NSORN: Designing a Benchmark Dataset for Neurosymbolic Ontology Reasoning with Noise

Julie Loesch [a,*], Gunjan Singh [b], Raghava Mutharaju [b] and Remzi Celebi [a]
[a] *Department of Advanced Computing Sciences, Maastricht University, Netherlands*
E-mails: julie.loesch@maastrichtunviersity.nl, remzi.celebi@maastrichtuniversity.nl
[b] *Knowledgeable Computing and Reasoning Lab, IIIT-Delhi, India*
E-mails: gunjans@iiitd.ac.in, raghava.mutharaju@iiitd.ac.in

Abstract. In the field of neurosymbolic computing, there is a lack of standardized benchmark datasets specifically designed for evaluating neurosymbolic ontology reasoning systems. Currently, no benchmarks or evaluation frameworks have been explicitly developed to assess the robustness of these systems to noise. Thus, this work aims to develop a mechanism for introducing noise into a ontology, particularly focusing on the ABox, and evaluate the performance of existing neurosymbolic reasoners on the commonly used ontologies under varying levels of noise. We developed NSORN (Neurosymbolic Ontology Reasoning with Noise), a framework that consists of three techniques to introduce noise into ontologies: logical, statistical, and random noise. Logical noise uses logical violations of disjoint axioms and domain/range constraints. While random noise corrupts existing triples by replacing either subject or object of a triple with random entity, statistical noise is introduced using Graph Neural Networks to add noisy facts with low-probability scores. We evaluated the performance of existing neurosymbolic reasoners by introducing noise to *OWL2Bench* and *Family* ontologies under these noise types with various levels. The resulting benchmarks were tested on two state-of-the-art neurosymbolic reasoners, *Box2EL* and *OWL2Vec**. We focus on reasoning tasks such as for instance membership and object property assertions to test how these reasoners handle noise. Our main finding is that logical noise creates a more challenging learning case, resulting in a significant decrease in the performance of both Box2EL and OWL2Vec*.

Keywords: Neurosymbolic Artificial Intelligence, Benchmark, Noise Injection, Ontology Reasoning

1. Introduction

Neurosymbolic computing has emerged as a prominent area of Artificial Intelligence in recent years, combining the robust learning capabilities of neural networks with the reasoning capabilities and interpretability of symbolic systems [21, 78]. Symbolic reasoners rely on formal logic, rules, and knowledge bases, such as ontologies to make inferences. They are often reliable and interpretable, offering traceable mechanisms for their inferences. However, they are sensitive to noise and struggle to handle incomplete or ambiguous data. Symbolic reasoners could fail to perform when faced with missing knowledge or errors in their knowledge base. Moreover, their reliance on a large number of predefined rules and axioms limits their scalability [44, 60]. In contrast, neural reasoners leverage deep learning models, which can generalize from large volumes of data, are robust to noise. However, their

[*]Corresponding author. E-mail: julie.loesch@maastrichtunviersity.nl.

primary limitation lies in their lack of interpretability [19] and handling tasks that require explicit logic or when dealing with rare or unseen examples. Neurosymbolic reasoners can address these shortcomings inherent in each paradigm [78]. By integrating symbolic reasoning with neural systems, these reasoners achieve a trade-off between interpretable logical reasoning and the scalable, data-driven capabilities of neural networks [44, 56]. Despite these advantages, neurosymbolic systems face unique challenges, particularly in the incorporation of domain ontologies while ensuring resilience against the noise and uncertainty that characterize real-world data.

Noise in ontologies encompasses various forms of disturbance that can affect their integrity, coherence, and interpretability. [1] et. al., presented a Semantic Web noise taxonomy, which distinguishes between two main categories of noise: TBox noise and ABox noise (i.e., propagable and non-propagable). TBox noise is the type of noise that resides within the ontology, such as in the class hierarchy, or domain and range properties. This type of noise will affect the inference over the entire dataset. While ABox noise is about corrupting an existing triple in an ontology by changing one of the triples' resources. This either changes the inference graph (i.e., propagable noise) or does not have any impact on the inference graph (i.e., non-propagable noise).

This work aims to develop NSORN (Neurosymbolic Ontology Reasoning with Noise), a framework designed to introduce noise into ontologies and create challenging benchmark datasets to test the effectiveness of neurosymbolic reasoners in handling noise. While numerous benchmark datasets exist for various AI tasks, such as image classification (i.e., MNIST [12], CIFAR-10 and CIFAR-100[1]), natural language processing (i.e., GLUE [75]), and reinforcement learning (i.e., OpenAI Gym [7]), there is a notable absence of standardized benchmark datasets specifically tailored for neurosymbolic reasoning, particularly evaluating their noise tolerance. Such a benchmark is essential to advance this field [54]. To the best of our knowledge, no benchmarks or evaluation frameworks have been explicitly designed to assess and compare the noise tolerance of neurosymbolic reasoning systems. Existing neurosymbolic benchmark datasets are predominantly designed to assess the performance of symbolic reasoners [62]. Furthermore, most reasoning systems are evaluated using various publicly available ontologies [5, 66, 68], which do not address the unique challenges of neurosymbolic integration. We developed three techniques to introduce noise into ontologies: logical, statistical, and random noise. Logical noise involves violations of disjoint axioms and domain/range constraints, statistical noise uses Graph Neural Networks (GNNs) to add low-probability links, and random noise corrupts existing triples by replacing either the subject or object of a triple with a random entity.

With this work, we have addressed the following research questions: how to characterize noise in ontologies, how to introduce noise into these structures, and how to evaluate the impact of noise on neurosymbolic reasoners. By exploring these questions, we aim to develop a framework for generating noisy benchmark datasets. This framework will facilitate the assessment of reasoners' robustness and effectiveness in handling noisy data, ultimately advancing the field of neurosymbolic AI [63, 67].

We run conventional reasoners on datasets with varying noise levels to illustrate their limitations in handling different noise, including logical inconsistencies. Subsequently, we evaluated the performance of neurosymbolic reasoners under these conditions. It should also be noted that most previous work has focused on tasks of ontology completion rather than ontology reasoning. The goal of ontology/link completion is to discover plausible relations that complement the original ontology, as was the task performed in the work of [10]. In contrast, our goal is to infer knowledge that logically follows from the given ontology. To achieve this, we adopt a method similar to that of Makni and Hendler [42].

The remainder of the paper is organized as follows: the existing literature on neurosymbolic ontology reasoners and benchmark data sets is reported in Section 2. Section 3 describes the process of designing the benchmark dataset, including noise injection techniques. Section 4 presents the experimental setup. Section 5 shows the results of the experiments, including performance metrics and analysis. Finally, Section 6 discusses the strengths and limitations of the designed benchmark datasets and explores potential extensions or improvements for future research, followed by Section 7 to conclude our work. The source code of the benchmark is available at https://github.com/jloe2911/NoisyBench under MIT License.

[1] https://www.cs.toronto.edu/~kriz/cifar.html

2. Related Work

Neurosymbolic approaches integrate diverse reasoning techniques, resulting in multiple variations in their evaluation. In Section 2.1, we provide a brief overview of neurosymbolic reasoning methods that are used for our experiments, followed by a discussion of most commonly used benchmark datasets in Section 2.2.

2.1. Reasoning Techniques

Henry Kautz, in his AAAI 2020 Robert S. Engelmore Memorial Award Lecture, discussed six categories of neurosymbolic AI systems as the "Future of AI" [34]. To showcase the variety in existing approaches, we categorize the neurosymbolic reasoning methods used in our experiments into one of those categories.

In [10], the authors introduced *OWL2Vec**, which involves converting the symbolic input (i.e., ontologies and RDF graphs) to vectors, giving rise to *Symbolic Neuro Symbolic*. The method leverages random walk and word embedding techniques to encode the semantics of OWL ontologies. Unlike traditional KG embedding methods, *OWL2Vec** considers not only the graph structure but also lexical information and logical constructors inherent in OWL ontologies. This comprehensive approach enables *OWL2Vec** to capture nuanced relationships between concepts, making it suitable for tasks requiring fine-grained reasoning, such as ontology completion and prediction. The empirical evaluation conducted with three real-world datasets, i.e., HeLis [15], FoodOn [14] and Gene Ontology (GO) [2], demonstrates that *OWL2Vec** outperforms the state-of-the-art methods in class membership and class subsumption prediction tasks. This suggests that *OWL2Vec** benefits from incorporating different aspects of ontology semantics, including graph structure, lexical information, and logical constructors.

In [33], the authors proposed a novel ontology embedding method called *Box2EL* for DL EL++. The approach embeds symbolic reasoning inside neural engines, representing symbolic information in geometric or vector spaces and employing neural methods for reasoning tasks, resulting in the *Neuro[Symbolic]* category. Specifically, they addressed the challenge of ontology completion in Description Logic (DL)-based OWL ontologies, which are widely used for knowledge representation. While classical deductive reasoning algorithms offer precise formal semantics for predicting missing facts in an ontology, recent years have seen a rise in interest in inductive reasoning techniques capable of deriving probable facts from an ontology. Inductive reasoning techniques, akin to those used in KG completion, involve learning ontology embeddings in a latent vector space while ensuring adherence to the semantics of the underlying DL. However, existing ontology embedding methods face shortcomings, particularly in faithfully modeling complex relations and role inclusion axioms, such as one-to-many, many-to-one, and many-to-many relations. This approach represents both concepts and roles as boxes (i.e., axis-aligned hyper-rectangles) and models inter-concept relationships using a bumping mechanism. The authors conduct an extensive experimental evaluation, achieving state-of-the-art results across a variety of datasets, i.e., GALEN [55], Gene Ontology (GO) [2] and Anatomy (a.k.a. Uberon) [50], on the tasks of subsumption prediction, role assertion prediction and approximating deductive reasoning.

2.2. Benchmark Datasets

There is a pressing need for standardized benchmark datasets for neurosymbolic reasoners to facilitate fair and consistent comparisons. Precisely, [67] et al., presented an overview of variations in neurosymbolic reasoning and evaluation approaches. Their overview reveals that similar works may differ significantly by employing distinct metrics and datasets to evaluate their contributions. For instance, the works of Makni et al. [42] and Ebrahimi et al. [17] focus on RDFS entailment reasoning, aiming to replicate deductive reasoning processes. However, they adopt different metrics and datasets to assess the effectiveness and performance of their approaches. Such variations in evaluation criteria can lead to diverse insights and perspectives on the contributions within the field.

