

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

Journal Title

XX(X):2-83

©The Author(s) 2016

Reprints and permission:

sagepub.co.uk/journalsPermissions.nav

DOI: 10.1177/ToBeAssigned

www.sagepub.com/

SAGE

Julie Loesch¹ and Michel Dumontier¹ and Remzi Celebi¹

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 Neurosymbolic artificial intelligence combines neural learning with symbolic knowledge representation and reasoning, yet the architectures and evaluation practices used for systems operating over knowledge graphs and ontologies remain heterogeneous. 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 that categorizes and characterizes neurosymbolic systems according to their dominant neural-symbolic integration mechanisms, using Kautz's taxonomy as a high-level organizational framework while further examining their neural components, symbolic components, knowledge representations, and information flows. We then catalogue the datasets used to evaluate these systems and analyze the 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 reported across the reviewed literature. Our synthesis identifies substantial fragmentation in current evaluation practices, including heterogeneous benchmark families, limited cross-system comparability, and insufficient evaluation of logical correctness, robustness, and explainability. Based on these limitations, we propose a multidimensional framework for more comprehensive evaluation of neurosymbolic systems over knowledge graphs and ontologies.

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. artificial intelligence combines neural learning with symbolic knowledge representation and reasoning (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. models are effective at learning statistical regularities from large,

¹Department of Advanced Computing Sciences, Maastricht University, Netherlands

Corresponding author:

Julie Loesch

Email: julie.loesch@maastrichtuniversity.nl

noisy, or unstructured datasets, but their internal decision processes are often difficult to interpret and they do not generally guarantee logically consistent outputs (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. approaches represent knowledge explicitly through entities, relations, concepts, axioms, and rules, enabling transparent and formally defined inference. However, purely symbolic systems can be sensitive to incomplete or noisy input and may face scalability limitations. Neurosymbolic systems seek to combine these complementary capabilities by integrating learned representations with explicit knowledge and reasoning mechanisms.

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 are two prominent forms of structured knowledge used in neurosymbolic systems. A knowledge graph can be represented as a set of relational facts, commonly expressed as triples of the form (s, p, o) , where a subject s is connected to an object o through a predicate p . Ontologies additionally provide an explicit conceptualization of a domain by defining classes, relations, individuals, and formal axioms that constrain their interpretation (Gruber, 1993; Hitzler and Sarker, 2022; Baader et al., 2007). Ontology knowledge is commonly separated into a terminological component, or TBox, which defines 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), and an assertional component, or ABox, which records facts about individuals. Formal ontology languages such as the Web Ontology Language (OWL) are grounded in Description Logics and support deductive reasoning tasks including consistency checking, concept satisfiability, classification, subsumption checking, and instance checking (Besold et al., 2017a; Hogan et al., 2020; Baader et al., 2007).

These structured representations support several additional tasks studied in neural and neurosymbolic research. Knowledge graph completion and link prediction aim to identify plausible missing relations or entities. Ontology completion and subsumption prediction aim to identify missing axioms or concept relations. Logical and complex query answering require systems to answer queries involving conjunction, disjunction, negation, or relational paths. Other tasks include rule learning, ontology alignment, entity classification, and the induction or verification of symbolic structures. The term “reasoning over ontologies and knowledge graphs” is therefore used in this review in a broad sense to include both deductive inference from explicit axioms and predictive inference over incomplete structured knowledge.

Several benchmark datasets have been developed for neural learning over KGs, including A wide range of benchmarks has been adopted for these tasks. FB15k and its variants (Bordes

et al., 2013; Toutanova and Chen, 2015), FB15k-237 are derived from Freebase and contain relational facts about entities such as people, organizations, locations, and cultural works (Bordes et al., 2013; Toutanova and Chen, 2015). WN18 and WN18RR (Bordes et al., 2013; Dettmers et al., 2018), and YAGO-based resources (Suchanek et al., 2007) are derived from WordNet and contain lexical-semantic relations between synsets (Bordes et al., 2013; Dettmers et al., 2018). YAGO3-10 is a knowledge graph completion benchmark derived from YAGO3 (Dettmers et al., 2018). These datasets are primarily assessed on embedding-based tasks, such as link prediction or predominantly used for 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 link prediction. In contrast, ontology-oriented benchmark resources such as LUBM (Guo et al., 2005), ORE (Matentzoglou and Parsia, 2014), and OWL2Bench (Singh et al., 2020) contain or generate OWL ontologies for evaluating ontology management, logical inference, scalability, and related reasoning tasks. Together, these benchmark families form a heterogeneous evaluation landscape spanning predictive and deductive tasks across different knowledge representations, reasoning objectives, and evaluation metrics.

Despite rapid progress in neurosymbolic AI, evaluation practices remain fragmented. There is no commonly accepted framework of benchmarks, reasoning tasks, and metrics for systematically comparing neurosymbolic reasoning systems across integration paradigms and reasoning settings. Many systems rely either on heterogeneous Systems are evaluated on heterogeneous resources, including public knowledge graphs designed for predictive tasks, such as link prediction/KG completion, or ontology-centric, logical query answering benchmarks, individual domain ontologies, and OWL/DL benchmarks that focus on deductive reasoning benchmark suites, which support different combinations of predictive and deductive reasoning tasks. 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 Several reviews address areas related to the present study. DeLong et al. (2025) survey neurosymbolic approaches to reasoning over knowledge graphs, with particular attention to the integration of neural representations and symbolic knowledge. Dai et al. (2020) provide a broader review of knowledge graph embedding methods, their applications, and commonly used benchmark datasets. Steinmetz and Sattler (2021) examine the challenges represented in benchmark datasets for question answering over knowledge graphs. Zamazal (2020) reviews ontology benchmarks developed for

evaluating Semantic Web ontology tools, while Seeliger et al. (2019) examine the use of Semantic Web technologies to support the explainability of machine-learning models.

These studies provide valuable analyses of individual method families, tasks, applications, or benchmark traditions. In contrast, the present review jointly examines neurosymbolic systems operating over both knowledge graphs and formal ontologies, their architectural integration according to Kautz's framework, the datasets used to evaluate them, the corresponding reasoning tasks and metrics, and limitations in current evaluation practices.

The present scoping review addresses this gap by connecting four levels of analysis. First, it categorizes neurosymbolic systems according to their dominant neural-symbolic integration mechanisms and further characterizes their neural components, symbolic components, underlying knowledge representations, and information flows. Kautz's six types are used as a high-level organizational framework, while architectural overlaps and finer-grained distinctions are considered in the system descriptions. Second, the review provides a systematic catalogue of datasets, analyze prevalent reasoning tasks and evaluation metrics, and identify limitations to guide the design of future benchmarks ontology- and knowledge graph-based datasets used in the included studies. Third, it maps the reasoning tasks performed on these datasets to the reported evaluation metrics. Fourth, it synthesizes limitations in existing benchmarks and evaluation practices and proposes a multidimensional framework for more comprehensive evaluation.

Specifically, this scoping review addresses the following objectives:

- Provide an overview of neurosymbolic reasoning systems and frameworks that integrate Categorize and characterize neurosymbolic systems that operate over ontologies and knowledge graphs, including their neural components, symbolic components, knowledge representations, and integration mechanisms.
- Present a systematic catalogue of benchmark datasets based on ontologies and knowledge graphs for evaluating neurosymbolic AI Ontology- and knowledge graph-based datasets used to evaluate neurosymbolic systems.
- Discuss prevalent ontology- and knowledge graph-centric Analyze the reasoning tasks and evaluation metrics used in the field reported for these systems and datasets.
- Identify key limitations in existing ontology- and knowledge graph-based datasets and evaluation practices , thereby improving future dataset development and benchmarking standards and derive recommendations for future benchmark development and system evaluation.

The remainder of this paper is structured as follows:

- **Background** defines neurosymbolic AI, its fundamental concepts , and key principles that underpin this interdisciplinary field introduces the principal knowledge representation and reasoning concepts required for the review and explains the high-level framework used to organize neurosymbolic architectures.

- **Methodology** 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 describes the search strategy, study-selection procedure, data-charting process, and synthesis methodology.
- **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 categorizes and characterizes the included neurosymbolic systems and presents the datasets, tasks, and evaluation metrics employed identified in the review.
- **Discussion** critically examines the limitations associated with current ontology- and knowledge graph-based analyzes limitations in current benchmarks and evaluation practices , and explores potential directions for future research and improvements and presents recommendations for more comprehensive neurosymbolic evaluation.
- Finally, **Conclusion** summarizes the key principal 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

Knowledge graphs and ontologies

A knowledge graph can be represented as a directed, edge-labelled graph $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is a set of entities, $\mathcal{R}$ is a set of relation types, and $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ is a set of factual triples (Hogan et al., 2020). A triple (h, r, t) states that a head entity h is related to a tail entity t through relation r . Depending on the underlying formalism, knowledge graphs may range from collections of relational facts with limited explicit semantics to semantically richer graphs governed by schemas, constraints, and logical axioms.

An ontology is an explicit specification of a conceptualization (Gruber, 1993). Ontologies define the concepts, relations, individuals, and axioms used to describe a domain. In Description Logic terminology Baader et al. (2007), the TBox contains terminological axioms concerning concepts and relations, such as explicit knowledge representation and logical inference, with the adaptability and pattern recognition capabilities of neural networks concept inclusions, equivalences, disjointness axioms, domain restrictions, and range restrictions. The ABox contains assertions about individuals, such as class-membership and object-property assertions. Some formalisms additionally distinguish an RBox containing role axioms, including role inclusions, transitivity, and role-chain axioms.

The Resource Description Framework (RDF) provides a graph-based data model in which statements are expressed as subject–predicate–object triples. RDF Schema

extends RDF with basic schema-level constructs. The Web Ontology Language (OWL) (Motik et al., 2008) supports richer semantic modelling and is grounded in Description Logics. Different OWL profiles and Description Logic fragments offer different trade-offs between expressivity and computational complexity. For example, the lightweight $\mathcal{EL}$ family supports large ontologies with extensive concept hierarchies and existential restrictions, while more expressive languages such as $\mathcal{ALC}$ and $\mathcal{SROIQ}$ support additional constructors including negation, disjunction, inverse roles, cardinality restrictions, and complex role axioms.

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, As a representative Description Logic, $\mathcal{ALC}$ constructs concept descriptions from atomic concepts using conjunction ($C \sqcap D$), disjunction ($C \sqcup D$), negation ($\neg C$), existential restrictions ($\exists r.C$), and universal restrictions ($\forall r.C$). Its semantics is defined with respect to an interpretation $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$, where concept names are interpreted as subsets of the domain $\Delta^{\mathcal{I}}$ and role names as binary relations over the domain. For example, $(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}$, while $(\exists r.C)^{\mathcal{I}}$ contains those individuals related through r to at least one individual in $C^{\mathcal{I}}$. An axiom $C \sqsubseteq D$ is satisfied when $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$.

