhuyan et al., 2024)
CodeX	Biomedical knowledge graph	Link Prediction (Shengyuan et al., 2023), Probabilistic Logic Reasoning (Shengyuan et al., 2023)
Datatourisme KG	Tourism linked-data knowledge graph (France)	Constraint Validation (Boulakbech and Wannous, 2025), Knowledge Graph Enrichment (Boulakbech and Wannous, 2025), Ontology Evolution (Boulakbech and Wannous, 2025)
DB100K	Entertainment/movie knowledge graph benchmark	Link Prediction (Purohit et al., 2025a)
DBbook	Books domain knowledge graph	Recommendation (Spillo et al., 2024)
DBP15K	Cross-lingual entity alignment benchmark (DBpedia)	Entity Alignment (Chen et al., 2025)
DBP1M	Large-scale entity alignment benchmark (DBpedia-derived)	Entity Alignment (Chen et al., 2025)
DBpedia20k	DBpedia subset knowledge graph benchmark	Link Prediction (Martinez Lorenzo et al., 2025)
Disease Ontology (DO)	Disease ontology	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
DisGeNET	Gene–disease association knowledge graph	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
Family Trees	Genealogy knowledge graph	Membership (Hohenecker and Lukasiewicz, 2020)
Family2	Family relations knowledge graph	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)

Dataset	**Domain**	**Task(s)**
Father	Family relations knowledge graph (father relation)	Concept Learning Support (Kamdem Teyou et al., 2025), Instance Retrieval (Kamdem Teyou et al., 2025), Robust Reasoning under Incompleteness/Inconsistency (Kamdem Teyou et al., 2025)
FB-AUTO	Automotive domain knowledge graph benchmark	Link Prediction (Abboud et al., 2020), Rule Injection (Abboud et al., 2020)
Food-Biomarker Ontology (FOBI)	Food biomarker ontology	Consistency Checking (Zhapa-Camacho and Hoehndorf, 2023), Link Prediction (PPI) (Zhapa-Camacho and Hoehndorf, 2023), Ontology Completion (Zhapa-Camacho and Hoehndorf, 2023), Subsumption (Zhapa-Camacho and Hoehndorf, 2023)
French Royalty	Historical genealogy knowledge graph	Link Prediction (Purohit et al., 2025a)
GlycoRDF	Glycobiology linked-data knowledge graph	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)
Hetionet	Integrative biomedical knowledge graph	Link Prediction (Liu et al., 2021b), Multi-hop Reasoning (Liu et al., 2021b)
JF17K	General-purpose knowledge graph benchmark	Link Prediction (Abboud et al., 2020), Rule Injection (Abboud et al., 2020)
KQA Pro	Knowledge-graph question answering (complex reasoning)	Knowledge Graph Question Answering (Agarwal and Bedathur, 2025)
LC-QuAD 2.1	Knowledge-graph question answering	Knowledge Graph Question Answering (Gomes et al., 2022), Semantic Parsing (Gomes et al., 2022), Template Classification (Gomes et al., 2022)
LUBM	OWL ontology reasoning benchmark (university domain)	Link Prediction (Martinez Lorenzo et al., 2025)
Lung cancer KG	Lung cancer clinical/biomedical knowledge graph	Link Prediction (Purohit et al., 2025b)
Lung cancer polypharmacy treatment KG (clinical records + DrugBank DDIs)	Clinical treatment KG with drug–drug interactions	Link Prediction (Rivas et al., 2024)

Dataset	Domain	Task(s)
Mutagenesis	Chemo-/bioinformatics relational benchmark (mutagenesis)	Concept Learning Support (Kamdem Teyou et al., 2025), Instance Retrieval (Kamdem Teyou et al., 2025), Robust Reasoning under Incompleteness/Inconsistency (Kamdem Teyou et al., 2025)
NCIT-DOID	Biomedical ontology mapping benchmark (cancer/disease)	Subsumption (Chen et al., 2023)
Neural Reprogramming Ontology (NRO)	Biomedical ontology (neural reprogramming)	Consistency Checking (Zhapa-Camacho and Hoehndorf, 2023), Link Prediction (PPI) (Zhapa-Camacho and Hoehndorf, 2023), Ontology Completion (Zhapa-Camacho and Hoehndorf, 2023), Subsumption (Zhapa-Camacho and Hoehndorf, 2023)
NL27k	Uncertain/weighted knowledge graph benchmark (NELL-derived)	Confidence Prediction (Chen et al., 2021b), Link Prediction (Chen et al., 2021b)
OGBL-WIKIKG2	Large-scale general-purpose knowledge graph benchmark (Wiki-based)	Inductive Complex Query Answering (Galkin et al., 2022)
Ontodm	Biomedical ontology (data mining)	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)
OpenEA	Entity alignment benchmark suite	Entity Alignment (Chen et al., 2025)
ORE	OWL reasoner evaluation resources	Subsumption (Mohapatra et al., 2021)
OWL2Bench	Synthetic OWL ontology benchmark generator	Subsumption (Mohapatra et al., 2021)
OWL-Centric Dataset	Based on Linked Data Cloud and Data Hub	Deductive Reasoning (Ebrahimi et al., 2021b), Knowledge Graph Entailment (Ebrahimi et al., 2021b)
Pizza ontology	Toy OWL ontology	Approximate Query Answering (Zhurov et al., 2025), Semantic Search (Zhurov et al., 2025)
QALD-8	Knowledge-graph question answering/semantic parsing benchmark	Knowledge Graph Question Answering (Vollmers et al., 2021), Semantic Parsing (Vollmers et al., 2021), Template Classification (Vollmers et al., 2021)