The existing traditional benchmarks such as LUBM (Lehigh University Benchmark) [26], UOBM (University Ontology Benchmark) [41] and OWL2Bench [64] lack suitability for evaluating neurosymbolic reasoners due to their narrow focus on conventional reasoning tasks. Traditional evaluations of reasoning systems often rely on metrics such as reasoning time, which may not align with the evaluation requirements of neurosymbolic reasoners.

Although the ontologies of these benchmarks, along with those from the OWL Reasoner Evaluation (ORE) Competition [53], can serve as initial datasets for neurosymbolic benchmarks, these datasets fall short of addressing the distinct challenges posed by neurosymbolic reasoning.

To our knowledge, no benchmarks or evaluation frameworks have been designed to evaluate and compare neurosymbolic reasoning systems. Most reasoner evaluations are performed on different publicly available ontologies, including but not restricted to SNOMED CT[2], Gene Ontology (GO) [2] and GALEN [55], as well as other ontologies available in public repositories such as DBpedia [40], YAGO [72], Wikidata [74], Claros[3], NCBO Bioportal[4] and AgroPortal[5]. However, these offer a limited set of ontologies for evaluation, which does not cover the full spectrum of possible scenarios.

3. Methodology

This section outlines the mechanisms used in NSORN (Neurosymbolic Ontology Reasoning with Noise) to introduce noise into ontologies, specifically targeting the ABox, which contains instance-level information. We devised three distinct techniques to introduce noise into an ontology: logical (see Section 3.1), statistical (see Section 3.2) and random noise (see Section 3.3). Each method was designed to simulate a unique form of inconsistency or error, enabling us to assess the performance and robustness of ontology reasoning under various noisy conditions.

1. **Logical Noise:** Logical noise is introduced by violating the formal constraints of the ontology. We implemented two approaches, as they can be easily used to create logical contradictions without altering the TBox of the ontology.
   (a) **Disjoint Axioms:** We introduce noise by asserting relationships or memberships that contradict declared disjoint axioms. This could be done by assigning an individual to two disjoint classes or linking two entities using disjoint object properties.
   (b) **Domain and Range Violations:** We generate noise by asserting relationships where the subject or object falls outside the defined domain or range of an object property. For example, linking an individual from an incompatible class as the subject or object of a property.
2. **Statistical Noise:** This approach leverages Graph Neural Networks to predict relationships within the ontology. Noise is introduced by adding links (triples) with the lowest probability scores, representing the most unlikely relationships. This method simulates errors arising from statistically improbable but plausible assertions.
3. **Random Noise:** Random inconsistencies are introduced by arbitrarily adding or modifying ABox assertions. This approach represents unpredictable errors that could occur in real-world data.

These techniques were specifically chosen to challenge the neurosymbolic reasoner's reasoning capabilities and to evaluate its resilience against varying levels and types of noise. By analyzing reasoning performance under such conditions, we can better understand the robustness and limitations of ontology-based systems.

3.1. Logical Noise

3.1.1. Contradictions based on Disjoint Axioms

This noise injection technique aims to test the robustness of reasoning engines by deliberately introducing contradictions into the ontology, thereby evaluating the system's ability to handle inconsistencies. To introduce ABox noise, particularly within disjoint axioms (i.e., disjoint classes and disjoint object properties), we developed the following approach.

[2] https://bioportal.bioontology.org/ontologies/SNOMEDCT
[3] https://www.clarosnet.org
[4] https://bioportal.bioontology.org/
[5] http://agroportal.lirmm.fr/

1. **Extracting Disjoint Class Axioms:** We first identified all disjoint class axioms from the ontology. A disjoint class axiom, denoted as $DisjointClasses(CE_1...CE_n)$, specifies that all class expressions CE_i $(1 \leqslant i \leqslant n)$ are pairwise disjoint[6]. This indicates that each axiom involves pairs of mutually exclusive classes. Extracted axioms are used to generate noise, which directly challenges the ontology's consistency.
2. **Introducing Noise:** To generate noise, we added k individuals to the ontology, assigning each to two disjoint classes CE_i and CE_j where $i \neq j$. For example, if `Male` and `Female` are disjoint classes, we would add `John rdf:type Male` and `John rdf:type Female`. This contradiction simulates real-world scenarios where data inconsistencies or conflicts occur, allowing us to measure the reasoner's performance under such conditions. The parameter k allows to control over the noise intensity.

Similarly, we extracted all disjoint object properties from the ontology. An object property axiom, denoted as $DisjointObjectProperties(OPE_1...OPE_n)$, asserts that all object property expressions OPE_i $(1 \leqslant i \leqslant n)$ are pairwise disjoint[7].

To further make ontology inconsistency, we added k individuals to the ontology, each possessing two disjoint object properties OPE_i and OPE_j where $i \neq j$. For example, if `like` and `dislike` are two disjoint properties, we would add `Emma likes mathematics` and `Emma dislikes mathematics`. This noise not only tests the reasoner's ability to handle conflicting object properties but also evaluates the scalability and stability of the ontology. By varying k, we can observe how different levels of noise affect the reasoning performance, providing insights into the system's resilience and accuracy.

3.1.2. Contradictions based on Range/Domain

Object properties in ontologies can have explicitly defined domains and ranges, which establish the types of individuals that are allowed to participate in a relationship. The domain specifies the class of individuals that can serve as the subject of the object property, while the range specifies the class of individuals that can serve as the object. Violations of these constraints lead to inconsistencies in the ontology, as they contradict the semantic rules established by the domain and range definitions.

For example, consider an object property `ownsPet` with a domain of `Person` and a range of `Animal`. This means:

1. The subject of the `ownsPet` relationship must be a `Person`.
2. The object of the `ownsPet` relationship must be an `Animal`.

If an assertion like `House ownsPet Dog` is made, it would violate the domain constraint because `House` is not an instance of the class `Person`. Similarly, if the property were used as `John ownsPet Chair`, this would violate the range constraint because `Chair` is not an instance of the class `Animal`.

Such violations undermine the logical consistency of the ontology, making reasoning unreliable. Clearly defining and enforcing domain and range constraints ensures that the relationships in the ontology align with its intended semantics, enabling accurate reasoning and error detection.

3.2. Statistical Contradictions

We utilized Relational Graph Convolutional Networks (R-GCN) [58] in our approach to model the complex relationships present in ontologies. R-GCN is particularly advantageous in handling multi-relational data as it extends the standard Graph Convolutional Network (GCN) [35] by incorporating relation-specific transformations for edges. This allows the model to capture the semantics of different types of relationships in the graph.

We trained the R-GCN on a link prediction task, where the model predicts missing links based on existing data. After training, we identified the top k triples with lowest prediction scores, which were then added as noise to the ontology. Specifically, we modified existing triples by replacing either the subject or the object with the entity that the R-GCN predicted to have the lowest probability score. This method assesses the impact of noise generated through a statistical model and provides insights into the reasoner's handling of statistically improbable but plausible assertions.

[6]https://www.w3.org/TR/owl2-syntax/#Disjoint_Classes
[7]https://www.w3.org/TR/owl2-syntax/#Disjoint_Object_Properties

3.3. Random Contradictions

We introduced k random triples to the ontology by corrupting either the object or the subject of existing triples. This method simulates random noise and evaluates the reasoner's resilience to arbitrary disruptions in the data. Unlike previous noise injection techniques, this random approach contrasts the effects of systematic versus random noise on ontology reasoning. By corrupting existing triples, this method helps to understand how well the reasoner manages unexpected and non-systematic errors, crucial for assessing its robustness in real-world scenarios with unpredictable data inconsistencies.

4. Experimental Setup

4.1. Datasets

We used OWL2Bench [64] and a modified Family ontology [71]. OWL2Bench includes a diverse set of axioms, such as Class Expression Axioms, Object Property Axioms, Data Property Axioms, and Assertions. OWL2Bench serves as a benchmark for assessing the coverage, scalability, and query performance of ontology reasoners across four OWL 2 profiles: EL, QL, RL and DL. OWL2Bench was extended from the well-known University Ontology Benchmark (UOBM) to create four TBoxes, one for each OWL 2 profile. Additionally, OWL2Bench includes an ABox generator and a set of 22 SPARQL queries involving reasoning tasks. For this paper, we modified *OWL2Bench1-DL*, where 1 is the number of universities and DL is the OWL 2 profile. *OWL2Bench-1* contains 60,573 axioms.

Furthermore, this work incorporates the Family ontology, a well-known ontology designed to represent family relationships and genealogical information. The Family ontology provides a foundational framework for reasoning about kinship terms, familial roles, and relationships such as parent-child, sibling, and spouse connections. *Family* contains 2,527 axioms. Table 1 lists the frequency of each axiom for each dataset.

Let G denote the original ontology and I the ontology inferred using Pellet reasoner [69]. For each resource r, we construct a subgraph g that includes all triples where either the subject or the object is r. We divide the original ontology into these smaller graphs g to improve Pellet's scalability. To ensure effective inference, each graph g is extended to two hops[8], denoted g', capturing all statements within two hops of r, and the TBox is added to each graph g. We then apply Pellet to the extended graphs g'_1, g'_2, ..., g'_R, where R represents the set of resources in the original ontology, resulting in the inference graphs i_1, i_2, ..., i_R. To extract only relevant inferred triples, we focus on membership and property assertion triples, removing any triples where the object is a `Literal` or `owl:Thing`, yielding refined graphs i^*_1, i^*_2, ..., i^*_R. Since our approach is unsupervised, the graphs g_1, g_2, ..., g_R are ultimately added to G_{train}, while i^*1, i^*_2, ..., i^*_R are assigned to G_{test} and G_{val} using a stratified splitting technique. The TBox is further added to G_{train}, G_{test} and G_{val}, ensuring that the reasoning tasks are based on a shared conceptual framework. Figure 1 illustrates this approach in detail[9].

Listing 1 contains an (simplified) extended graph about the resource `richard_john_bright_1962`, and Listing 2 contains the refined inference graph generated using Pellet.

In many domains, obtaining perfectly clean data is impractical or costly, particularly for ontologies derived from unstructured data. In addition, real-world datasets often contain errors, inconsistencies, or irrelevant information. By modeling noise, we can develop systems that are more robust and better suited to real-world scenarios. In this work, our aim is to introduce the noise generated by our approach into the training set to test the resilience of reasoners in real-world environments.

[8]We utilized two hops because the graph is sufficiently rich for making inferences and compact enough to apply the reasoner effectively.

[9]In our implementation, the validation set was not used since the reasoners did not require it, and as a result, it was eventually incorporated into the training set.