Embeddings

Embedding methods represent entities, relations, concepts, or axioms as continuous vectors or geometric objects that can be processed by machine-learning models. In knowledge graph embedding, entities and relations are typically mapped to a low-dimensional space, and a scoring function is learned to estimate the plausibility of triples. These representations are commonly used for predictive tasks such as link prediction and knowledge graph completion.

Representative knowledge graph embedding approaches include TransE, which models relations as translations between entity vectors, and TransH, which extends this idea by representing relations on relation-specific hyperplanes. DistMult applies a bilinear scoring function to entity and relation embeddings, whereas RotatE represents relations as rotations in a complex vector space. These methods differ in the relational patterns that they can represent, but they are primarily data-driven approaches and do not necessarily include an explicit symbolic reasoning component.

Ontology embedding methods extend representation learning to formally defined concepts, relations, individuals, and axioms. Depending on the approach, ontology semantics may be incorporated through logical constraints, ontology-derived graph structures, geometric representations of concepts, or reasoner-generated inferences. Such embeddings can support tasks including subsumption prediction, membership prediction, ontology completion, ontology alignment, and logical query answering.

Embedding methods constitute important building blocks for several systems examined in this review. However, standalone embedding models are not classified as neurosymbolic systems unless they are integrated with an explicit symbolic

component, such as ontology axioms, logical rules, symbolic inference, constraint checking, or reasoner-generated knowledge.

Reasoning over structured knowledge

Symbolic ontology reasoners derive logical consequences from asserted axioms. Common reasoning services include consistency checking, concept satisfiability, ontology classification, subsumption checking, instance checking, and query answering. Established reasoners include ELK, which is optimized for ontologies in the OWL 2 EL profile, and HermiT (Glimm et al., 2014), which supports expressive OWL 2 ontologies through hypertableau-based reasoning.

Neural and neurosymbolic systems also perform predictive reasoning over structured knowledge. Link prediction and knowledge graph completion estimate missing triples. Entity or concept classification predicts class membership. Subsumption and ontology-completion tasks predict missing relations between concepts or other axioms. Logical query answering evaluates queries containing relational paths and logical operations such as conjunction, disjunction, and negation. Rule learning and structure induction seek reusable symbolic regularities, while ontology alignment identifies correspondences between independently developed schemas.

Deductive and predictive tasks should not be treated as mutually exclusive properties of ontology- and knowledge graph-based systems. Ontology embeddings may, for example, operate under monotonic logic, meaning that adding new rules can only increase the knowledge base without retracting previously encoded beliefs be evaluated by ranking missing subsumptions or class-membership assertions, while knowledge graph systems may incorporate rules and perform forms of logical inference. The relevant distinction is therefore the nature of the task and evaluation protocol rather than whether the resource is labelled a knowledge graph or an ontology.

Neurosymbolic integration

Neurosymbolic systems combine learned statistical components with explicit symbolic representations or reasoning processes (Susskind et al., 2021; Besold et al., 2017b; Garcez et al., 2019). Neural components may encode entities and concepts, score candidate facts or axioms, guide symbolic search, learn rule parameters, or translate unstructured input into symbolic representations. Symbolic components may supply logical constraints, produce deductive closure, verify candidate predictions, perform query evaluation, or generate structured explanations.

Many classical rule-based systems rely on manually specified rules and can be brittle when confronted with incomplete, noisy, or conflicting information (Russell and Norvig, 2016). Non-monotonic formalisms can address some of these limitations, but introduce additional modelling and computational considerations (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 proposed six broad types of neural–symbolic integration based on the relationship between neural and symbolic processing (Kautz, 2022). These types provide a recognizable high-level vocabulary for organizing the field, but they do not define mutually exclusive or exhaustive architectural classes. Multi-stage systems may exhibit characteristics associated with more than one type, particularly when their training and inference procedures use different forms of interaction. In this review, Kautz’s types are therefore used as broad organizational categories rather than as a definitive classification. The systems are additionally characterized through their knowledge representations, neural components, symbolic components, and directions of information flow.

Table 1 presents this taxonomy, summarizes the operational interpretation used in this review.

Table 1. Kautz’s neural–symbolic integration types and their operational interpretation in this review.

Integration type	Kautz notation	Operational interpretation
Neural Processing in a Symbolic input–output 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.
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.

Integration type	Kautz notation	Operational interpretation
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 sub-components during execution, while remaining primarily neural architectures.

Note: These operational interpretations support consistent presentation but do not imply that the types are mutually exclusive. Each system is organized according to its dominant integration mechanism, while substantial overlaps with other types are discussed where relevant.

Neurosymbolic learning systems exhibit several key advantages : may offer advantages in efficiency, generalization, and interpretability, depending on the architecture and task. In some settings, neural components can reduce reliance on exhaustive symbolic search or provide approximate guidance for reasoning. 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 Symbolic knowledge can also act as structured supervision or constraints that may improve generalization beyond purely data-driven models. Finally, architectures that retain explicit symbolic representations or inference traces can provide more interpretable intermediate reasoning steps than fully opaque 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).

Architectural description and Boxology

Because high-level integration types do not fully expose the internal structure of a system, this review also adopts component-oriented terminology inspired by the Boxology framework of van Bekkum et al. (2021). Boxology distinguishes the data, models, and processes involved in hybrid learning and reasoning systems and makes the direction of information flow explicit. We use this terminology descriptively to explain what information enters each component, what transformation or reasoning operation is performed, and how outputs are transferred between neural and symbolic components.

This component-level description is particularly important for systems that share a broad Kautz type but differ in their training procedure, symbolic formalism, neural

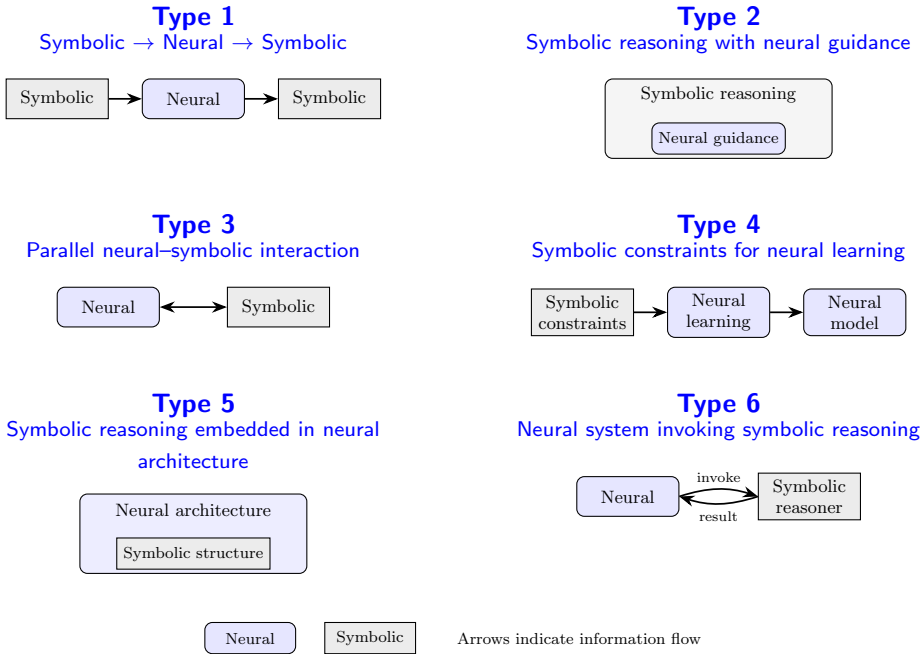


Figure 1. Schematic comparison of the six neural-symbolic integration architectures used in this review. The representation uses a consistent component-oriented notation inspired by the Boxology framework of [van Bekkum et al. \(2021\)](#). Rounded blue boxes denote neural components and grey boxes denote symbolic components. The diagrams emphasize the organization and information flow associated with each Kautz integration type.

architecture, or inference process. It also exposes boundary cases in which a system exhibits characteristics of several integration types.

To make these architectural differences explicit, Figure 1 provides a schematic comparison of the six integration types using a consistent component-oriented notation inspired by Boxology. The diagrams distinguish neural and symbolic components and emphasize their relative position and direction of information flow within each architecture.

Methodology

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, categorize and characterize neurosymbolic systems operating over ontologies and knowledge graphs, examine how they are evaluated, and systematically analyze limitations in the corresponding datasets and evaluation methodologies. This objective gives rise to the following research questions (RQs):

RQ 1 What neurosymbolic reasoning Which neurosymbolic systems have been developed for predictive or deductive reasoning over ontologies and knowledge graphs, and how can their neural components, symbolic components, knowledge representations, and integration mechanisms be characterized?

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 systems?

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 and eligibility assessment

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 eligibility criteria were applied during the screening process. :

Inclusion criteria:

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

Exclusion criteria: published in English.

- 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 in which neural and symbolic components are integrated, rather than purely neural or purely symbolic approaches considered in isolation.
- 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).

Reviews, surveys, and conceptual papers identified through the search strategy, particularly through Query 5, were used to contextualize prior work and identify gaps in existing syntheses, but were not included in the 63-study analytical corpus unless they reported an original empirical contribution satisfying the eligibility criteria.

For the purpose of study selection, an ontology or knowledge graph had to constitute an operative structured knowledge representation within the proposed system or its evaluation, rather than being mentioned only as background information. Knowledge graphs were considered in scope when entities and relations were represented explicitly as graph-structured facts or triples and were used by the learning or reasoning process. Ontologies were considered in scope when classes, relations, axioms, or constraints provided an explicit conceptual or logical structure that was used for learning, inference, prediction, validation, or evaluation. Studies in which such representations were only mentioned, but did not participate in the neurosymbolic approach, were excluded.

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 non-peer-reviewed or out-of-scope formats); (2) absence of an original empirical contribution relevant to the research questions; (3) omission of required components such as ontologies, knowledge graphs, or reasoning mechanisms absence of an operative ontology or knowledge graph, or of a neural-symbolic integration mechanism; 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. 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 reports that were sought for full-text retrieval. Twenty-six reports could not be retrieved; 249 reports were assessed for eligibility, and 63 met the eligibility criteria and were included in the scoping review. The study selection process was conducted 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

PRISMA-ScR guidelines (Tricco et al., 2018) and is summarized in the PRISMA-ScR flow diagram shown in Figure 2.

Stage 4: Charting the data

To structure our review, we develop a multi-dimensional taxonomy (Figure 3) 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).

Purely neural and purely symbolic systems are outside the scope of this review, which focuses specifically on systems that integrate neural and symbolic components.

Literature Search

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.

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.