Dataset	**Domain**	**Task(s)**
Semantic Bible	Religious/historical knowledge graph	Concept Learning Support (Kamdem Teyou et al., 2025), Instance Retrieval (Kamdem Teyou et al., 2025), Robust Reasoning under Incompleteness/Inconsistency (Kamdem Teyou et al., 2025)
SIDER	Drug adverse effect knowledge graph	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
SNOMED CT	Clinical terminology ontology	Subsumption (Mondal et al., 2021)
SportsNELL	Sports domain knowledge graph (NELL-derived)	Link Prediction (Abboud et al., 2020), Rule Injection (Abboud et al., 2020)
Sso	Taxonomy ontology	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)
STITCH	Chemical–protein interaction knowledge graph	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
STRING	Protein–protein interaction knowledge graph/network	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
SwissProt GO annotations	Protein–GO annotation graph	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
Time	Temporal ontology/temporal knowledge graph	Link Prediction (Zhu et al., 2023), Membership (Zhu et al., 2023), Subsumption (Zhu et al., 2023)
Vicodi	Historical/cultural linked-data knowledge graph	Concept Learning Support (Kamdem Teyou et al., 2025), Instance Retrieval (Kamdem Teyou et al., 2025), Robust Reasoning under Incompleteness/Inconsistency (Kamdem Teyou et al., 2025)
WatDiv	RDF/SPARQL benchmark knowledge graph	Inductive Complex Query Answering (Pflueger et al., 2022), Node Classification (Pflueger et al., 2022)
WikiTopics-QA	Knowledge-base/wiki-focused question answering dataset	Complex Query Answering (Galkin et al., 2024)

Dataset	Domain	Task(s)
Yeast PPI dataset	Yeast protein–protein interaction knowledge graph/network	Consistency Checking (Zhapa-Camacho and Hoehndorf, 2023), Link Prediction (PPI) (Zhapa-Camacho and Hoehndorf, 2023), Ontology Completion (Zhapa-Camacho and Hoehndorf, 2023), Subsumption (Zhapa-Camacho and Hoehndorf, 2023)

Table 12. Task-to-metric mapping.

Task	Supported metric(s)	Neuro role	Symbolic role
(G1) Grounding & parsing			
Semantic Parsing: Map natural language to an executable logical form (often SPARQL) over a KG/ontology	Accuracy; F1; F1 (QALD); Precision; Recall; Runtime	Maps NL to logical form/SPARQL; entity/relation linking	Executes logical form; checks type constraints and logical validity
Template Classification: Classify an NL question into a query template/operator pattern for downstream parsing/execution	Accuracy; F1; F1 (QALD); Precision; Recall; Runtime	Classifies question/query template from NL	Template corresponds to symbolic query structure; ensures feasability
Semantic Search: Retrieve relevant entities/concepts using semantic similarity while respecting ontology constraints	Hits@K; MRR; Overlap@k; Similarity vs. distance curve; Subsumption violation rate	Embedding-based retrieval for concepts/entities	Ontology constraints
(G2) Representation learning with logic			
Recommendation: Recommend items to users using KG/ontology (entities/relations/types as context)	Diversity; F1; Hits@K; MAP; Mean Average Recall (MAR); Novelty; Precision; Recall; nDCG	Learns user/item representations; predicts relevance	Uses KG constraints and paths to justify

Task	Supported metric(s)	Neuro role	Symbolic role
Confidence Prediction: Predict confidence/uncertainty of KG inferences (triples/answers) for calibration or ranking	Mean Absolute Error; Mean Squared Error; nDCG	Learns calibration/uncertainty for predictions	Optional symbolic structures (rule-based confidence, constraint-aware calibration)
Disjointness: Predict or assess disjointness-related behavior (e.g., ensure disjoint classes don't overlap)	F1; Precision; Recall	Learns embeddings that separate disjoint classes	Encodes disjointness axioms; checks violations
Hierarchical Multi-label Classification: Predict multiple labels with hierarchical dependencies (parent–child constraints)	F1; Precision; Recall	Learns multi-label classifier	Enforces hierarchy closure, parent-child constraints, disjointness
Node Classification: Assign labels/types to nodes in a graph using structure/attributes	Average Precision; Precision; Recall	Learns node representations & labels	Optional label constraints from ontology
Rule Injection: Improve a neural model by injecting known symbolic rules as constraints/features	Hits@K; MRR; Mean Rank	Learns predictor regularized by injected rules	Rules define constraints; checks rule satisfaction
(G3) Structured reasoning over learned representations			
Path Reasoning: Infer answers by composing relations along paths (explicit path scoring/composition)	Hits@K; MAP; MRR; Precision; Recall; nDCG	Learns path composition or neural path ranking	Path validity constrained by schema/rules; (often) explanation via symbolic paths