Table 1
Number of axioms in *OWL2Bench-1* and *Family*.

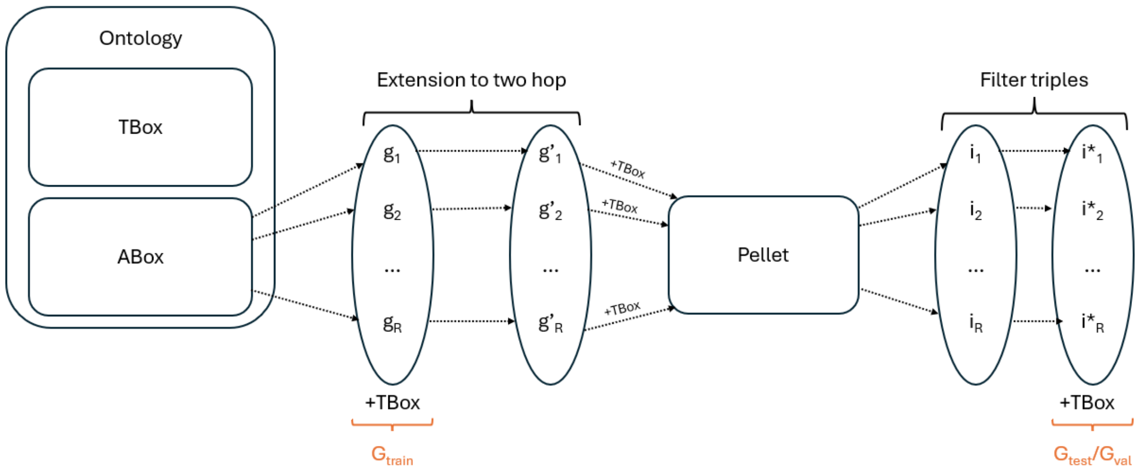
	OWL2Bench-1	Family
Class Expression Axioms		
Subclass Axioms	128	9
Equivalent Classes	21	5
Disjoint Classes	6,118	5
Object Property Axioms		
Object Subproperties	67	20
Equivalent Object Properties	4	1
Disjoint Object Properties	1	14
Inverse Object Properties	29	15
Object Property Domain	62	11
Object Property Range	57	13
Functional Object Properties	2	3
Inverse-Functional Object Properties	1	0
Reflexive Object Properties	2	0
Irreflexive Object Properties	2	2
Symmetric Object Properties	3	2
Asymmetric Object Properties	1	0
Transitive Object Properties	6	2
Role Composition	4	4
Data Property Axioms		
Data Subproperties	2	0
Equivalent Data Properties	1	0
Disjoint Data Properties	1	0
Data Property Domain	7	0
Data Property Range	1	0
Functional Data Properties	3	0
Assertions		
Individual Equality	2	0
Individual Inequality	4	1
Class Assertions	3,885	3
Positive Object Property Assertions	27,794	1,337
Negative Object Property Assertions	2	0
Positive Data Property Assertions	18,446	0
Negative Data Property Assertions	0	0

4.2. Metrics, Tasks and Reasoners

We used Mean Reciprocal Rank (MRR) and Hits@N to compare the performance of different neurosymbolic reasoners. MRR represents the average reciprocal rank, calculated by taking the reciprocal of the rank (1/rank) of the first relevant item retrieved. Hits@N measures the percentage of positive examples that appear in the top-k ranked predictions.

To assess how reasoners respond to noise, we focused on specific reasoning tasks: the first involves class assertions (also known as realization or membership), which determine whether an individual belongs to a specific class based on the logical definitions and constraints within the ontology, for example, `Alice rdf:type Person`. The second task involves object property assertions, that infer new relationships between two individuals in the ontology, such as `Alice hasSibling Bob`.

Fig. 1. Creation of train, test and validation graphs.



Listing 1 Input graph g'

```
ns1:richard_john_bright_1962 a ns1:Man,
    owl:NamedIndividual ;
ns1:hasFather ns1:david_bright_1934 ;
ns1:hasMother ns1:margaret_grace_rever_1934 ;
ns1:isBrotherOf ns1:robert_david_bright_1965 .

ns1:peter_william_bright_1941 ns1:isBrotherOf ns1:david_bright_1934 .
```

Listing 2 Inference graph i^*

```
ns1:peter_william_bright_1941 ns1:isUncleOf ns1:richard_john_bright_1962 .
```

This experimental framework analyzes the impact of noise on reasoning outcomes, as well as to evaluate the performance and robustness of ontology reasoning under different levels and types of noise. For our exploration into neurosymbolic reasoning, we have selected state-of-the-art neurosymbolic reasoners such as *Box2EL* [33] and *OWL2Vec** [10]. This work used the implementation of these methods provided by the mOWL library [80].

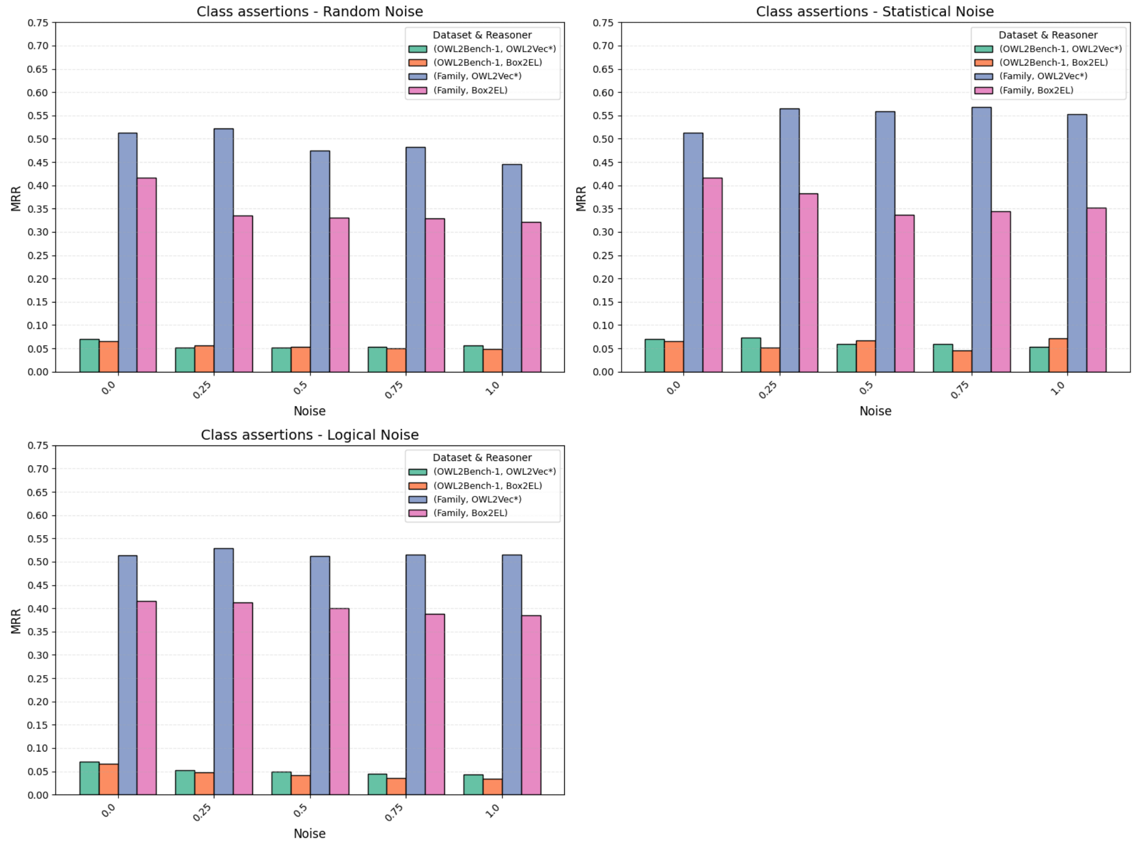
5. Results

Figures 2 and 3 demonstrate the impact of introducing noise into the ABox of *OWL2Bench-1* and *Family* ontologies on the performance of two reasoners: *OWL2Vec** and *Box2EL*. For each noise injection technique, we set a parameter k to control the number of 'noisy' triples added to the ontology. We represented this parameter as a percentage of the total triples in the ontology, providing a relative measure of the amount of noise introduced. The detailed results, reporting various evaluation metrics (including Mean Reciprocal Rank (MRR) and Hit@N) across different noise generation techniques, can be found in the supporting material (see Tables 2–5)[10]. To ensure reliable results, we ran each experiment 5 times, averaging out randomness to obtain a robust performance evaluation. The variability of the MRR is detailed in the supporting material (see Figures 4–7).

In Figures 2 and 3, which show the performance of *OWL2Vec** on *OWL2Bench-1*, the following trends can be observed. The MRR for both class and object property assertions decreases as various types of noise (i.e., random,

[10]Except for OWL2Bench-OWL2Vec*, MRR and Hits@10 exhibit a significant correlation: OWL2Bench-OWL2Vec* ($r = 0.1153$, $p = 0.5749$), OWL2Bench-Box2EL ($r = 0.9844$, $p = 0.000$), Family-OWL2Vec* ($r = 0.8349$, $p = 0.0000$), and Family-Box2EL ($r = 0.9963$, $p = 0.0000$).

Fig. 2. Results on class assertions.