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

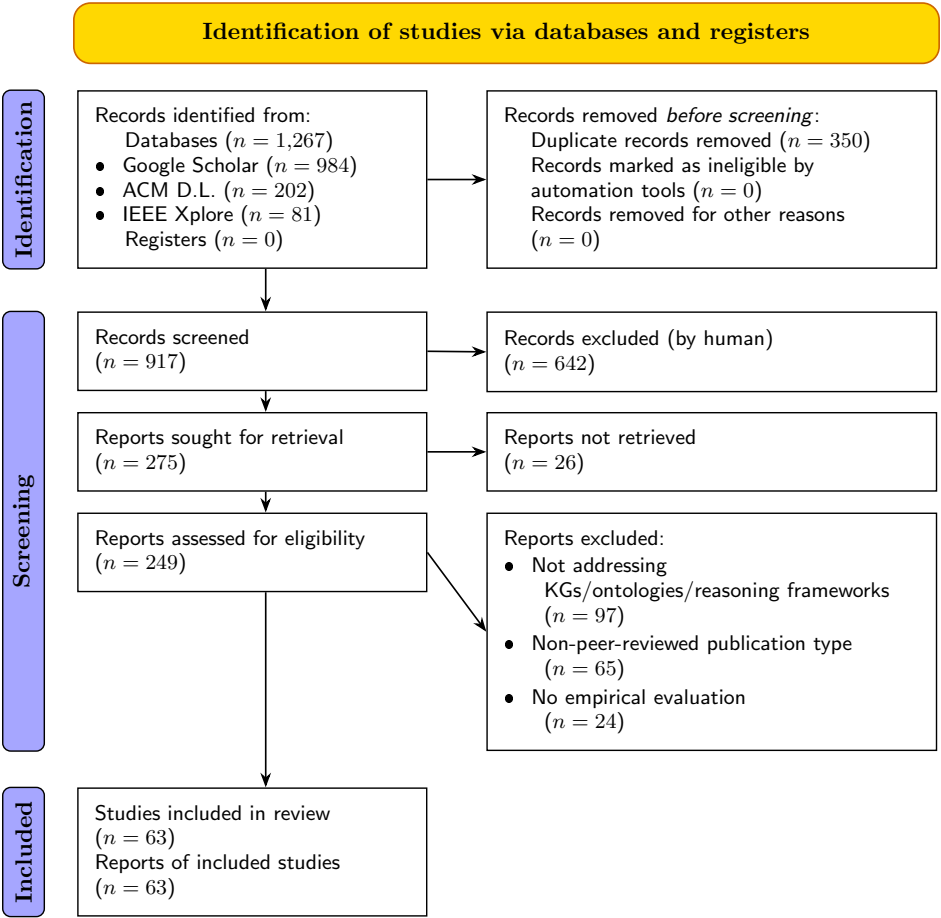


Figure 2. PRISMA-ScR (Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews) flow diagram for the study selection process. Arrows indicate the progression of records through the identification, screening, eligibility assessment, and inclusion stages.

complementary charting tables were developed. The first table characterizes the included neurosymbolic reasoning systems by documenting the symbolic formalism, neural learning component, neural-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 2symbolic integration paradigm, datasets, reasoning tasks, and evaluation metrics.

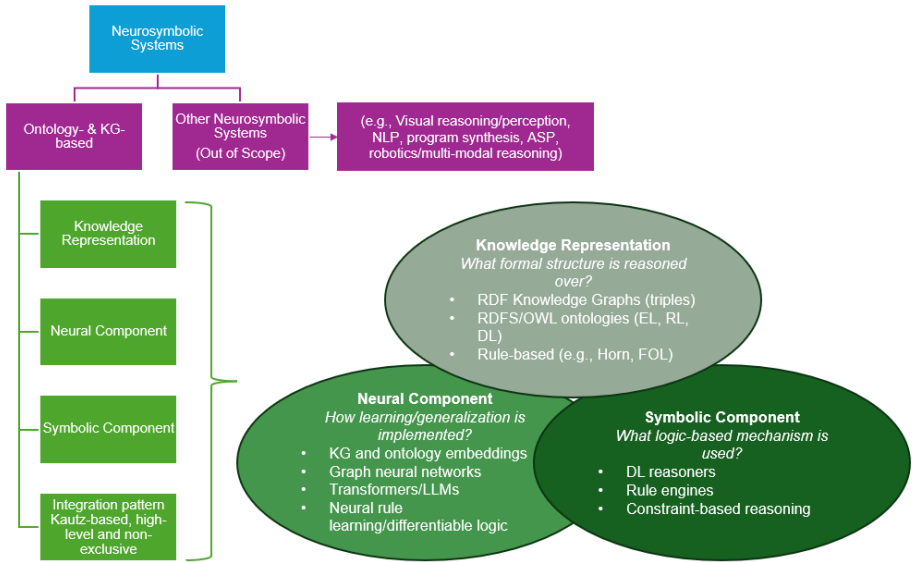


Figure 3. Taxonomy of neurosymbolic reasoning systems operating over ontologies and knowledge graphs.

Within each Kautz type, the included systems were subsequently compared according to their neural and symbolic components and their interaction mechanisms. Recurring integration mechanisms were identified inductively from the reviewed corpus and used to organize systems within each Kautz category. These mechanisms were not predefined and were not treated as mutually exclusive, since some systems combine multiple forms of neural-symbolic interaction.

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.

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-when their primary evaluation setting did not involve reasoning over an ontology or knowledge graph-centric reasoning tasks, for example, benchmarks focused exclusively on image-based or unstructured text/lexical tasks without an operative structured knowledge representation. 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).

Stage 5: Collating, summarizing, and reporting the results

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

Neurosymbolic Systems (RQ 1)

To address RQ 1, we categorize organize the identified systems using Kautz’s taxonomy, which distinguishes six paradigms for integrating neural learning with broad paradigms according to the relationship between neural learning and symbolic reasoning. The core distinction across these paradigms is where primary basis for this organization is the dominant neural–symbolic integration mechanism: 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. System, the respective roles of the neural and symbolic components, and the direction and timing of information flow between them. Because contemporary systems may combine mechanisms associated with more than one paradigm, the resulting categories are used as primary organizational clusters rather than as mutually exclusive architectural classes.

Within each Kautz type, the systems are further grouped according to integration mechanisms observed in the reviewed corpus. These mechanisms are descriptive architectural patterns used to distinguish how neural and symbolic components interact within a Kautz type; they do not necessarily represent recurrent themes across multiple studies, and a mechanism may be instantiated by one or several systems. Each subsection therefore first describes the defining architectural organization of the corresponding type, then identifies the integration mechanisms and architectural variations observed within that cluster, and finally discusses the individual systems together with the tasks and evaluation settings in which they are applied.

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) are 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 The Type 1 systems were grouped into four integration mechanisms. One integration mechanism observed in the corpus is *template-based neural classification over symbolic output selection templates*, 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

A second integration mechanism is *learning to approximate neural approximation of 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

A third integration mechanism is *predicting ontology relations, especially subsumption ontology-informed representation learning*, 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 *integration mechanism is embedding-based reasoning constrained by axioms axiom-constrained representation learning*, where EmEL++ (Mondal et al., 2021) uses symbolic 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 derive approximate instance sets for instance retrieval and downstream concept learning. Its performance is evaluated against established symbolic reasoners used as state-of-the-art baselines.

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 systems: Neural Processing in a Symbolic I/O Pipeline.

System	Dataset(s)	Task(s)	Metric(s)
<i>Neural classification over symbolic templates</i>			
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
<i>Neural approximation of deductive closure</i>			

System	Dataset(s)	Task(s)	Metric(s)
CFR (ChunfyReasoner) (Zhu et al., 2023)	Family; Time	Subsumption; Membership; Link Prediction	Precision; Recall; F1; Runtime
Ontology-informed representation learning			
BERTSubs (Subsumption prediction method) (Chen et al., 2023)	NCIT-DOID; HeLiS; FoodOn	Subsumption	MRR; Hits@K

System	Dataset(s)	Task(s)	Metric(s)
<i>Axiom-constrained representation learning</i>			
EBR (Embedding Based Reasoner) (Kamdern Teyou et al., 2025)	Carcinogenesis; Mutagenesis; Semantic Bible; Vicodi; Family; Father	Instance Retrieval; Robust Reasoning under Incomplete- ness/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 , or rule-based deduction, or probabilistic logic inference) remains responsible for the overall reasoning and for producing the final output) provides the primary reasoning structure or transforms the knowledge on which subsequent prediction operates. Neural components are integrated as auxiliary modules that *guide* guide, parameterize, or consume the output of this symbolic process, for example by proposing candidates, scoring alternatives, learning rule parameters, or prioritizing search decisions or making predictions over symbolically enriched representations. 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 symbolic reasoning remains an explicit upstream or controlling component of the reasoning pipeline rather than being internalized within the neural architecture.

In our corpus, one *theme* integration mechanism 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

A second integration mechanism 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

A third integration mechanism is *neural generation paired with symbolic validation* for knowledge graph enrichment and ontology evolution, i.e., the controlled update or extension of an ontology to incorporate new or changed knowledge: 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 *integration mechanism is neural parameterization and constraint-based filtering in prediction with symbolic inference* constraint filtering. DPLogic (Li et al., 2025) and DiffLogic (Shengyuan et al., 2023) retain symbolic probabilistic

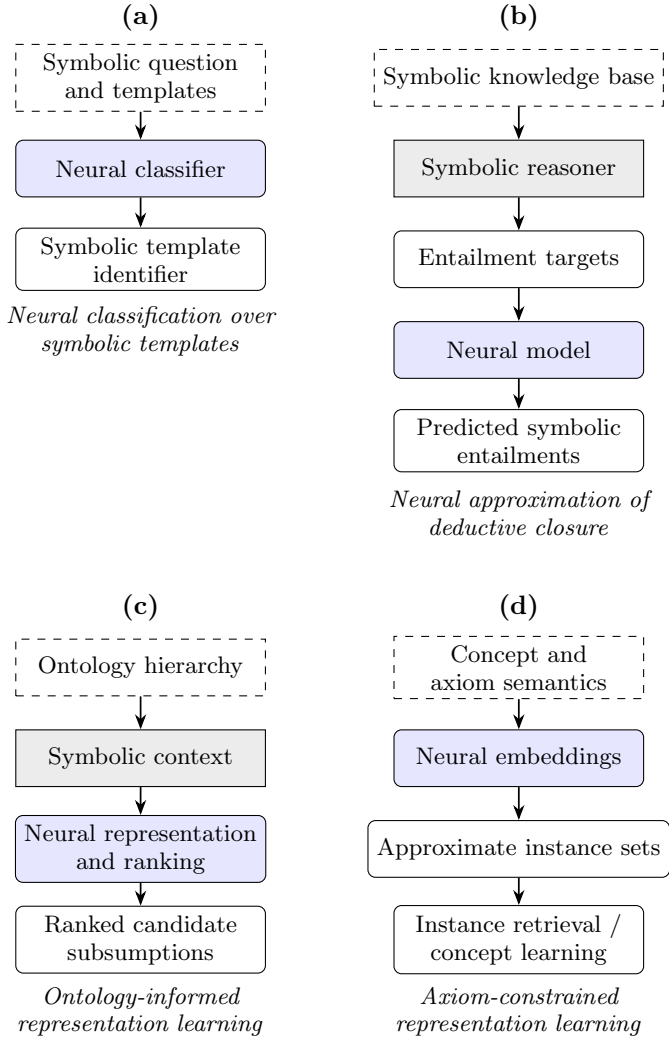


Figure 4. Schematic representation of the integration mechanisms observed within Type 1 systems (Neural Processing in a Symbolic I/O Pipeline): (a) neural classification over symbolic templates, (b) neural approximation of deductive closure, (c) ontology-informed representation learning, and (d) axiom-constrained representation learning. Rounded blue boxes denote neural components, gray boxes denote symbolic components or processes, dashed boxes denote structured symbolic knowledge or input, and arrows indicate the principal direction of information flow.