Task	**Supported metric(s)**	**Neuro role**	**Symbolic role**
Higher-order Relational Reasoning: Reason over relations beyond pairwise edges (e.g., relational patterns/compositions)	AUC-ROC; Accuracy; F1	Learns compositional reasoning over relations	Encodes higher-order constraints/rules; validates relational structure
Instance Retrieval: Retrieve/rank individuals that satisfy a concept/query (often ontology-defined)	F1; Jaccard Similarity; Runtime	Ranks individuals for a concept/query via learned scoring	Concept definition comes from ontology; checks membership conditions
Knowledge Graph Entailment: Predict whether a fact is entailed by KG/ontology/rules (entailment classification)	Accuracy; F1; Precision; Recall	Learns entailment classifier	Provides entailment rules/semantics; evaluates correctness
Probabilistic Logic Reasoning: Perform reasoning with uncertainty using weighted rules/soft constraints	Hits@K; MRR	Learns weights; amortizes inference	Performs probabilistic logical inference (PSL/MLN-style constraints)
(G4) Symbolic structure induction			
Ontology Alignment: Identify correspondences between entities (classes/properties/instances) across ontologies	F1; Hits@K; MRR; Precision; Recall	Learns lexical/structural similarity; generates candidate matches	Enforces logical coherence; repairs incoherent alignments
Concept Learning Support: Learn or support learning DL concept descriptions (class expressions) from examples	F1; Jaccard Similarity; Runtime	Proposes candidate concepts/features or scores hypotheses	Constructs/validates class expressions; ensures DL satisfiability

Task	Supported metric(s)	Neuro role	Symbolic role
Entity Alignment: Align entities across KGs (typically instance-level), often using seeds	Hits@K; MRR	Learns similarity/matching scores across KGs	Enforces 1–1 constraints, type coherence, alignment consistency/repair
Equivalence: Predict/validate equivalence relations (e.g., equivalent classes or entities)	AUC-ROC; Hits@K; Mean Rank	Scores candidate equivalences	Confirms via logical equivalence constraints/coherence
Overlap: Measure/predict overlap between classes/sets (often ontology-driven similarity/overlap)	F1; Precision; Recall	Estimates similarity/overlap of concept sets	Uses symbolic set semantics/subsumption relations
	(G5) Search & planning with neural guidance		
Consistency Checking: Detect contradictions/unsatisfiable parts of a KG/ontology (TBox/ABox consistency)	AUC-ROC; Hits@K; Mean Rank	Detects likely contradictions	Performs logical consistency checking; identifies violated axioms
Constraint Validation: Validate whether predicted/added facts satisfy constraints (e.g., SHACL/OWL restrictions)	Constraint Satisfaction Rate; F1; Ontology Extension Accuracy; Ontology Match Rate; Precision; Recall	Predicts missing facts/suggests repairs under constraints	SHACL/OWL constraints define validity
Knowledge Graph Enrichment: Add new facts/schema elements while maintaining constraint satisfaction	Constraint Satisfaction Rate; F1; Ontology Extension Accuracy; Ontology Match Rate; Precision; Recall	Proposes new edges/classes/properties	Validates against ontology/constraints; repairs inconsistencies

Task	Supported metric(s)	Neuro role	Symbolic role
Masked ABox Revision: Revise/impute missing ABox assertions while preserving satisfiability/minimal change	KL Divergence; Precision; Recall; Success Rate	Suggests imputations/edits; ranks revisions	Constraint-based revision/minimal change; ensures satisfiable ABox
Ontology Evolution: Update ontology/KG over time (add/remove axioms/facts) while preserving consistency	Constraint Satisfaction Rate; F1; Ontology Extension Accuracy; Ontology Match Rate; Precision; Recall	Predicts changes/suggests updates from data	Ensures version consistency, constraint satisfaction, minimal change
Robust Reasoning under Incompleteness/Inconsistency: Reason and predict reliably under missing facts or inconsistent/noisy axioms	F1; Jaccard Similarity; Runtime	Learns noise-aware scoring; detects uncertainty	Repair/constraint handling; inconsistency-tolerant semantics
(G6) Explainable inference			
Rule-Path Reasoning: Answer queries with explicit rule/path chains serving as proof-like explanations	Hits@K; MRR	Selects useful rules/paths for a query (neural scoring)	Produces explicit rule/path chain; validates chain logically

Acknowledgments

During the preparation of this manuscript, the authors used ChatGPT (OpenAI) to improve the readability of the text and correct grammatical errors.