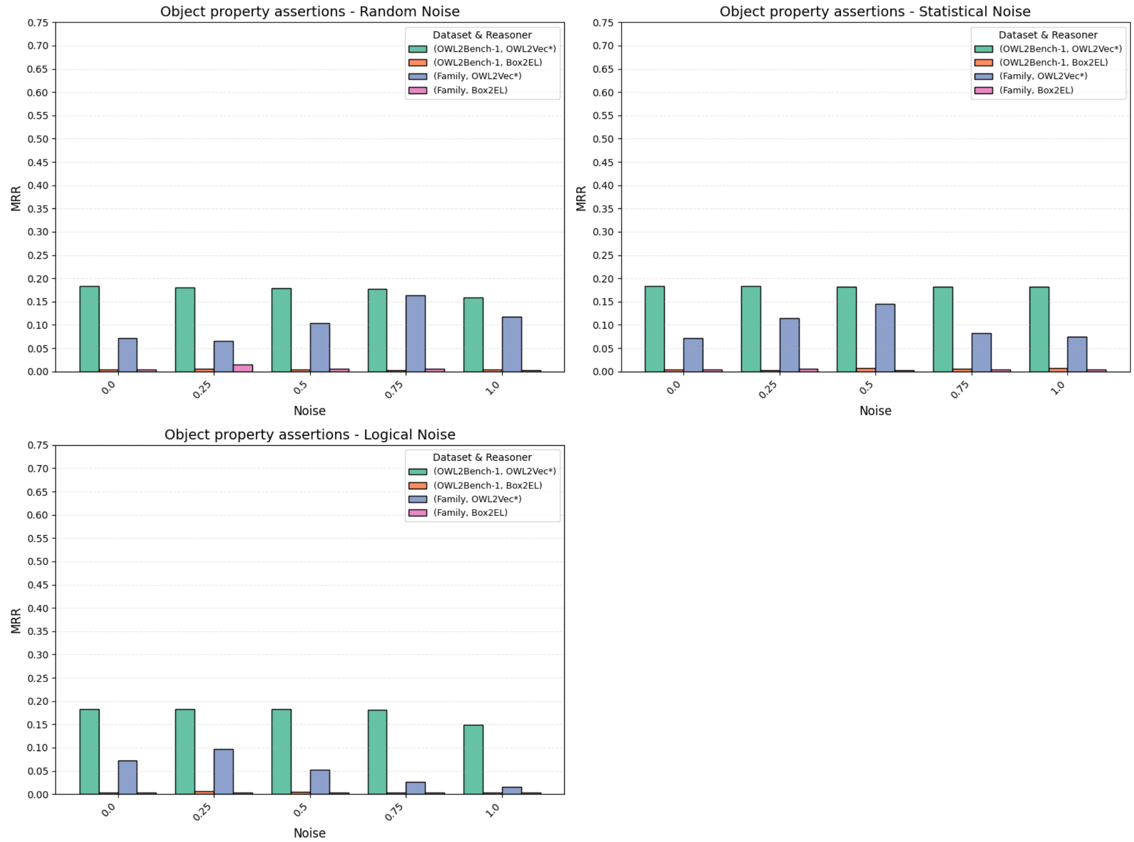
statistical, and logical) are introduced. Among these, logical noise has the greatest impact. Under the 100% logical noise scenario, the MRR for class assertions drops from 0.070 (without noise) to 0.043, while the MRR for object property assertions falls from 0.183 (without noise) to 0.149.

A similar trend can be observed in *Box2EL*. Logical noise has the most pronounced effect on class assertions, with the MRR decreasing from 0.066 (without noise) to 0.034. Moreover, *Box2EL* consistently underperforms on object property assertions, even in the absence of noise. Overall, the class assertion task in the *OWL2Bench-1* ontology proves to be particularly challenging. The average MRR scores range from 0.066 to 0.070 without noise, but can drop to 0.034 when logical noise is introduced.

The results reveal that the performance of both *OWL2Vec** and *Box2EL* (Figures 2 and 3) on the *Family* ontology exhibits slightly different trends compared to those observed on the *OWL2Bench-1* dataset. The object property task proves particularly challenging, with the MRR score of 0.072 in the absence of noise dropping to its lowest value of 0.015 when 100% logical noise is introduced. In contrast, class assertions appear less challenging and more resilient to all types of noise, achieving an MRR score of 0.513 without noise and dropping to its lowest value of 0.446 under 100% random noise.

Similarly, we observe that *Box2EL* consistently underperforms *OWL2Vec** in both tasks, with random noise having the most significant impact on class assertions. The MRR score decreases from 0.416 without noise to 0.322 under 100% random noise. For object property assertions, it is difficult to identify any clear trend, as the values are already low, even without the introduction of noise.

Fig. 3. Results on object property assertions.



6. Discussion

Our study investigates the application of noise injection methods to ontologies, examining their impact on various reasoning tasks. The proposed noise injection techniques are designed to be applicable across a wide range of ontologies. Based on our findings, we observed that class assertions are most affected by either logical or random noise, depending on the ontology. Logical noise, in particular, leads to a significant decrease in object property assertions, especially in the case of *OWL2Bench-1*. Another important finding is that certain tasks are particularly challenging. For example, in the Family ontology, the object property assertion task is particularly difficult, with neurosymbolic reasoners achieving the highest MRR score of 0.072 without noise. With noise, this score can drop to 0.004. Similarly, for the OWL2Bench ontology, the class assertion task presents considerable difficulty. The average MRR scores range from 0.066 to 0.070 without noise but drop to 0.049 when noise is introduced.

However, the specific characteristics of each ontology significantly influence the effectiveness of noise injection, highlighting the need for tailored approaches in certain scenarios. For example, the specific relations in the test set may not effectively show the influence of noise introduced as these relations inherently resist noise. In *OWL2Bench-1*, `knows` relation is defined as reflexive (i.e., every individual 'knows' themselves), making it less sensitive to object property assertion inferences. These inferences hold regardless of corrupted assertions unless the TBox is modified. This raises questions about the validity of evaluating noise effects in scenarios where axiomatic properties dominate reasoning outcomes. Future work should consider refining testing sets or introducing variations in TBox definitions to better capture the influence of noise.

Furthermore, it should be noted that the results from previous works, such as the work of [79], are not comparable to ours due to the fact that our proposed benchmark focuses on evaluating ontology reasoning rather than ontology completion. Ontology reasoning refers to inferring logically consistent relationships from existing data and rules, which is inherently more complex. This complexity arises because reasoning requires the system to consider all possible logical implications of the data, making it more sensitive to inconsistencies and noise in the dataset. Consequently, the metrics may reflect this added difficulty, leading to poorer results compared to approaches that focus solely on completing the ontology.

While our initial exploration centered on introducing noise through the addition of logical contradictions or corruption of triples with low probability of occurrence, many other types of axioms and noise patterns merit investigation. Future research could involve examining various inconsistencies, contradictions, and errors that frequently occur in real-world ontologies, thereby enhancing the diversity of noise generation techniques. In particular, introducing noise in the TBox (e.g., modifying class hierarchies, altering domain and range constraints, or introducing invalid equivalence axioms) could offer valuable insights into how structural and logical inconsistencies impact reasoning outcomes. Furthermore, future work could focus on establishing standardized metrics and evaluation frameworks to consistently measure the performance of neurosymbolic reasoning systems.

7. Conclusion

This paper presents NSORN (Neurosymbolic Ontology Reasoning with Noise), a framework for generating noisy benchmark datasets, with a specific focus on the generation of noisy ABox assertions for an ontology. We developed three techniques for introducing noise into the ABox: logical noise, statistical noise, and random noise. Logical noise is introduced by contradicting disjoint axioms or violating domain/range constraints of object properties. Statistical noise, on the other hand, leverages Graph Neural Networks to add new links with low probability scores. Random noise involves arbitrarily altering ABox assertions. These methods were designed to evaluate the robustness and performance of ontology-based neurosymbolic reasoners under various noise conditions.

We evaluated the performance of existing neurosymbolic reasoners on *OWL2Bench* and *Family* under different noise levels. The resulting benchmarks were tested on state-of-the-art neurosymbolic reasoners, *Box2EL* and *OWL2Vec**. The reasoning tasks considered include class assertions and object property assertions, with the aim of evaluating how effectively these reasoners handle noise. Our findings suggest that class assertions are primarily influenced by either logical or random noise, depending on the ontology. Logical noise causes a considerable decline in object property assertions, with a more pronounced effect observed in *OWL2Bench*. Furthermore, our study highlights that most previous work has mainly focused on ontology completion, whereas our emphasis is on ontology reasoning, which is a more difficult task. The source code of NSORN is available at https://github.com/jloe2911/NoisyBench under MIT License.

Gunjan Singh and Raghava Mutharaju would like to acknowledge the partial support of the Infosys Centre for Artificial Intelligence (CAI), IIIT-Delhi, India, in this work.

References

[1] L.E. Anke, T. Declerck, D. Gromann, B. Makni, J. Hendler, D. Gromann, L. Espinosa Anke and T. Declerck, Deep learning for noise-tolerant RDFS reasoning, *Semant. Web* **10**(5) (2019), 823–862–. doi:10.3233/SW-190363.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin and G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet* **25**(1) (2000), 25–29. doi:10.1038/75556. http://www.ncbi.nlm.nih.gov/pubmed/10802651.

[3] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi and P.F. Patel-Schneider (eds), *The Description Logic Handbook: Theory, Implementation, and Applications*, Cambridge University Press, 2003. ISBN 0-521-78176-0.

[4] S. Badreddine, A.S. d'Avila Garcez, L. Serafini and M. Spranger, Logic Tensor Networks, *Artif. Intell.* **303** (2022), 103649. doi:10.1016/j.artint.2021.103649.

[5] D. Banerjee, R. Usbeck, N. Mihindukulasooriya, M.Y. Jaradeh, S. Auer, G. Singh, R. Mutharaju and P. Kapanipathi (eds), Joint Proceedings of Scholarly QALD 2023 and SemREC 2023, in: *22nd International Semantic Web Conference ISWC 2023*, CEUR Workshop Proceedings, Aachen, 2023. ISSN 1613-0073. http://ceur-ws.org/Vol-3592/.

[6] A. Bordes, N. Usunier, A. Garcia-Durán, J. Weston and O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, pp. 2787–2795–.

[7] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang and W. Zaremba, OpenAI Gym, 2016.

[8] J.J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne and K. Wilkinson, Jena: implementing the semantic web recommendations, in: *Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters*, WWW Alt. '04, Association for Computing Machinery, New York, NY, USA, 2004, pp. 74–83–. ISBN 1581139128. doi:10.1145/1013367.1013381.

[9] J. Chen, F. Lécué, Y. Geng, J.Z. Pan and H. Chen, Ontology-guided Semantic Composition for Zero-shot Learning, in: *Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning, KR 2020, Rhodes, Greece, September 12-18, 2020*, D. Calvanese, E. Erdem and M. Thielscher, eds, 2020, pp. 850–854. doi:10.24963/kr.2020/87.

[10] J. Chen, P. Hu, E. Jimenez-Ruiz, O.M. Holter, D. Antonyrajah and I. Horrocks, OWL2Vec*: Embedding of OWL Ontologies, 2021.

[11] A.S. d'Avila Garcez, T.R. Besold, L.D. Raedt, P. Földiák, P. Hitzler, T. Icard, K. Kühnberger, L.C. Lamb, R. Miikkulainen and D.L. Silver, Neural-Symbolic Learning and Reasoning: Contributions and Challenges, in: *2015 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 22-25, 2015*, AAAI Press, 2015. http://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10281.