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 using 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 support, 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 systems: Symbolic Reasoning with Neural Guidance.

System	Dataset(s)	Task(s)	Metric(s)
<i>Neural guidance for symbolic query/program execution</i>			
NS-KGQA (Agarwal and Bedathur, 2025)	KQA Pro; MetaQA; WebQSP	Knowledge Graph Question Answering	Accuracy; Hits@K; F1
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
<i>Symbolic enrichment followed by neural prediction</i>			
A neuro-symbolic link prediction system (Rivas et al., 2024)	Lung cancer polypharmacy treatment KG (clinical records + DrugBank DDIs)	Link Prediction	Precision; Recall; F1
<i>Neural generation paired with symbolic validation</i>			
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
<i>Neural prediction with symbolic constraint filtering</i>			
KOSMOS (Knowledge Oriented Symbolic learning for Medical Ontology-based decision System) (Purohit et al., 2025b)	Lung cancer KG	Link Prediction	MRR; Hits@K

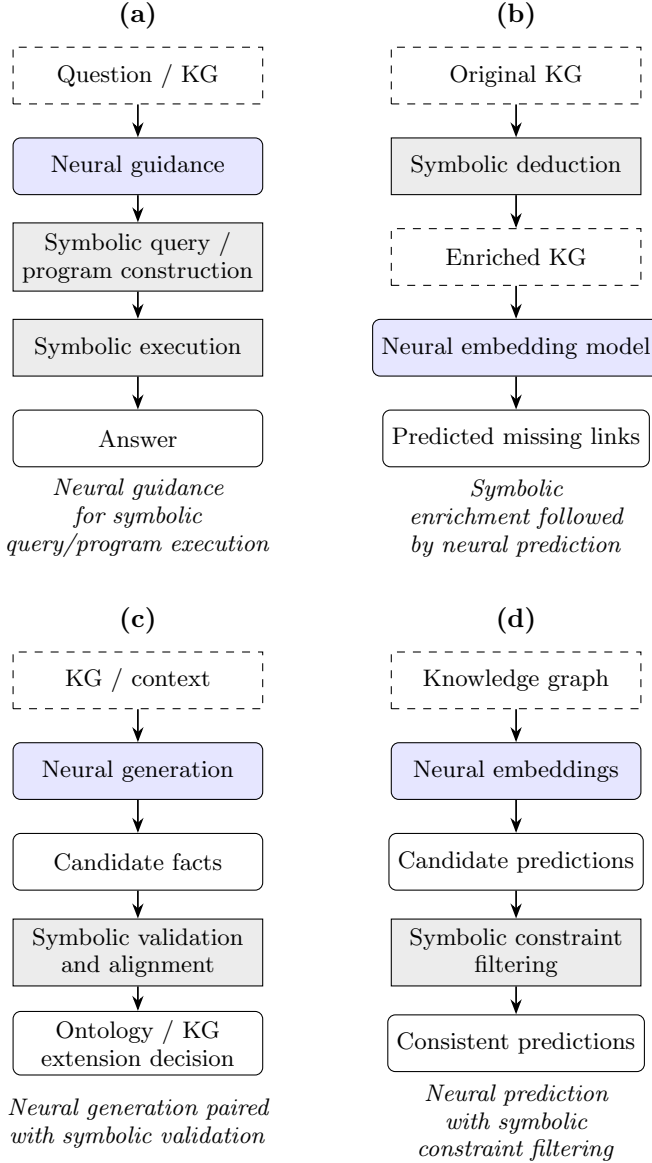


Figure 5. Schematic representation of the integration mechanisms observed within Type 2 systems (Symbolic Reasoning with Neural Guidance): (a) neural guidance for symbolic query/program execution, (b) symbolic enrichment followed by neural prediction, (c) neural generation paired with symbolic validation, and (d) neural prediction with symbolic constraint filtering.

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 or inferred information. The two components may contribute complementary predictions or interact iteratively, with outputs from one component informing subsequent computation in the other. This preserves distinct neural and symbolic reasoning mechanisms while enabling bidirectional information exchange. 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 *integration mechanism is parallel neural scoring and symbolic rule reasoning for knowledge graph completion* *iterative neural-symbolic coupling*. 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. embedding-learning rounds. DPLoGic (Li et al., 2025) and DiffLoGic (Shengyuan et al., 2023) likewise combine knowledge graph embeddings with probabilistic logical reasoning, with the neural and symbolic components exchanging information during iterative optimization rather than operating in a one-way pipeline. 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 *integration mechanism is alignment as modular candidate generation, with neural refinement, and symbolic consolidation* *refinement*. 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* *iterative neural projection and symbolic refinement*: 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 distinct neural prediction modules and symbolic refinement reasoning 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.

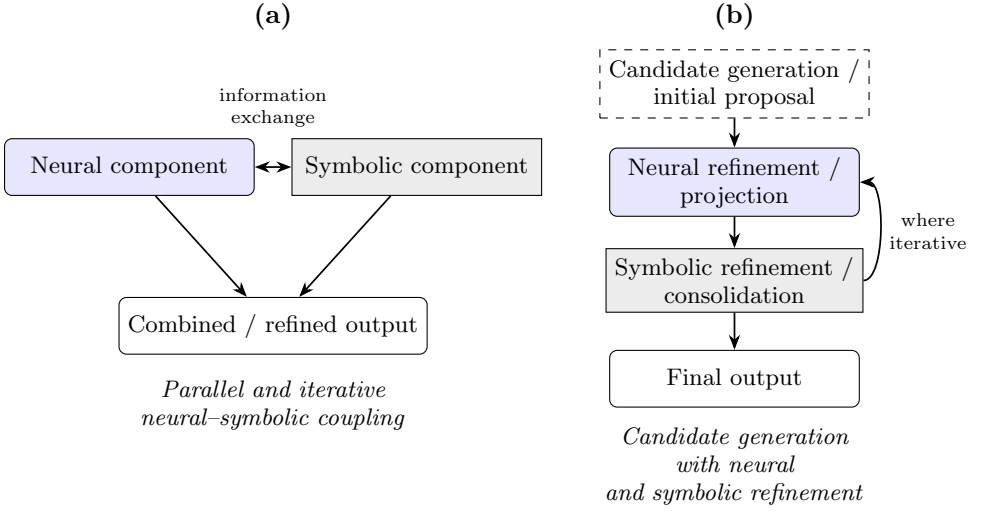


Figure 6. Schematic representation of the integration mechanisms observed within Type 3 systems (Parallel Neuro-Symbolic Systems): (a) parallel and iterative neural-symbolic coupling and (b) candidate generation with neural and symbolic refinement. Feedback arrows indicate iterative interaction where applicable rather than a mandatory processing step for every system.

Table 4. Type 3 neurosymbolic systems: Parallel Neuro-Symbolic Systems.

System	Dataset(s)	Task(s)	Metric(s)
<i>Parallel and iterative neural-symbolic coupling</i>			
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 (%)
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
Poderoso (Martinez Lorenzo et al., 2025)	LUBM; DBpedia20k	Link Prediction	MRR; Runtime
<i>Candidate generation with neural and symbolic refinement</i>			
NeSyMatch (Sharma and Jain, 2025)	OAEI Bio-ML 2023	Ontology Alignment	Precision; Recall; F1; MRR; Hits@K
NeuSymEA (Chen et al., 2025)	DBP15K; OpenEA; DBP1M	Entity Alignment	Hits@K; MRR
ENeSy (Xu et al., 2022)	FB15k-237; NELL-995	Complex Query Answering	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.

The Type 4 systems were organized around five mechanisms through which symbolic knowledge shapes neural learning: constraint-driven grounding and revision,

deduction-guided supervision and evaluation, logic-regularized representation learning, symbolic guidance of neural reasoning policies, and symbolic knowledge injection into representation learning.

In our corpus, one **theme integration mechanism** 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 *ACC* constraints, using the ontology as a regularizer of the final grounding. A second theme is

A second integration mechanism is *using deductive closure deduction-guided supervision 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 **integration mechanism is logic-regularized neural representations for complex query answering representation learning**. 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

A fourth integration mechanism uses symbolic structure as *training symbolic guidance for multi-hop of neural reasoning agents policies*. 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 A fifth integration mechanism is *ontology-level prediction and alignment symbolic knowledge injection into representation learning*, 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 represents entities as probabilistic axis-aligned boxes and relations as affine transformations over the head and tail boxes. The model learns uncertain knowledge graph representations for confidence prediction and fact ranking, while relational constraints are incorporated into the geometric representation to promote 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 systems: Symbolic Constraints for Neural Learning.

System	Dataset(s)	Task(s)	Metric(s)
<i>Constraint-driven grounding and revision under description logic</i>			
EmALC (Wu and Zhao, 2025)	Ontodm; Nifdys; Nihss; GlycoRDF; Sso; Family; Family2	Masked ABox Revision; Complex Query Answering	Success Rate; KL Divergence; Precision; Recall
<i>Deduction-guided supervision and evaluation</i>			
CoPCA (Purohit et al., 2025a)	DB100K; YAGO3-10; French Royalty	Link Prediction	MRR; Hits@K
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
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
<i>Logic-regularized representation learning</i>			
DAGE (He et al., 2025)	FB15k-237; FB15k; NELL	Complex Query Answering	MRR; Runtime

System	Dataset(s)	Task(s)	Metric(s)
<i>Symbolic guidance of neural reasoning policies</i>			
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
RRN (Recursive Reasoning Network) (Hohenecker and Lukasiewicz, 2020)	Family Trees; Countries; DBpedia; Claros; UMLS	Membership	Accuracy; F1
<i>Symbolic knowledge injection into representation learning</i>			
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
KGTORe (Mancino et al., 2023)	MovieLens 1M; Yahoo! Movies; Facebook Books	Recommendation	nDCG; Precision; Hits@K; Recall
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 One prominent integration mechanism observed in the corpus is neural execution of symbolic queries via learned operators.

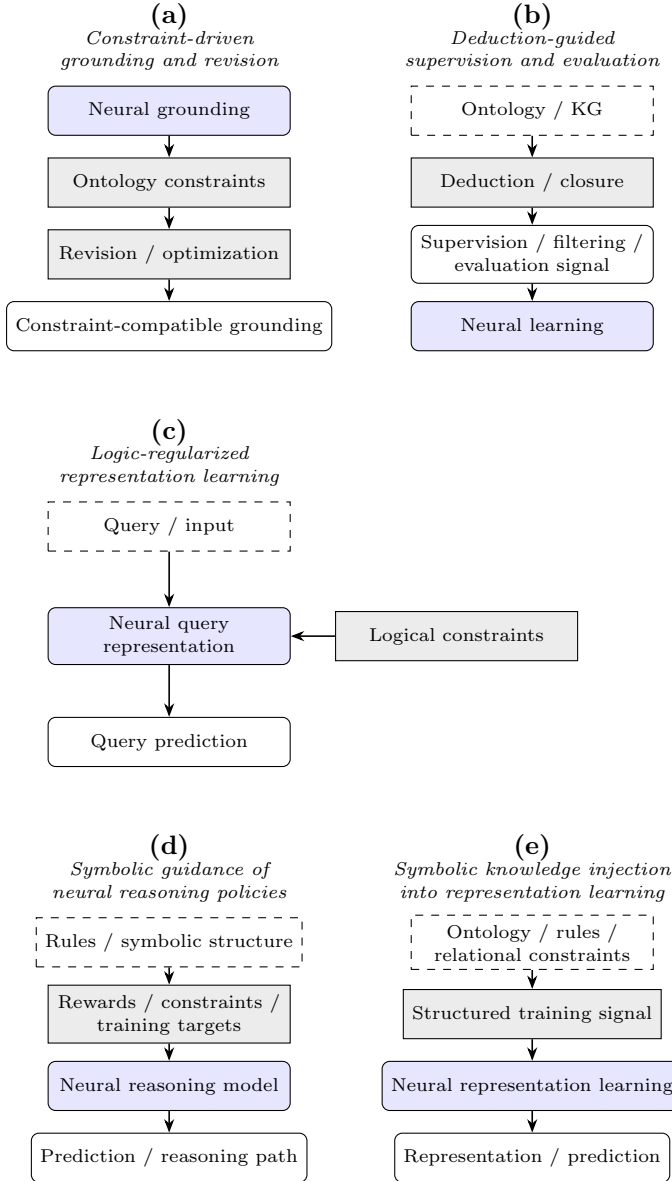


Figure 7. Schematic representation of the mechanisms through which symbolic knowledge shapes neural learning in Type 4 systems (Symbolic Constraints for Neural Learning): (a) constraint-driven grounding and revision, (b) deduction-guided supervision and evaluation, (c) logic-regularized representation learning, (d) symbolic guidance of neural reasoning policies, and (e) symbolic knowledge injection into representation learning.

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). FuzzyVis (Zhurov et al., 2025) follows a related query-oriented integration pattern: users compose symbolic fuzzy concepts using ontology constructors, while a fuzzy membership-based embedding layer retrieves approximate top- k matches and tracks satisfaction or violations of ontology axioms.

A second theme is integration mechanism is ontology semantics embedded as differentiable constraints for completion and approximate deduction representation learning. 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), EmEL++ (Mondal et al., 2021), 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 learning representations that preserve aspects of ontology semantics and support the ranking of candidate axioms. In particular, EmEL++ extends the geometric embedding formulation to additional $\mathcal{EL}^{++}$ constructs, including role hierarchies and role chains. The box-embedding approach of (Memariani et al., 2025) follows a closely related geometric strategy: molecular entities are represented as vectors derived from SMILES strings, while ontology classes are represented as boxes, from which symbolic relations such as subsumption, disjointness, and overlap are predicted. LNN-MP and LNN-CM (Sen et al., 2021) instantiate a related idea form of integration by coupling neural representations with logic-inspired mixture /and chain constructions and training them to satisfy closure-style under closure-oriented logical constraints.

A third theme integration mechanism 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 a fourth integration mechanism is explainability neuralized logical reasoning and transfer rule-aware

prediction. 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 systems: Neural Architectures with Embedded Symbolic Reasoning.

System	Dataset(s)	Task(s)	Metric(s)
<i>Neural execution of symbolic queries via learned operators</i>			
ULTRAQUERY (Galkin et al., 2024)	FB15k-237; NELL-995; FB15k; WikiTopics-QA	Complex Query Answering	MRR; Hits@K; AUC-ROC; Mean Absolute Percentage Error
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)
CLMPT (Conditional Logical Message Passing Transformer) (Zhang et al., 2024)	FB15k; FB15k-237; NELL-995	Complex Query Answering	MRR; Hits@K
NewLook (Liu et al., 2021a)	FB15k; FB15k-237; NELL	Complex Query Answering	Hits@K; MRR
Query2Box (Ren et al., 2020)	FB15k; FB15k-237; NELL-995	Complex Query Answering	MRR; Hits@K

System	Dataset(s)	Task(s)	Metric(s)
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
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
<i>Ontology semantics embedded as differentiable constraints for representation learning</i>			
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
ELEm ($\mathcal{EL}$ Embeddings) (Kulmanov et al., 2019)	GO+STRING	Link Prediction (PPI)	Hits@K; Mean Rank; AUC-ROC
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
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
EmELvar (Mohapatra et al., 2021)	ORE; OWL2Bench	Subsumption	Hits@K; Median Rank; Percentile Rank
TransBox (Yang et al., 2025)	GALEN; Gene Ontology (GO); Anatomy	Link Prediction; Subsumption	Hits@K; Median Rank; MRR; Mean Rank; AUC-ROC

System	Dataset(s)	Task(s)	Metric(s)
A box-embedding approach (Memariani et al., 2025)	ChEBI	Subsumption; Disjointness; Overlap; Hierarchical Multi-label Classification	Precision; Recall; F1
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
<i>Learning rules and symbolic functions as internal neural structure</i>			
DegreEmbed (Li et al., 2023)	FB15k-237; WN18; UMLS; Kinship; Family	Link Prediction; Rule Learning	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
<i>Neutralized logical reasoning and rule-aware prediction</i>			
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
KG-LRR (Knowledge Graphs-based Logic Reasoning Recommendation) (Wang et al., 2025)	Yelp2018; Amazon-book; Amazon-electronics	Recommendation (Explainable)	Precision; Recall; nDCG
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

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

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 *integration mechanism 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 *Neuralsymbolic Commonsense Reasoner with Relation Predictors proposed in (Moghimi 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

A second integration mechanism is neural prediction coupled with symbolic constraint checking and explanation constraint-based inference. 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

A third integration mechanism is structured logical query answering structure combined with symbolic query materialization neural representations. 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. *uses an inductive neural entity encoder to represent unseen entities and a fixed logical query language together with a non-parametric fuzzy/logical executor to compute query answers from the learned scores.*

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.

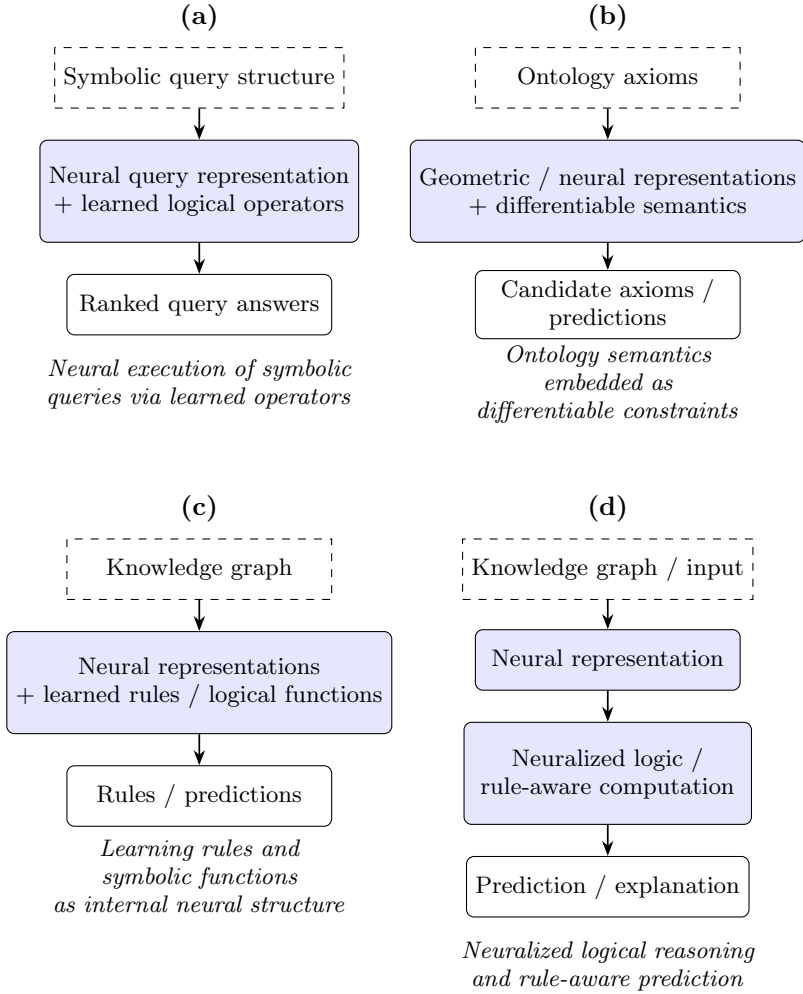


Figure 8. Schematic representation of the integration mechanisms observed within Type 5 systems (Neural Architectures with Embedded Symbolic Reasoning): (a) neural execution of symbolic queries via learned operators, (b) ontology semantics embedded as differentiable constraints for representation learning, (c) learning rules and symbolic functions as internal neural structure, and (d) neuralized logical reasoning and rule-aware prediction.

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).

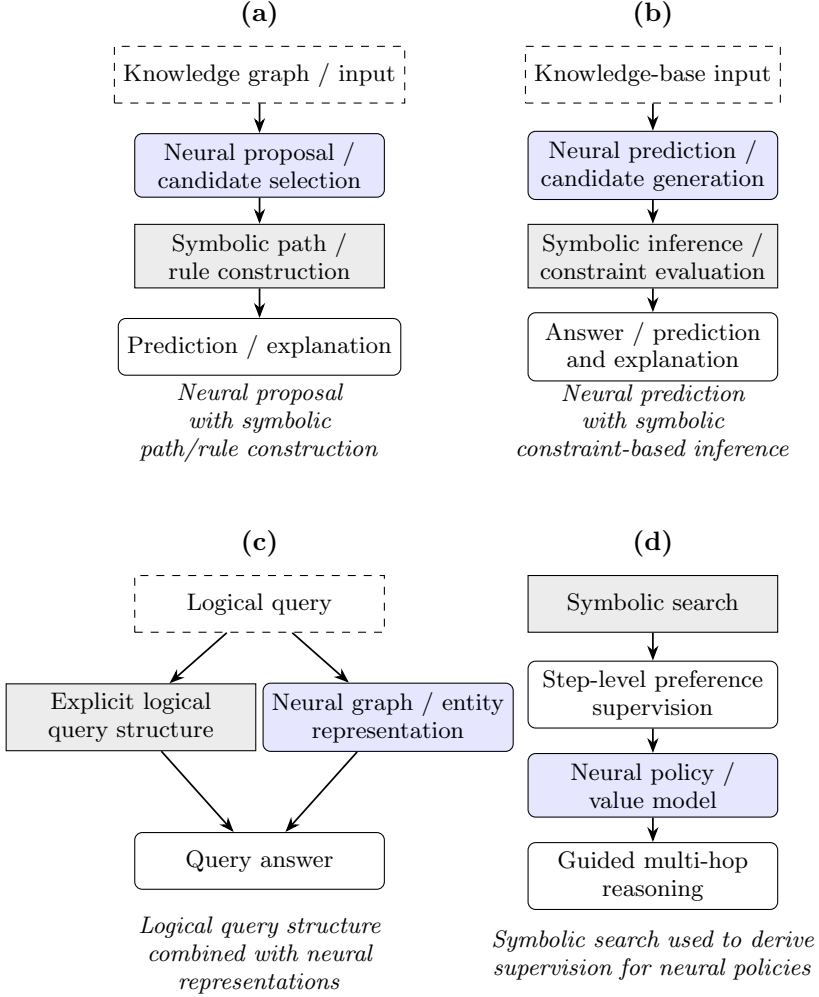


Figure 9. Schematic representation of the integration mechanisms observed within Type 6 systems (Neural Networks with Logical Reasoning Layers): (a) neural proposal with symbolic path/rule construction, (b) neural prediction with symbolic constraint-based inference, (c) logical query structure combined with neural representations, and (d) symbolic search used to derive supervision for neural policies.