[12] L. Deng, The mnist database of handwritten digit images for machine learning research, *IEEE Signal Processing Magazine* **29**(6) (2012), 141–142.

[13] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran and T. Solorio, eds, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

[14] D.M. Dooley, E.J. Griffiths, G.S. Gosal, P.L. Buttigieg, R. Hoehndorf, M.C. Lange, L.M. Schriml, F.S.L. Brinkman and W.W.L. Hsiao, FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, *npj Science of Food* **2**(1) (2018), 23–. doi:10.1038/s41538-018-0032-6.

[15] M. Dragoni, T. Bailoni, R. Maimone and C. Eccher, HeLiS: An Ontology for Supporting Healthy Lifestyles, in: *The Semantic Web – ISWC 2018*, D. Vrandečić, K. Bontcheva, M.C. Suárez-Figueroa, V. Presutti, I. Celino, M. Sabou, L.-A. Kaffee and E. Simperl, eds, Springer International Publishing, Cham, 2018, pp. 53–69. ISBN 978-3-030-00668-6.

[16] A. Eberhart, M. Ebrahimi, L. Zhou, C. Shimizu and P. Hitzler, Completion Reasoning Emulation for the Description Logic $\mathcal{EL}^+$, in: *Proceedings of the AAAI 2020 Spring Symposium on Combining Machine Learning and Knowledge Engineering in Practice, AAAI-MAKE 2020, Palo Alto, CA, USA, March 23-25, 2020, Volume I*, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle and F. van Harmelen, eds, CEUR Workshop Proceedings, Vol. 2600, CEUR-WS.org, 2020. http://ceur-ws.org/Vol-2600/paper5.pdf.

[17] M. Ebrahimi, A. Eberhart and P. Hitzler, On the Capabilities of Pointer Networks for Deep Deductive Reasoning, *CoRR* **abs/2106.09225** (2021). https://arxiv.org/abs/2106.09225.

[18] M. Ebrahimi, M.K. Sarker, F. Bianchi, N. Xie, D. Doran and P. Hitzler, Reasoning over RDF Knowledge Bases using Deep Learning, *CoRR* **abs/1811.04132** (2018). http://arxiv.org/abs/1811.04132.

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

[20] S. Farzana, Q. Zhou and P. Ristoski, Knowledge Graph-Enhanced Neural Query Rewriting, in: *Companion Proceedings of the ACM Web Conference 2023, WWW 2023, Austin, TX, USA, 30 April 2023 - 4 May 2023*, ACM, 2023, pp. 911–919. doi:10.1145/3543873.3587678.

[21] A.d. Garcez, T.R. Besold, L. De Raedt, P. Földiak, P. Hitzler, T. Icard, K.-U. Kühnberger, L.C. Lamb, R. Miikkulainen and D.L. Silver, Neural-symbolic learning and reasoning: contributions and challenges, in: *2015 AAAI Spring Symposium Series*, 2015.

[22] D. Garg, S. Ikbal, S.K. Srivastava, H. Vishwakarma, H.P. Karanam and L.V. Subramaniam, Quantum Embedding of Knowledge for Reasoning, in: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox and R. Garnett, eds, 2019, pp. 5595–5605. https://proceedings.neurips.cc/paper/2019/hash/cb12d7f933e7d102c52231bf62b8a678-Abstract.html.

[23] B. Glimm, I. Horrocks, B. Motik, G. Stoilos and Z. Wang, HermiT: An OWL 2 Reasoner, *J. Autom. Reason.* **53**(3) (2014), 245–269–. doi:10.1007/s10817-014-9305-1.

[24] B.C. Grau, I. Horrocks, B. Motik, B. Parsia, P.F. Patel-Schneider and U. Sattler, OWL 2: The next step for OWL, *Journal of Web Semantics* **6**(4) (2008), 309–322. doi:10.1016/j.websem.2008.05.001.

[25] J. Gray (ed.), *The Benchmark Handbook for Database and Transaction Systems (2nd Edition)*, Morgan Kaufmann, 1993. ISBN 1-55860-292-5.

[26] Y. Guo, Z. Pan and J. Heflin, LUBM: A Benchmark for OWL Knowledge Base Systems, *Journal of Web Semantics.* 3(2–3) (2005), 158–182.

[27] P. Hitzler, M. Krötzsch and S. Rudolph, *Foundations of Semantic Web Technologies*, Chapman and Hall/CRC Press, 2010. ISBN 9781420090505. http://www.semantic-web-book.org/.

[28] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, S. Kirrane, J.E.L. Gayo, R. Navigli, S. Neumaier, A.N. Ngomo, A. Polleres, S.M. Rashid, A. Rula, L. Schmelzeisen, J.F. Sequeda, S. Staab and A. Zimmermann, Knowledge Graphs, *ACM Computing Surveys* **54**(4) (2022), 71:1–71:37. doi:10.1145/3447772.

[29] P. Hohenecker and T. Lukasiewicz, Deep Learning for Ontology Reasoning, *CoRR* **abs/1705.10342** (2017). http://arxiv.org/abs/1705.10342.

[30] P. Hohenecker and T. Lukasiewicz, Ontology Reasoning with Deep Neural Networks, *Journal of Artificial Intelligence Research* **68** (2020), 503–540. doi:10.1613/jair.1.11661.
[31] I. Horrocks, O. Kutz and U. Sattler, The Even More Irresistible SROIQ, in: *Proceedings, Tenth International Conference on Principles of Knowledge Representation and Reasoning, Lake District of the United Kingdom, June 2-5, 2006*, P. Doherty, J. Mylopoulos and C.A. Welty, eds, AAAI Press, 2006, pp. 57–67. http://www.aaai.org/Library/KR/2006/kr06-009.php.
[32] M. Jackermeier, J. Chen and I. Horrocks, Dual box embeddings for the description logic EL++, Association for Computing Machinery, 2024.
[33] M. Jackermeier, J. Chen and I. Horrocks, Dual Box Embeddings for the Description Logic EL++, 2024.
[34] H.A. Kautz, The Third AI Summer: AAAI Robert S. Engelmore Memorial Lecture, *AI Magazine* **43**(1) (2022), 93–104. doi:10.1609/aimag.v43i1.19122.
[35] T.N. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, *CoRR* **abs/1609.02907** (2016). http://arxiv.org/abs/1609.02907.
[36] G. Klyne, J. J. Carroll and B. McBride, Resource Description Framework (RDF), 2015. https://www.w3.org/TR/rdf11-concepts/.
[37] M. Krötzsch, F. Simancik and I. Horrocks, A Description Logic Primer, *CoRR* **abs/1201.4089** (2012). http://arxiv.org/abs/1201.4089.
[38] M. Kulmanov, W. Liu-Wei, Y. Yan and R. Hoehndorf, EL Embeddings: Geometric Construction of Models for the Description Logic $\mathcal{EL}^{++}$, in: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, S. Kraus, ed., ijcai.org, 2019, pp. 6103–6109. doi:10.24963/ijcai.2019/845.
[39] G. Lample and F. Charton, Deep Learning For Symbolic Mathematics, in: *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, OpenReview.net, 2020. https://openreview.net/forum?id=S1eZYeHFDS.
[40] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. Mendes, S. Hellmann, M. Morsey, P. Van Kleef, S. Auer and C. Bizer, DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, *Semantic Web Journal* **6** (2014). doi:10.3233/SW-140134.
[41] L. Ma, Y. Yang, G. Qiu Z .and Xie, Y. Pan and S. Liu, Towards a Complete OWL Ontology Benchmark, in: *The Semantic Web: Research and Applications*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 125–139.
[42] B. Makni and J.A. Hendler, Deep learning for noise-tolerant RDFS reasoning, *Semantic Web* **10**(5) (2019), 823–862. doi:10.3233/SW-190363.
[43] B. Makni, I. Abdelaziz and J.A. Hendler, Explainable Deep RDFS Reasoner, *CoRR* **abs/2002.03514** (2020). https://arxiv.org/abs/2002.03514.
[44] B. Makni, M. Ebrahimi, D. Gromann and A. Eberhart, Neuro-Symbolic Semantic Reasoning, in: *Neuro-Symbolic Artificial Intelligence: The State of the Art*, P. Hitzler and M.K. Sarker, eds, Frontiers in Artificial Intelligence and Applications, Vol. 342, IOS Press, 2021, pp. 253–279. ISBN 978-1-64368-244-0. doi:10.3233/FAIA210358.
[45] J. Mao, C. Gan, P. Kohli, J.B. Tenenbaum and J. Wu, The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision, in: *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. https://openreview.net/forum?id=rJgMlhRctm.
[46] T. Mikolov, K. Chen, G.S. Corrado and J. Dean, Efficient Estimation of Word Representations in Vector Space, in: *International Conference on Learning Representations*, 2013. https://api.semanticscholar.org/CorpusID:5959482.
[47] B. Mohapatra, S. Bhatia, R. Mutharaju and G. Srinivasaraghavan, Why Settle for Just One? Extending $\mathcal{EL}^{++}$ Ontology Embeddings with Many-to-Many Relationships, *CoRR* **abs/2110.10555** (2021). https://arxiv.org/abs/2110.10555.
[48] S. Mondal, S. Bhatia and R. Mutharaju, EmEL^{++}: Embeddings for $\mathcal{EL}^{++}$ Description Logic, in: *Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Stanford University, Palo Alto, California, USA, March 22-24, 2021*, A. Martin, K. Hinkelmann, H. Fill, A. Gerber, D. Lenat, R. Stolle and F. van Harmelen, eds, CEUR Workshop Proceedings, Vol. 2846, CEUR-WS.org, 2021. http://ceur-ws.org/Vol-2846/paper19.pdf.
[49] B. Motik, B.C. Grau, I. Horrocks, Z. Wu, A. Fokoue and C. Lutz, OWL 2 Web Ontology Language Profiles (Second Edition), 2012. https://www.w3.org/TR/owl2-profiles/.
[50] C.J. Mungall, C. Torniai, G.V. Gkoutos, S.E. Lewis and M.A. Haendel, Uberon, an integrative multi-species anatomy ontology, *Genome-Biology.com* **13**(1) (2012). doi:10.1186/gb-2012-13-1-r5.
[51] J. Ott, A. Ledaguenel, C. Hudelot and M. Hartwig, How to Think About Benchmarking Neurosymbolic AI?, in: *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning, La Certosa di Pontignano, Siena, Italy, July 3-5, 2023*, A.S. d'Avila Garcez, T.R. Besold, M. Gori and E. Jiménez-Ruiz, eds, CEUR Workshop Proceedings, Vol. 3432, CEUR-WS.org, 2023, pp. 248–254. https://ceur-ws.org/Vol-3432/paper22.pdf.
[52] Ö.L. Özçep, M. Leemhuis and D. Wolter, Cone Semantics for Logics with Negation, in: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, C. Bessiere, ed., ijcai.org, 2020, pp. 1820–1826. doi:10.24963/ijcai.2020/252.
[53] B. Parsia, N. Matentzoglu, R.S. Gonçalves, B. Glimm and A. Steigmiller, The OWL Reasoner Evaluation (ORE) 2015 Competition Report, *Journal of Automated Reasoning* **59**(4) (2017), 455–482. doi:10.1007/s10817-017-9406-8.
[54] I.D. Raji, E.M. Bender, A. Paullada, E. Denton and A. Hanna, AI and the Everything in the Whole Wide World Benchmark, *CoRR* **abs/2111.15366** (2021). https://arxiv.org/abs/2111.15366.
[55] A.L. Rector, The GALEN High Level Ontology, 2008. https://api.semanticscholar.org/CorpusID:73607647.
[56] M.K. Sarker, L. Zhou, A. Eberhart and P. Hitzler, Neuro-Symbolic Artificial Intelligence, *AI Communications* **34**(3) (2021), 197–209.
[57] F. Scarselli, M. Gori, A.C. Tsoi, M. Hagenbuchner and G. Monfardini, The Graph Neural Network Model, *IEEE Transactions on Neural Networks* **20**(1) (2009), 61–80. doi:10.1109/TNN.2008.2005605.