Table 7. Type 6 neurosymbolic systems: Neural Networks with Logical Reasoning Layers.

System	Dataset(s)	Task(s)	Metric(s)
<i>Neural proposal with symbolic path/rule construction</i>			
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
Commonsense reasoner (relation prediction) (Moghimifar et al., 2021)	ATOMIC; ConceptNet-100K	Link Prediction; Multi-hop Reasoning; Rule-Path Reasoning	MRR; Hits@K
<i>Neural prediction with symbolic constraint-based inference</i>			
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
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)
<i>Logical query structure combined with neural representations</i>			
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

System	Dataset(s)	Task(s)	Metric(s)
Symbolic search used to derive supervision for neural policies			
FD-PORT (Flow-guided Direct Preference Optimization for knowledge graph reasoning with trees) (Shen et al., 2025)	WebQSP; Com- plexWebQuestions; MetaQA	Knowledge Graph Question Answering; Multi-hop Reasoning	Hits@K

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 systems (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 Within the reviewed corpus, neurosymbolic evaluation spans both historically embedding-oriented KG benchmarks and ontology-centric resources with stronger formal semantics, rather than a single standardized benchmark suite. We therefore organize the dataset landscape according to dataset type and domain, recurrence across the reviewed studies, and the reasoning tasks for which the resources are used.

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 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 Their repeated use provides established experimental settings that facilitate comparison

across neurosymbolic systems. 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).

Many of the remaining datasets occur in only one of the reviewed studies. These resources are typically domain-specific, task-specific, or introduced or adapted for the particular system being evaluated, rather than established as general-purpose neurosymbolic benchmarks. Their limited reuse in the reviewed corpus therefore primarily reflects the fragmentation and specialization of current evaluation practices. This also limits direct comparison across systems, as approaches evaluated on application-specific resources cannot readily be compared with methods assessed on widely reused benchmarks.

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. In contrast to conventional KG completion benchmarks, these resources provide explicit ontological semantics and richer logical structures that can support both deductive and predictive evaluation settings. Their use in the reviewed corpus therefore does not correspond to a strict separation between deductive ontology reasoning and predictive knowledge graph reasoning; rather, ontology-based resources are also used for predictive tasks 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 link prediction.

Ontology-oriented benchmark suites are less frequently reused in the reviewed corpus but provide complementary evaluation settings. LUBM models the university domain through classes and relations describing entities such as universities, departments, students, faculty, and courses, together with synthetically generated instance data. In the reviewed systems, LUBM is used by Poderoso (Martinez Lorenzo et al., 2025) for link prediction. ORE provides collections of naturally occurring OWL ontologies originally assembled for the OWL Reasoner Evaluation initiative. In the reviewed corpus, ORE is used by EmELvar (Mohapatra et al., 2021) for subsumption prediction. OWL2Bench is a configurable university-domain ontology benchmark derived from UOBM, with TBoxes corresponding to the OWL 2 EL, QL, RL, and DL profiles, an ABox generator, and reasoning-oriented SPARQL queries. It is likewise used by EmELvar (Mohapatra et al., 2021) for subsumption prediction. These uses illustrate that ontology benchmark suites originally developed around symbolic reasoning can also serve predictive neurosymbolic evaluations.

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 tasks such as link prediction assess generalization over observed graph patterns, whereas ontology reasoning tasks emphasize semantic faithfulness to evaluations involving explicit ontology semantics can additionally assess whether predictions respect 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	Resource type / domain	DL expressivity	Main evaluation task families	Studies per task family (n)
FB15k-237	Knowledge graph benchmark; general-purpose (Freebase-derived)	–	Complex query answering; link prediction/KG completion; multi-hop and path reasoning; rule learning; node classification	10; 6; 3; 3; 1
Family	Ontology benchmark; family relations	–	Ontology reasoning, retrieval, learning, and repair; link prediction; complex query answering; rule learning; robust reasoning	3; 5; 1; 1; 1
FB15k	Knowledge graph benchmark; general-purpose (Freebase-derived)	–	Complex query answering; link prediction/KG completion; rule learning	6; 1; 1

Dataset	Resource type / domain	DL expressivity	Main evaluation task families	Studies per task family (n)
Gene Ontology (GO)	Ontology; biomedicine (gene function)	$\mathcal{SRI}$	Ontology reasoning and completion; link prediction, including protein–protein interaction prediction; drug repurposing	6; 5; 1
NELL-995	Knowledge graph benchmark; general-purpose (NELL-derived)	–	Complex query answering; link prediction; multi-hop and path reasoning	6; 1; 2
WN18RR	Knowledge graph benchmark; lexical-semantic (WordNet-derived)	–	Link prediction; complex query answering; multi-hop and path reasoning; probabilistic logic and rule learning	6; 1; 2; 3
FoodOn	Ontology; food	$\mathcal{SRIQ}$	Ontology reasoning and completion; link prediction, including protein–protein interaction prediction	4; 2
GALEN	Ontology; biomedicine (clinical/anatomical knowledge)	$\mathcal{SHF}$ (small) / $\mathcal{SHIF}$ (full)	Ontology reasoning and subsumption; link prediction	4; 3
GO+STRING	Integrated ontology/KG benchmark; biomedicine (gene function and protein interactions)	$\mathcal{EL}^{++}$	Ontology reasoning and subsumption; link prediction, particularly protein–protein interaction prediction	2; 4

Dataset	Resource type / domain	DL expressivity	Main evaluation task families	Studies per task family (n)
Kinship	Knowledge graph benchmark; kinship/family relations	–	Link prediction; probabilistic logic reasoning; rule learning	4; 1; 2
UMLS	Biomedical terminology/ontology resource; biomedicine	–	Link prediction; membership prediction; rule learning	3; 1; 2
Anatomy	Ontology benchmark; biomedicine (anatomy)	<i>SRIQ</i> (Uberon)	Ontology reasoning and subsumption; link prediction	3; 2
HPO	Ontology; biomedicine (human phenotypes)	<i>AL</i>	Approximate query answering and semantic search; biomedical link prediction and drug repurposing	1; 1
LC-QuAD 1.0	KGQA benchmark; general-purpose (DBpedia-based)	–	KG question answering; semantic parsing and template classification; complex query answering	3; 2; 1
NELL	Knowledge graph; general-purpose (web-extracted knowledge)	–	Complex query answering; multi-hop reasoning	3; 1
WebQSP	KGQA benchmark; general-purpose (Freebase-based)	–	KG question answering; semantic parsing and template classification; multi-hop reasoning	3; 1; 1
WN18	Knowledge graph benchmark; lexical-semantic (WordNet-derived)	–	Link prediction; probabilistic logic reasoning; rule learning	3; 1; 2

Dataset	Resource type / domain	DL expressivity	Main evaluation task families	Studies per task family (n)
YAGO3-10	Knowledge graph benchmark; general-purpose (YAGO-derived)	–	Link prediction; probabilistic logic reasoning; rule injection	3; 1; 1
ComplexWeb-Questions	KGQA benchmark; general-purpose (Freebase-based)	–	KG question answering; semantic parsing and template classification; multi-hop reasoning	2; 1; 1
Countries	Knowledge graph benchmark; geopolitical relations	–	Explainable link prediction; membership prediction	1; 1
DBpedia	Knowledge graph; general-purpose (Wikipedia-derived)	–	KG and complex query answering; membership prediction	1; 1
HeLiS	Ontology/KG resource; health and lifestyle	$\mathcal{ALCHI}Q(\mathcal{D})$	Ontology reasoning and completion, including membership and subsumption	2
LC-QuAD 2.0	KGQA benchmark; general-purpose (DBpedia / Wikidata-based)	–	KG question answering; semantic parsing; template classification	2; 2; 2
MetaQA	KGQA benchmark; movies	–	KG question answering; multi-hop reasoning	2; 1
OAEI Bio-ML 2023	Ontology-matching benchmark; biomedicine	Varies by ontology pair	Ontology alignment	2
QALD-9	KGQA/semantic-parsing benchmark; general-purpose	–	KG and complex query answering; semantic parsing; template classification	2; 1; 1

Note 1: DL expressivity is reported for ontology datasets and ontology-based knowledge bases. “–” denotes datasets that are not represented as DL ontologies.

For GALEN, the established small and full ontology variants have different DL expressivities. OAEI Bio-ML 2023 comprises multiple source ontologies and therefore has no single dataset-level DL expressivity.

Note 2: The values in “Studies per task family (n)” correspond, in order, to the task families listed in the preceding column and count distinct studies. Task families are non-exclusive; a study may contribute to more than one family. “—” denotes resources for which DL expressivity is not applicable.

Tasks and Metrics (RQ 2.2). The reported metrics are not uniformly applicable across all reasoning tasks. Some are task-specific, whereas others are general predictive measures used across several task families. Ranking metrics such as MRR, Mean Rank, and Hits@(*K*) are primarily used for link prediction, knowledge graph completion, axiom ranking, and related retrieval tasks. Classification metrics such as accuracy, precision, recall, F1-score, and AUC are applied more broadly to binary or multi-class prediction tasks. Similarity measures such as Jaccard similarity are commonly used for set-valued outputs, including instance retrieval and concept learning.

To characterize evaluation practice in the reviewed literature, we extracted all reported evaluation metrics and computed their frequencies across studies. We then report the the 63 included studies. The ten most frequently used metrics in ontology- and knowledge-graph-centric evaluations of neurosymbolic reasoning systems reported metrics were Hits@*K* (40 studies), MRR (31), Precision (16), Recall (16), F1-score (14), AUC-ROC (14), Mean Rank (9), Accuracy (7), Runtime (7), and nDCG (5). Their interpretation is summarized below:

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.

Because the overall frequency of a metric is influenced by the prevalence of the corresponding task, we additionally examined metric usage within the most frequently

represented task categories. Among the 28 studies evaluating link prediction, including task variants such as protein–protein interaction and explainable link prediction, 22 reported Hits@K and 17 reported MRR. For subsumption prediction, Hits@K was reported in 8 of 10 studies and MRR in 4 of 10. Among the 13 studies evaluating complex query answering, including inductive and negation-aware variants, 8 reported MRR and 6 reported Hits@K. For multi-hop reasoning, both Hits@K and MRR were reported in 5 of 6 studies. In knowledge graph question answering, Precision and Recall were each reported in 3 of 5 studies, while Accuracy, F1-score, and Hits@K were each reported in 2 of 5 studies. These task-conditioned frequencies illustrate that the prevalence of individual metrics should be interpreted in relation to the reasoning task rather than as a task-independent measure of evaluation practice.

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.

These metrics capture only selected aspects of system performance. Ranking metrics such as MRR and Hits@K indicate whether correct answers receive high scores, but they do not establish that predictions are logically sound, consistent with the ontology, or supported by a valid reasoning process. Accuracy and F1-score may obscure performance differences across classes and can be misleading on imbalanced datasets. Set-overlap measures evaluate agreement with reference answers but do not distinguish between logically justified and unsupported predictions. Moreover, runtime results are difficult to compare when studies use different hardware, implementations, and dataset sizes. A comprehensive evaluation of neurosymbolic reasoning systems should therefore combine predictive metrics with measures of logical correctness, consistency, robustness, explanation quality, and computational efficiency.

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. 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.

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

A third cluster addressed structured reasoning over learned representations (G3 Structured reasoning over learned representations), including complex query answering and , multi-hop reasoning, and membership (realization). In complex query answering and multi-hop reasoning: neural components learned query operators, neural components learn query operators, representations, or traversal policies, while symbolic components provided query semantics and constraint checking. Ontology-centric structure induction was also prominent. In subsumption constraints on the reasoning process. For membership, neural models score whether an individual belongs to a class, while ontology axioms or rules define the target membership relation and may support entailment or consistency checking.

G4 (Symbolic structure induction) groups subsumption prediction and rule learning (G4 Symbolic structure induction), neural models ranked , where neural models rank 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. are used to assess their validity through coherence or constraint-violation checks.

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).

G5, *Constrained Search, Repair & Evolution*, covers tasks in which symbolic constraints guide the correction, completion, or adaptation of knowledge under incompleteness, noise, or inconsistency.

G6, *Explainable & Auditable Inference*, covers tasks that produce human-interpretable reasoning evidence, such as rules, proof traces, or graph paths, and may assess the coherence or faithfulness of these explanations.

We include this mapping in the appendix because these task types were least commonly adopted in the reviewed neurosymbolic reasoners systems, 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			
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

Task	Supported metric(s)	Neuro role	Symbolic role
(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)
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

Task	Supported metric(s)	Neuro role	Symbolic role
(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 systems exhibit several limitations. First, the benchmark landscape is fragmented across two partially distinct traditions: widely adopted KG completion datasets(e.g., Freebase- several partially overlapping traditions, including widely adopted knowledge graph completion datasets, logical query answering benchmarks, and WordNet-derived benchmarks) and ontology-centric OWL/DL benchmarks (e.g., resources ranging from individual domain ontologies to benchmark suites such as 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.

The limitations summarized in Table 10 summarizes these limitations and their consequences were synthesized from the full set of included studies, with Query 4 specifically supporting the identification of literature focused on benchmark and dataset limitations.

Table 10. Benchmark limitations and their consequences.

Limitation	Consequence
Fragmentation across partially overlapping benchmark families (e.g., KG completion, logical/complex query answering, domain ontologies, and OWL/DL benchmark suites)	Limited cross-study comparability; gains may not transfer across representations, tasks, or semantic settings
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.

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

Dimensions of Neurosymbolic Evaluation

To address these limitations, we proposed a five-dimensional evaluation framework, as illustrated in Figure 10, encompassing: . The Figure synthesizes the evaluation practices observed across the reviewed literature, we organize them along five complementary dimensions: dataset characteristics, reasoning capabilities, evaluation protocols, evaluation metrics, and robustness and generalization. These dimensions capture distinct aspects of how neurosymbolic systems are evaluated and provide a common structure for comparing evaluation practices across otherwise heterogeneous systems and benchmarks.

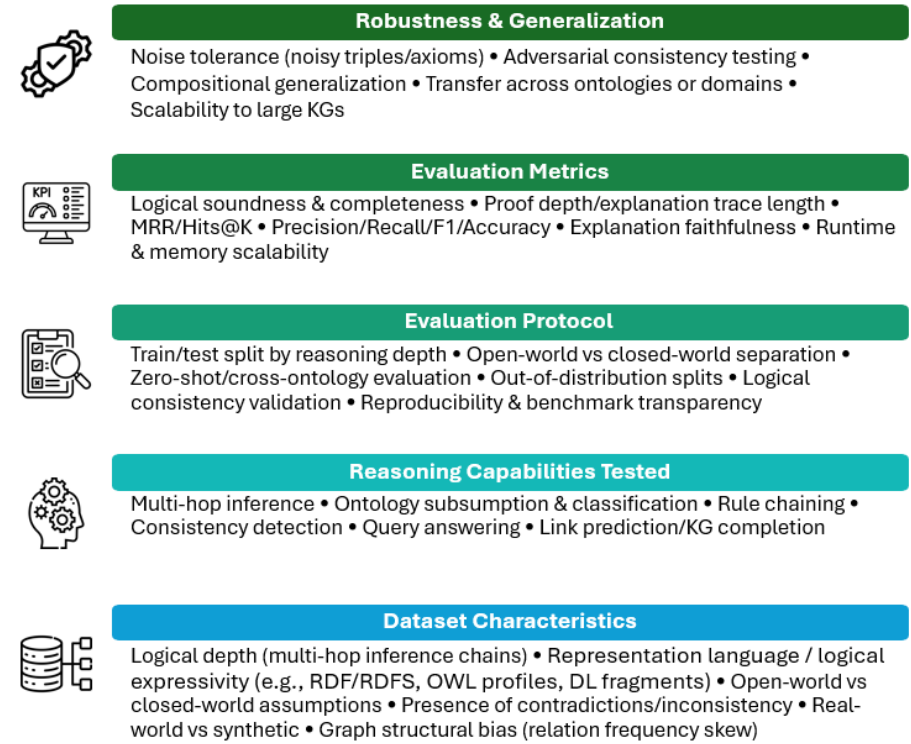


Figure 10. A five-dimensional framework for evaluating ontology- and knowledge graph-based neurosymbolic systems.

The five dimensions provide a general framework for evaluating neurosymbolic systems over both knowledge graphs and ontologies, but their operationalization should be adapted to the representation and task. Knowledge graph prediction commonly relies on ranking metrics such as MRR and Hits@(*K*), together with filtered evaluation protocols, whereas ontology-oriented reasoning may additionally require

measures of logical soundness and completeness, consistency preservation, supported Description Logic expressivity, reasoning depth, and the correctness of inferred axioms or proof traces.

The framework comprises the following five dimensions:

1. **Dataset Characteristics:** Evaluates whether datasets are suitable for reasoning, including logical depth (multi-hop inference chains), **ontology expressivity** (RDF/OWL representation language and logical expressivity (e.g., RDF/RDFS, OWL profiles, and DL fragments), open- vs closed-world assumptions, presence 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.
MRR and Hits@K are included as representative metrics for ranking-based tasks, whereas precision, recall, F1-score, and accuracy represent commonly used measures for classification, retrieval, and answer-set evaluation. Their inclusion reflects their prevalence in the reviewed literature and does not imply that they are sufficient for all neurosymbolic reasoning tasks.
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.

Discussion

This scoping review has some methodological limitations. Search-result screening was truncated for each database/query because of the large retrieval volume, and title/abstract screening was conducted by a single reviewer. Although multiple databases and an extended terminology query were used, relevant studies may therefore have been missed, particularly work not indexed under the selected neural-symbolic or logic-aware terminology. The review was also restricted to English-language peer-reviewed publications available by the final search dates. The resulting catalogue should consequently be interpreted as a structured map of the retrieved literature rather than an exhaustive census of all neurosymbolic systems.

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 Which neurosymbolic reasoning systems have been developed for predictive or deductive reasoning over ontologies and knowledge graphs, and how can their neural components, symbolic components, knowledge representations, and integration mechanisms be characterized?***

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 neurosymbolic 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 reasonersystems?*** From the 63 included neurosymbolic **reasonersystems**, 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 benchmark families that differ in their knowledge representation, semantic richness, domain, and supported reasoning tasks. General-purpose KG benchmarks are predominantly used for predictive tasks such as link prediction and knowledge graph completion, whereas ontology-centric OWL/DL benchmarks targeting deductive reasoning (e.g., subsumption and consistency). This split resources support both deductive reasoning and predictive tasks such as subsumption prediction and ontology completion. This heterogeneity constrains cross-study comparability and weakens claims about overall progress when systems are assessed within only one benchmark family

or task setting. 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/>.
- Dai Y, Wang S, Xiong NN and Guo W (2020) A survey on knowledge graph embedding: Approaches, applications and benchmarks. *Electronics* 9(5): 750. DOI:10.3390/electronics9050750.
- d’Avila Garcez A, Lamb L and Gabbay D (2020) Neuro-symbolic ai: The 3rd wave. *Communications of the ACM* 63(11): 58–66.
- DeLong LN, Mir RF and Fleuriot JD (2025) Neurosymbolic ai for reasoning over knowledge graphs: A survey. *IEEE Transactions on Neural Networks and Learning Systems* 36(5): 7822–7842. DOI:10.1109/tnnls.2024.3420218. URL <http://dx.doi.org/10.1109/TNNLS.2024.3420218>.
- 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.
- Guo Y, Pan Z and Heflin J (2005) LUBM: A Benchmark for OWL Knowledge Base Systems. *Journal of Web Semantics*. 3(2-3): 158–182.
- 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, Mihindukulasoorya 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, Perfiljev 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>.
- Matentzoglou N and Parsia B (2014) Ore 2014 reasoner competition dataset. DOI: 10.5281/zenodo.10791. URL <https://doi.org/10.5281/zenodo.10791>.
- 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.
- Seeliger A, Pfaff M and Krcmar H (2019) Semantic web technologies for explainable machine learning models: A literature review. In: *Joint Proceedings of the Workshops on Profiling and Searching Data on the Web and Semantic*