[58] M. Schlichtkrull, T.N. Kipf, P. Bloem, R. van den Berg, I. Titov and M. Welling, Modeling Relational Data with Graph Convolutional Networks, 2017.

[59] P. Sen, B.W.S.R. de Carvalho, R. Riegel and A.G. Gray, Neuro-Symbolic Inductive Logic Programming with Logical Neural Networks, in: *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, AAAI Press, 2022, pp. 8212–8219. https://ojs.aaai.org/index.php/AAAI/article/view/20795.

[60] A. Sheth, K. Roy and M. Gaur, Neurosymbolic Artificial Intelligence (Why, What, and How), *IEEE Intelligent Systems* **38**(3) (2023), 56–62. doi:10.1109/MIS.2023.3268724.

[61] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T.P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel and D. Hassabis, Mastering the game of Go with deep neural networks and tree search, *Nat.* **529**(7587) (2016), 484–489. doi:10.1038/nature16961.

[62] G. Singh, Benchmarking Symbolic and Neuro-Symbolic Description Logic Reasoners, Doctoral Consortium at International Semantic Web Conference, 2023.

[63] G. Singh, Benchmarking Symbolic and Neuro-Symbolic Description Logic Reasoners, in: *Proceedings of the Doctoral Consortium at ISWC 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023*, C. d'Amato and J.Z. Pan, eds, CEUR Workshop Proceedings, Vol. 3678, CEUR-WS.org, 2023. https://ceur-ws.org/Vol-3678/paper11.pdf.

[64] G. Singh, S. Bhatia and R. Mutharaju, OWL2Bench: A Benchmark for OWL 2 Reasoners, in: *The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, J.Z. Pan, V.A.M. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne and L. Kagal, eds, Lecture Notes in Computer Science, Vol. 12507, Springer, 2020, pp. 81–96. doi:10.1007/978-3-030-62466-8_6.

[65] G. Singh, S. Bhatia and R. Mutharaju, Neuro-Symbolic RDF and Description Logic Reasoners: The State-Of-The-Art and Challenges, in: *Compendium of Neurosymbolic Artificial Intelligence*, P. Hitzler and M.K. Sarker, eds, Frontiers in Artificial Intelligence and Applications, Vol. 369, IOS Press, 2023, pp. 29–63. ISBN 978-1-64368-407-9. doi:10.3233/FAIA230134.

[66] G. Singh, R. Mutharaju and P. Kapanipathi (eds), Proceedings of the Semantic Reasoning Evaluation Challenge (SemREC 2021), in: *20th International Semantic Web Conference (ISWC 2021)*, CEUR Workshop Proceedings, Aachen, 2021. ISSN 1613-0073. http://ceur-ws.org/Vol-3123/.

[67] G. Singh, R. Mutharaju, S. Bhatia and R. Tommasini, Benchmarking Neuro-Symbolic Reasoners: Existing Challenges and A Way Forward. https://neurosymbolic-ai-journal.com/system/files/nai-paper-774.pdf.

[68] G. Singh, R. Mutharaju, P. Kapanipathi, N. Mihindukulasooriya, M. Dubey, R. Usbeck and D. Banerjee (eds), Joint Proceedings of SemREC 2022 and SMART 2022, in: *21st International Semantic Web Conference (ISWC 2022)*, CEUR Workshop Proceedings, Aachen, 2022. ISSN 1613-0073. http://ceur-ws.org/Vol-3337/.

[69] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur and Y. Katz, Pellet: A practical OWL-DL reasoner, *Journal of Web Semantics* **5**(2) (2007), 51–53, Software Engineering and the Semantic Web. doi:https://doi.org/10.1016/j.websem.2007.03.004. https://www.sciencedirect.com/science/article/pii/S1570826807000169.

[70] A. Steigmiller, T. Liebig and B. Glimm, Konclude: System description, *Journal of Web Semantics* **27-28** (2014), 78–85, Semantic Web Challenge 2013. doi:https://doi.org/10.1016/j.websem.2014.06.003. https://www.sciencedirect.com/science/article/pii/S157082681400047X.

[71] R. Stevens and M. Stevens, A Family History Knowledge Base Using OWL 2, in: *OWL: Experiences and Directions*, 2008. https://api.semanticscholar.org/CorpusID:10478581.

[72] F.M. Suchanek, G. Kasneci and G. Weikum, YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, in: *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 697–706–. ISBN 9781595936547. doi:10.1145/1242572.1242667.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin, Attention is All you Need, in: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S.V.N. Vishwanathan and R. Garnett, eds, 2017, pp. 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[74] D. Vrandečić, Wikidata: A New Platform for Collaborative Data Collection, in: *Proceedings of the 21st International Conference on World Wide Web*, WWW '12 Companion, Association for Computing Machinery, New York, NY, USA, 2012, pp. 1063–1064–. ISBN 9781450312301. doi:10.1145/2187980.2188242.

[75] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy and S.R. Bowman, GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding, 2019.

[76] B. Xiong, N. Potyka, T. Tran, M. Nayyeri and S. Staab, Faithful Embeddings for $\mathcal{EL}^{++}$ Knowledge Bases, in: *The Semantic Web - ISWC 2022 - 21st International Semantic Web Conference, Virtual Event, October 23-27, 2022, Proceedings*, U. Sattler, A. Hogan, C.M. Keet, V. Presutti, J.P.A. Almeida, H. Takeda, P. Monnin, G. Pirrò and C. d'Amato, eds, Lecture Notes in Computer Science, Vol. 13489, Springer, 2022, pp. 22–38. doi:10.1007/978-3-031-19433-7_2.

[77] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao and J. Zhu, Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges, in: *Natural Language Processing and Chinese Computing - 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9-14, 2019, Proceedings, Part II*, J. Tang, M. Kan, D. Zhao, S. Li and H. Zan, eds, Lecture Notes in Computer Science, Vol. 11839, Springer, 2019, pp. 563–574. doi:10.1007/978-3-030-32236-6_51.

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

[79] F. Zhapa-Camacho and R. Hoehndorf, Evaluating Different Methods for Semantic Reasoning Over Ontologies, in: *QALD/SemREC@ ISWC*, 2023.

[80] F. Zhapa-Camacho, M. Kulmanov and R. Hoehndorf, mOWL: Python library for machine learning with biomedical ontologies, *Bioinformatics* (2022), btac811. doi:10.1093/bioinformatics/btac811.

Appendix A. Supporting material

		MRR	Hits@1	Hits@5	Hits@10
No Noise	Class assertions	0.070	0.001	0.151	0.230
	Object property assertions	0.183	0.166	0.190	0.212
25% Random Noise	Class assertions	0.052	0.000	0.111	0.200
	Object property assertions	0.180	0.164	0.185	0.206
50% Random Noise	Class assertions	0.052	0.000	0.088	0.210
	Object property assertions	0.179	0.163	0.184	0.206
75% Random Noise	Class assertions	0.053	0.000	0.097	0.224
	Object property assertions	0.178	0.162	0.183	0.204
100% Random Noise	Class assertions	0.056	0.000	0.113	0.270
	Object property assertions	0.159	0.138	0.169	0.193
25% Statistical Noise	Class assertions	0.073	0.000	0.157	0.250
	Object property assertions	0.183	0.167	0.188	0.212
50% Statistical Noise	Class assertions	0.060	0.000	0.103	0.228
	Object property assertions	0.182	0.166	0.186	0.208
75% Statistical Noise	Class assertions	0.060	0.000	0.101	0.221
	Object property assertions	0.182	0.167	0.188	0.210
100% Statistical Noise	Class assertions	0.053	0.000	0.066	0.248
	Object property assertions	0.182	0.167	0.187	0.210
25% Logical Noise	Class assertions	0.053	0.002	0.075	0.170
	Object property assertions	0.183	0.166	0.189	0.210
50% Logical Noise	Class assertions	0.049	0.001	0.068	0.155
	Object property assertions	0.182	0.167	0.188	0.208
75% Logical Noise	Class assertions	0.044	0.003	0.048	0.115
	Object property assertions	0.181	0.167	0.187	0.206
100% Logical Noise	Class assertions	<u>0.043</u>	0.002	0.046	0.114
	Object property assertions	<u>0.149</u>	0.137	0.155	0.171

Table 2

Results on *OWL2Bench1* using *OWL2Vec** [10]. The lowest MRR values are underlined.

		MRR	Hits@1	Hits@5	Hits@10
No Noise	Class assertions	0.066	0.003	0.070	0.221
	Object property assertions	0.004	0.001	0.003	0.006
25% Random Noise	Class assertions	0.056	0.002	0.055	0.189
	Object property assertions	0.005	0.001	0.004	0.008
50% Random Noise	Class assertions	0.053	0.002	0.054	0.186
	Object property assertions	0.004	0.001	0.003	0.006
75% Random Noise	Class assertions	0.050	0.002	0.049	0.168
	Object property assertions	0.003	0.001	0.002	0.005
100% Random Noise	Class assertions	0.049	0.002	0.045	0.174
	Object property assertions	0.004	0.001	0.003	0.006
25% Statistical Noise	Class assertions	0.052	0.004	0.045	0.156
	Object property assertions	<u>0.003</u>	0.000	0.002	0.003
50% Statistical Noise	Class assertions	0.067	0.011	0.098	0.190
	Object property assertions	0.007	0.001	0.008	0.014
75% Statistical Noise	Class assertions	0.045	0.003	0.042	0.125
	Object property assertions	0.006	0.002	0.007	0.014
100% Statistical Noise	Class assertions	0.071	0.019	0.103	0.181
	Object property assertions	0.007	0.001	0.008	0.015
25% Logical Noise	Class assertions	0.048	0.002	0.046	0.145
	Object property assertions	0.006	0.002	0.006	0.012
50% Logical Noise	Class assertions	0.041	0.002	0.038	0.112
	Object property assertions	0.005	0.000	0.006	0.010
75% Logical Noise	Class assertions	0.035	0.002	0.032	0.086
	Object property assertions	0.004	0.001	0.004	0.009
100% Logical Noise	Class assertions	<u>0.034</u>	0.002	0.029	0.078
	Object property assertions	0.004	0.000	0.004	0.009

Table 3

Results on *OWL2Bench1* using *Box2EL* [33]. The lowest MRR values are underlined.

		MRR	Hits@1	Hits@5	Hits@10
No Noise	Class assertions	0.513	0.297	1.000	1.000
	Object property assertions	0.072	0.000	0.100	0.400
25% Random Noise	Class assertions	0.522	0.285	0.946	1.000
	Object property assertions	0.066	0.000	0.100	0.360
50% Random Noise	Class assertions	0.474	0.230	0.908	0.995
	Object property assertions	0.103	0.000	0.300	0.460
75% Random Noise	Class assertions	0.482	0.235	0.901	0.993
	Object property assertions	0.164	0.000	0.400	0.500
100% Random Noise	Class assertions	<u>0.446</u>	0.190	0.848	0.972
	Object property assertions	0.118	0.000	0.300	0.400
25% Statistical Noise	Class assertions	0.565	0.340	0.991	1.000
	Object property assertions	0.115	0.000	0.200	0.500
50% Statistical Noise	Class assertions	0.559	0.332	0.958	1.000
	Object property assertions	0.145	0.000	0.400	0.400
75% Statistical Noise	Class assertions	0.568	0.332	0.989	1.000
	Object property assertions	0.083	0.000	0.120	0.500
100% Statistical Noise	Class assertions	0.553	0.335	0.981	1.000
	Object property assertions	0.074	0.000	0.200	0.300
25% Logical Noise	Class assertions	0.529	0.335	0.860	1.000
	Object property assertions	0.097	0.000	0.300	0.400
50% Logical Noise	Class assertions	0.512	0.329	0.844	1.000
	Object property assertions	0.053	0.000	0.000	0.240
75% Logical Noise	Class assertions	0.515	0.326	0.843	1.000
	Object property assertions	0.026	0.000	0.000	0.000
100% Logical Noise	Class assertions	0.516	0.327	0.843	1.000
	Object property assertions	<u>0.015</u>	0.000	0.000	0.000

Table 4

Results on *Family* using *OWL2Vec** [10]. The lowest MRR values are underlined.

		MRR	Hits@1	Hits@5	Hits@10
No Noise	Class assertions	0.416	0.220	0.668	0.928
	Object property assertions	0.004	0.000	0.000	0.000
25% Random Noise	Class assertions	0.335	0.182	0.519	0.654
	Object property assertions	0.015	0.000	0.020	0.020
50% Random Noise	Class assertions	0.331	0.174	0.539	0.662
	Object property assertions	0.005	0.000	0.000	0.000
75% Random Noise	Class assertions	0.329	0.165	0.546	0.679
	Object property assertions	0.005	0.000	0.000	0.000
100% Random Noise	Class assertions	<u>0.322</u>	0.171	0.513	0.641
	Object property assertions	<u>0.003</u>	0.000	0.000	0.000
25% Statistical Noise	Class assertions	0.382	0.199	0.597	0.866
	Object property assertions	0.005	0.000	0.000	0.000
50% Statistical Noise	Class assertions	0.337	0.163	0.531	0.768
	Object property assertions	0.003	0.000	0.000	0.000
75% Statistical Noise	Class assertions	0.344	0.173	0.516	0.775
	Object property assertions	0.004	0.000	0.000	0.000
100% Statistical Noise	Class assertions	0.352	0.190	0.534	0.731
	Object property assertions	0.004	0.000	0.000	0.000
25% Logical Noise	Class assertions	0.412	0.224	0.646	0.934
	Object property assertions	0.004	0.000	0.000	0.000
50% Logical Noise	Class assertions	0.400	0.210	0.631	0.932
	Object property assertions	0.003	0.000	0.000	0.000
75% Logical Noise	Class assertions	0.388	0.205	0.610	0.916
	Object property assertions	0.004	0.000	0.000	0.000
100% Logical Noise	Class assertions	0.385	0.206	0.591	0.912
	Object property assertions	0.004	0.000	0.000	0.000

Table 5

Results on *Family* using *Box2EL* [33]. The lowest MRR values are underlined.

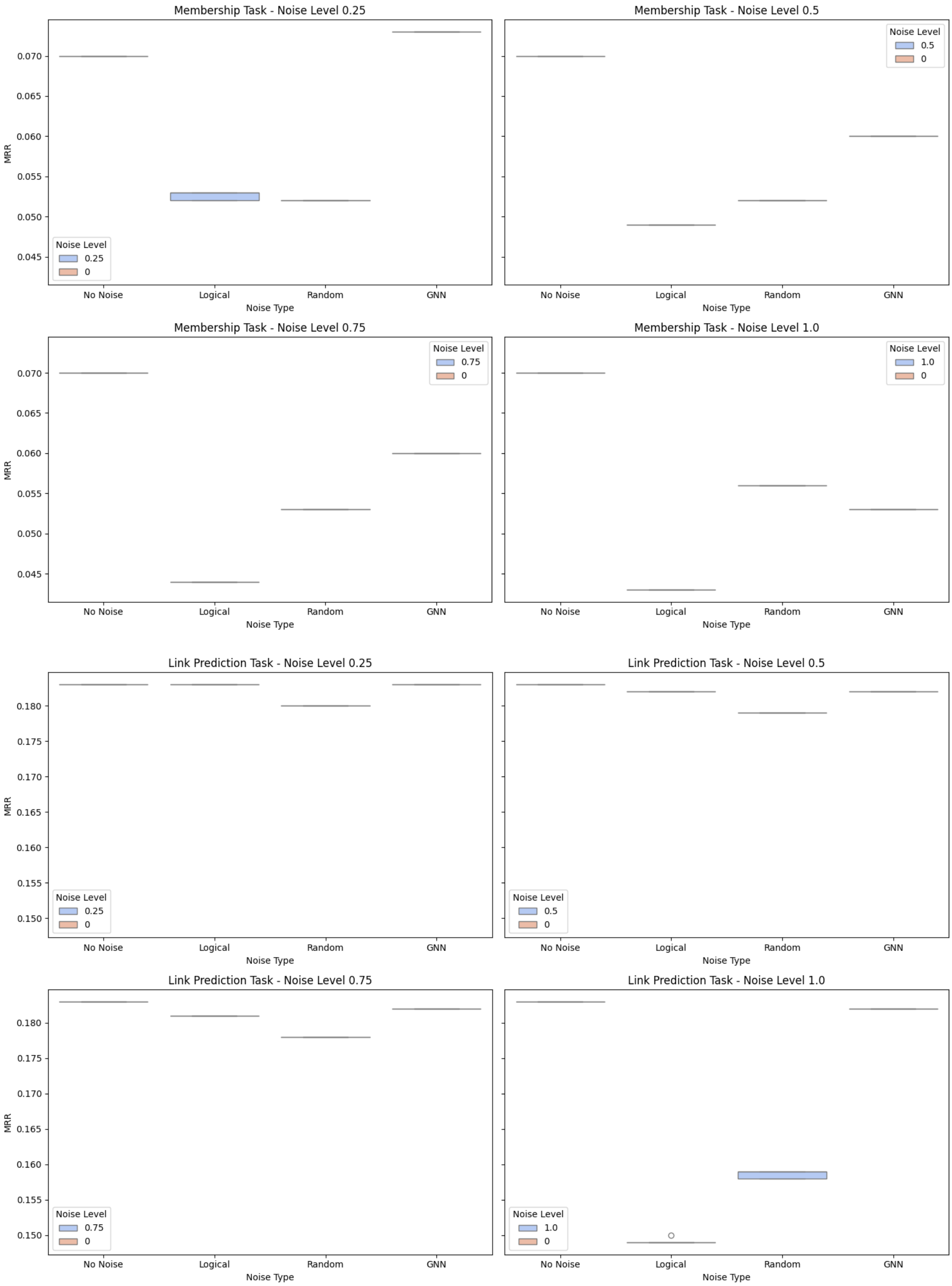


Fig. 4. Variability of MRR on *OWL2Bench1* using *OWL2Vec** [10].