- Explainability*, *CEUR Workshop Proceedings*, volume 2465. CEUR-WS.org, pp. 30–45. URL https://ceur-ws.org/Vol-2465/semex_paper1.pdf.
- 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.
- Singh G, Bhatia S and Mutharaju R (2020) *OWL2Bench: A Benchmark for OWL 2 Reasoners*. ISBN 978-3-030-62465-1, pp. 81–96. DOI:10.1007/978-3-030-62466-8_6.
- 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
- Steinmetz N and Sattler KU (2021) What is in the kgqa benchmark datasets? survey on challenges in datasets for question answering on knowledge graphs. pp. 697–706. *Journal on Data Semantics* 10(3): 241–265. DOI:10.1007/s13740-021-00128-9.
- 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.
- van Bekkum M, de Boer M, van Harmelen F, Meyer-Vitali A and ten Teije A (2021) Modular design patterns for hybrid learning and reasoning systems: A taxonomy, patterns and use cases. *Applied Intelligence* 51(9): 6528–6546. DOI: 10.1007/s10489-021-02394-3.
- 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/S08933608023003398>.
- Zamazal O (2020) A survey of ontology benchmarks for semantic web ontology tools. *International Journal on Semantic Web and Information Systems* 16(1): 47–68. DOI:10.4018/IJSWIS.2020010103.
- 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. In: *Neural-Symbolic Learning and Reasoning*. pp. 355–369. 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")
```

Search-scope note. The generic terms “knowledge graph embedding” and “link prediction” were not included as standalone search terms because they retrieve a broad literature on purely subsymbolic methods. Query 6 instead used embedding and reasoning terms with an explicit logical or ontological focus. Studies applying neurosymbolic approaches to link prediction remained eligible when retrieved through the other search queries.

Supplementary Tables

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

Dataset	Resource type / domain	DL expressivity	Task(s)
Biological Interaction Network (Bio)	Graph-query benchmark; biomedicine (biological interactions)	–	Complex Answering (Hamilton et al., 2018) Query

Dataset	Resource type / domain	DL expressivity	Task(s)
Carcinogenesis	Relational / ontology benchmark; biomedicine / toxicology	–	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	Ontology; biomedicine/chemistry (chemical entities)	$SR\mathcal{OIQ}(\mathcal{D})$	Disjointness (Memariani et al., 2025), Hierarchical Multi-label Classification (Memariani et al., 2025), Overlap (Memariani et al., 2025), Subsumption (Memariani et al., 2025)
Claros	RDF / knowledge graph resource; cultural heritage / archaeology	–	Membership (Hohenecker and Lukasiewicz, 2020)
CMHR (Chinese Medical High-order Relational dataset)	Hyper-relational / hypergraph benchmark; medicine	–	Higher-order Relational Reasoning (Bhuyan et al., 2024), Hypergraph Link Prediction (Bhuyan et al., 2024)
CodeX	Knowledge graph completion benchmark; general-purpose (Wikidata / Wikipedia-derived)	–	Link Prediction (Shengyuan et al., 2023), Probabilistic Logic Reasoning (Shengyuan et al., 2023)
Datatourisme KG	Knowledge graph; tourism (France)	–	Constraint Validation (Boulakbech and Wannous, 2025), Knowledge Graph Enrichment (Boulakbech and Wannous, 2025), Ontology Evolution (Boulakbech and Wannous, 2025)
DB100K	Knowledge graph completion benchmark; general-purpose	–	Link Prediction (Purohit et al., 2025a)

Dataset	Resource type / domain	DL expressivity	Task(s)
DBbook	Recommendation dataset with linked-data mappings; books	–	Recommendation (Spillo et al., 2024)
DBP15K	Entity-alignment benchmark; multilingual DBpedia	–	Entity Alignment (Chen et al., 2025)
DBP1M	Large-scale entity-alignment benchmark; multilingual DBpedia	–	Entity Alignment (Chen et al., 2025)
DBpedia20k	Knowledge graph benchmark; general-purpose (DBpedia subset)	–	Link Prediction (Martinez Lorenzo et al., 2025)
Disease Ontology (DO)	Ontology; biomedicine (diseases)	OWL 2 EL	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
DisGeNET	Biomedical knowledge resource / KG; gene–disease associations	–	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
Family Trees	Relational / KG benchmark; family/genealogical relations	–	Membership (Hohenecker and Lukasiewicz, 2020)
Family2	Ontology benchmark; family relations	–	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)
Father	Ontology / relational benchmark; family relations	–	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)

Dataset	Resource type / domain	DL expressivity	Task(s)
FB-AUTO	Hyper-relational knowledge graph benchmark; general-purpose	–	Link Prediction (Abboud et al., 2020), Rule Injection (Abboud et al., 2020)
Food-Biomarker Ontology (FOBI)	Ontology; biomedicine / nutrition (food biomarkers)	$\mathcal{ALC}$	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	Knowledge graph benchmark; genealogy / French royalty	–	Link Prediction (Purohit et al., 2025a)
GlycoRDF	RDF / ontology resource; biomedicine (glycomics)	$\mathcal{ALE}/\mathcal{RDFS}$	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)
Hetionet	Biomedical knowledge graph; biomedicine	–	Link Prediction (Liu et al., 2021b), Multi-hop Reasoning (Liu et al., 2021b)
JF17K	Hyper-relational knowledge graph benchmark; general-purpose	–	Link Prediction (Abboud et al., 2020), Rule Injection (Abboud et al., 2020)
KQA Pro	KGQA benchmark; general-purpose (Wikidata-derived)	–	Knowledge Graph Question Answering (Agarwal and Bedathur, 2025)
LC-QuAD 2.1	KGQA benchmark; general-purpose	–	Knowledge Graph Question Answering (Gomes et al., 2022), Semantic Parsing (Gomes et al., 2022), Template Classification (Gomes et al., 2022)

Dataset	Resource type / domain	DL expressivity	Task(s)
LUBM	Synthetic ontology benchmark; university domain	$\mathcal{AL}\mathcal{E}\mathcal{H}\mathcal{I}(\mathcal{D})$	Link Prediction (Martinez Lorenzo et al., 2025)
Lung cancer KG	Biomedical knowledge graph; oncology	–	Link Prediction (Purohit et al., 2025b)
Lung cancer polypharmacy treatment KG (clinical records + DrugBank DDIs)	Clinical / biomedical knowledge graph; lung cancer and polypharmacy	–	Link Prediction (Rivas et al., 2024)
Mutagenesis	Relational / ontology benchmark; biomedicine / toxicology	–	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	Ontology benchmark; biomedicine (cancer and disease)	$\mathcal{SH}$ (NCIT) / OWL 2 EL (DOID)	Subsumption (Chen et al., 2023)
Neural Reprogramming Ontology (NRO)	Ontology; biomedicine (neural reprogramming)	$\mathcal{ALC}$	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 knowledge graph benchmark; general-purpose (NELL-derived)	–	Confidence Prediction (Chen et al., 2021b), Link Prediction (Chen et al., 2021b)

Dataset	Resource type / domain	DL expressivity	Task(s)
OGBL-WIKIKG2	Knowledge graph completion benchmark; general-purpose (Wikidata-derived)	–	Inductive Answering (Galkin et al., 2022) Complex Query (Galkin et al., 2022)
Ontodm	Ontology; data mining	$SHOIQ(\mathcal{D})$	Complex Answering (Wu and Zhao, 2025), Masked Revision (Wu and Zhao, 2025) Query ABox
OpenEA	Entity-alignment benchmark suite; general-purpose / multilingual KGs	–	Entity Alignment (Chen et al., 2025)
ORE	Ontology-reasoning benchmark collection; cross-domain	Multiple ontologies (OWL 2 DL / EL)	Subsumption (Mohapatra et al., 2021)
OWL2Bench	Synthetic ontology benchmark; university domain	OWL 2 EL / QL / RL / DL	Subsumption (Mohapatra et al., 2021)
OWL-Centric Dataset	Ontology / KG reasoning benchmark; general-purpose (Linked Data-derived)	Multiple OWL ontologies	Deductive Reasoning (Ebrahimi et al., 2021b), Knowledge Graph Entailment (Ebrahimi et al., 2021b)
Pizza ontology	Ontology benchmark; food	$SHOIN$	Approximate Query Answering (Zhurov et al., 2025), Semantic Search (Zhurov et al., 2025)
QALD-8	KGQA / semantic-parsing benchmark; general-purpose	–	Knowledge Graph Question Answering (Vollmers et al., 2021), Semantic Parsing (Vollmers et al., 2021), Template Classification (Vollmers et al., 2021)

Dataset	Resource type / domain	DL expressivity	Task(s)
Semantic Bible	Ontology / relational benchmark; biblical domain	$\mathcal{SHOIN}(\mathcal{D})$	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	Biomedical data resource / KG; drug adverse effects	–	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
SNOMED CT	Clinical terminology / ontology; biomedicine	$\mathcal{ELH}$ (OWL 2 EL)	Subsumption (Mondal et al., 2021)
SportsNELL	Knowledge graph benchmark; sports (NELL-derived)	–	Link Prediction (Abboud et al., 2020), Rule Injection (Abboud et al., 2020)
Sso	Ontology; public health (syndromic surveillance)	$\mathcal{ALCH}(\mathcal{D})$	Complex Query Answering (Wu and Zhao, 2025), Masked ABox Revision (Wu and Zhao, 2025)
STITCH	Biomedical interaction resource / KG; chemical–protein interactions	–	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
STRING	Biomedical interaction network / KG; protein–protein interactions	–	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
SwissProt GO annotations	Protein-annotation dataset; biomedicine (protein function)	–	Drug Repurposing (Alshahrani et al., 2017), Link Prediction (Alshahrani et al., 2017)
Time	Ontology benchmark; temporal relations	$\mathcal{SRI\mathcal{F}}$	Link Prediction (Zhu et al., 2023), Membership (Zhu et al., 2023), Subsumption (Zhu et al., 2023)

Dataset	Resource type / domain	DL expressivity	Task(s)
Vicodi	Ontology / knowledge base; European history	RDFS(DL)	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	Synthetic RDF benchmark; general-purpose	–	Inductive Complex Query Answering (Pflueger et al., 2022), Node Classification (Pflueger et al., 2022)
WikiTopics-QA	Complex-query-answering benchmark; general-purpose, topic-specific Wikidata graphs	–	Complex Query Answering (Galkin et al., 2024)
Yeast dataset	PPI Biological interaction network; biomedicine (yeast protein–protein interactions)	–	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)

Note: DL expressivity is reported only for ontology datasets or ontology benchmark collections. “–” denotes datasets that are not themselves ontology benchmarks. “Not determined” indicates an ontology dataset for which a single canonical DL expressivity could not be established with sufficient confidence. For ontology collections or generators, the applicable OWL profiles or constituent ontology types are reported instead of a single DL expressivity. Expressivity may depend on the ontology release; where applicable, the table reports a documented canonical release.

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 feasibility
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
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

Task	Supported metric(s)	Neuro role	Symbolic role
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
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

Task	Supported metric(s)	Neuro role	Symbolic role
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
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

Task	Supported metric(s)	Neuro role	Symbolic role
(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
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

Task	Supported metric(s)	Neuro role	Symbolic role
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.