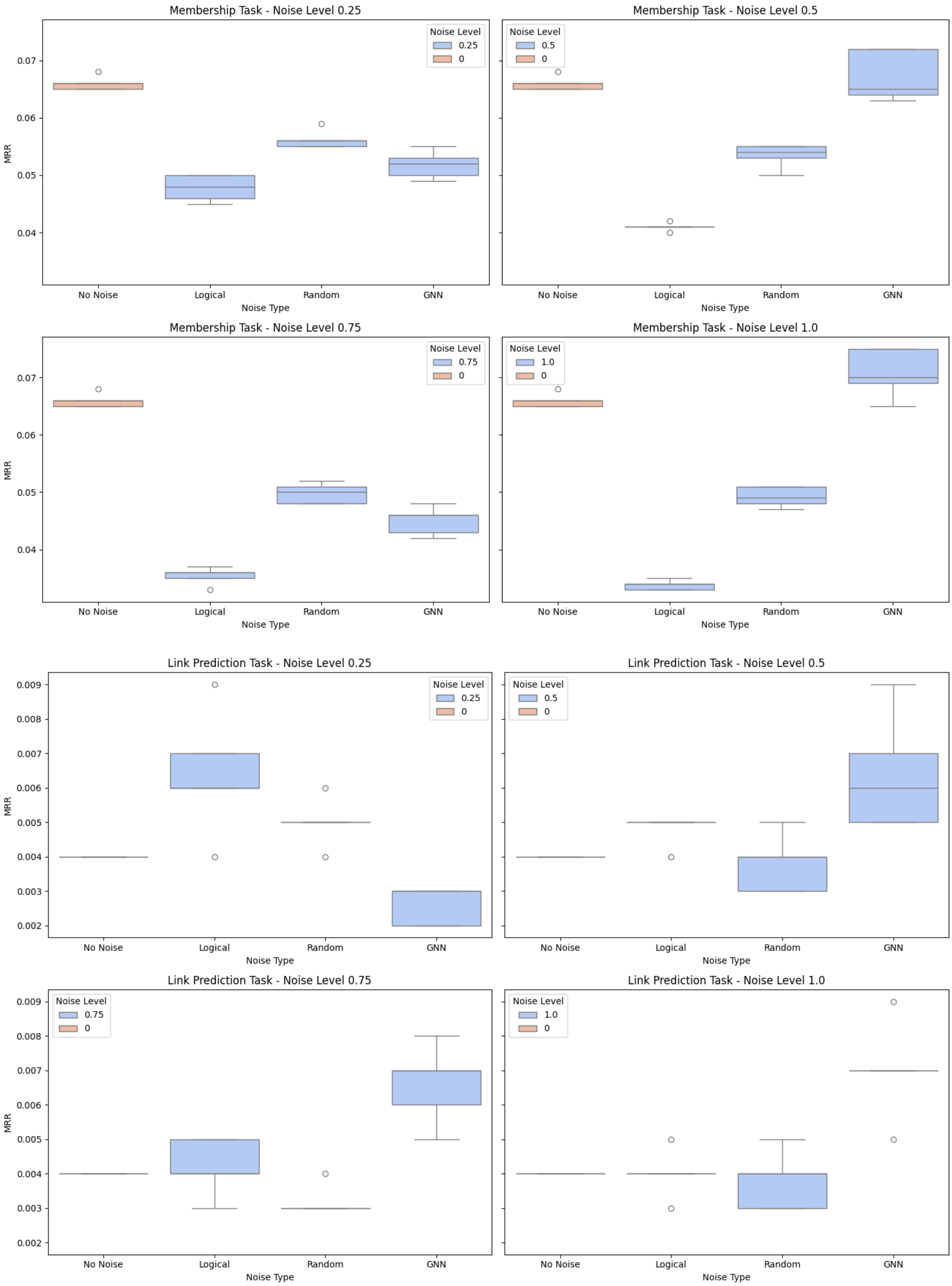


Fig. 5. Variability of MRR on *OWL2Bench1* using *Box2EL* [33].

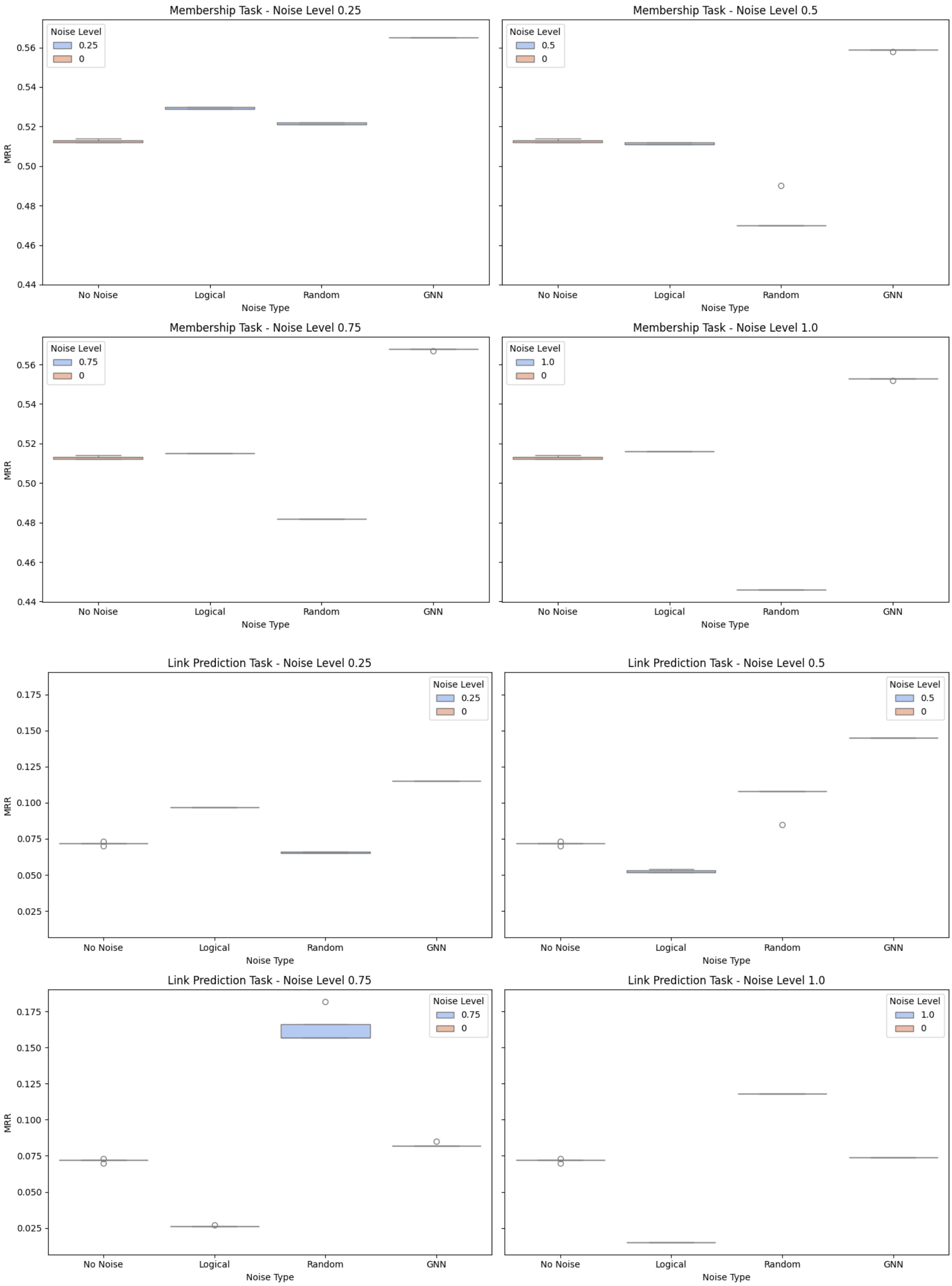


Fig. 6. Variability of MRR on *Family* using *OWL2Vec** [10].

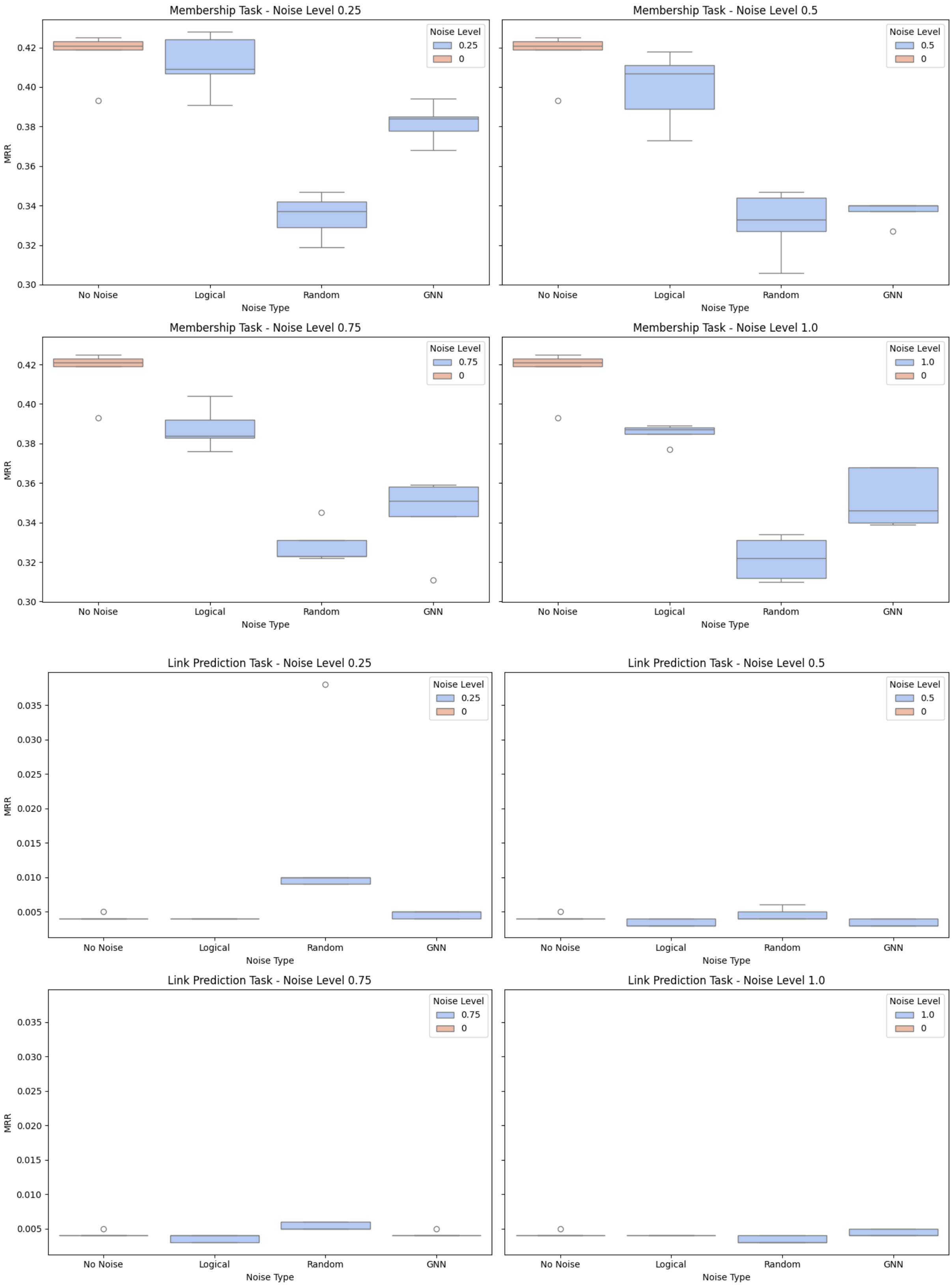


Fig. 7. Variability of MRR on *Family* using *Box2EL* [33].