

# Using Rule Mining and Link Prediction to Understand Species Interaction Networks and to Understand and Predict Invasive Species Behavior

Barbara Meulenbelt<sup>1</sup>  
Supervised by Lise Stork<sup>2</sup>

<sup>1</sup> Vrije Universiteit Amsterdam, [b.a.meulenbelt@student.vu.nl](mailto:b.a.meulenbelt@student.vu.nl)

<sup>2</sup> Vrije Universiteit Amsterdam, [l.stork@vu.nl](mailto:l.stork@vu.nl)

**Abstract.** Intricate networks of interactions between organisms and their environment form the ecosystems that sustain life on earth. Our planet is changing at an alarming rate. This causes disruptions to ecosystems and a major loss of biodiversity. Understanding species interaction networks and understanding and predicting invasive species behavior is a growing imperative for next-generation biodiversity monitoring. The current research uses rule mining on a large species interaction network to detect interesting patterns that can help understand species interaction networks. Link prediction is used as a tool to understand and predict invasive species behavior. Domain expert knowledge was used to evaluate the interestingness of these patterns. Results for rule mining show that the found patterns do well at explaining the interactions in the training data, with an average quality of 0.61 and 0.52 over the top-5 and top-100 patterns, respectively. The patterns have low generalizability, with an average precision of 0.13 and 0.12 over the top-5 and top-100 patterns, respectively. This indicates that our model is overfitting to the training data. The domain experts judged the interestingness of the patterns with with an overall score of 3.06 out of 5. A slight inter-rater agreement is found with an  $\alpha_K$  of 0.105. Results for link prediction show that in 10% of the cases, our model correctly predicts the answer to a query. Rule mining and link prediction are promising tools for the analysis of species interactions networks and invasive species behavior. Further research is needed to enhance model performance.

**Keywords:** Species Interaction Networks · Rule Mining · Link Prediction

## 1 Introduction

*"I am tempted to give one more instance showing how plants and animals, most remote in the scale of nature, are bound together by a web of complex relations."*

– Charles Darwin, *Origins of The Species* [1]

Sharing information about biological diversity is deep seated. Humans have been fascinated by the animal kingdom throughout history. Early examples of humans capturing knowledge of the living world are oral traditions about medicinal plants and animal cave paintings. The emerging computer technology has transformed the way that biodiversity information is shared. The systematic representation, organising, naming, and describing organisms is now done digitally [2].

AI can provide powerful tools for analyzing this widely accessible biological data. Machine learning techniques are emerging as the new standard in computational ecology [3] and particularly in network ecology [4]. Network theory is a branch of mathematics and computer science that studies the dynamics of networks. Network-thinking is permeating studies in ecology and evolution and is one of the fastest growing ecological disciplines [5, 6]. By using network theory, we can investigate questions ranging from the species level to the community level within a formal mathematical framework [7].

Network approaches can be used for a variety of ecological systems, including ecological networks. An ecological network is a representation of all biotic interactions within an ecosystem where species are connected by pairwise interactions [2]. Species interact with other species and their environment directly and indirectly. These networks drive ecological and evolutionary dynamics, and maintain the coexistence, diversity, and functioning of ecosystems [2, 8, 9].

The information captured in these networks can be analysed to find general patterns that give insight into the underlying mechanisms. With a better understanding of these interaction networks, ecologists and biologists can make better informed predictions about the ways different environmental factors will impact ecosystems.

These insights can support the field of invasive biology. This is the interdisciplinary study of the patterns, processes and consequences of the redistribution of biodiversity [10]. Invasive species act as a massive drain on global resources and are one of the major drivers of biodiversity loss worldwide [11]. Understanding and predicting invasive species behavior enhances our ability to understand the spread of invasive species and predict future spread. This way, we can help preserve biodiversity in the face of global environmental challenges.

To reason over these networks, we need enough data. Collecting all possible species interactions is time consuming and expensive given the number of species on our planet. This is why, knowledge on species interactions is one of the biggest bio-diversity data shortfalls [12]. The data that is available on species interactions tends to be biased and noisy [13] and interactions are generally measured as a binary variable, being either present or absent. Species interactions are not binary but occur probabilistically due to variation of ecological interaction networks through space and time [14]. Therefore, assessing the likelihood of species interactions is an imperative for several fields of ecology [15].

The current research aims to address this gap by using rule mining to find general patterns, in the form of logic rules, based on information in a large species interaction network. The species interaction networks will be represented in a

structured and machine-readable format using knowledge graphs. Reasoning over knowledge graphs aims to reveal implicit knowledge through the understanding of existing facts and is therefore becoming a hot research topic [16]. Rule mining can identify patterns and make predictions based on incomplete data. Additionally, rule mining uses a probabilistic approach. This way, we can account for incomplete data, the uncertainty and complexity of ecological systems, and provide estimates of interaction strengths.

The mined rules will be used to serve as general patterns or motifs, species behaving in certain patterns together with other species. These can help us gain a deeper understanding of the relationships between species and reveal the underlying structure of species interaction networks.

We explore the potential of using the tool of link prediction to predict and understand how invasive species behave in relation with other species. Link prediction can help to identify the links or pathways through which invasive species are most likely to spread and which links are most likely to form between invasive species and native species in a network.

### 1.1 Research question & Approach

The main question we aim to answer is: How can rule mining and link prediction help understand species interaction networks and understand and predict invasive species behavior?

Two hypotheses are used in order to answer this question:

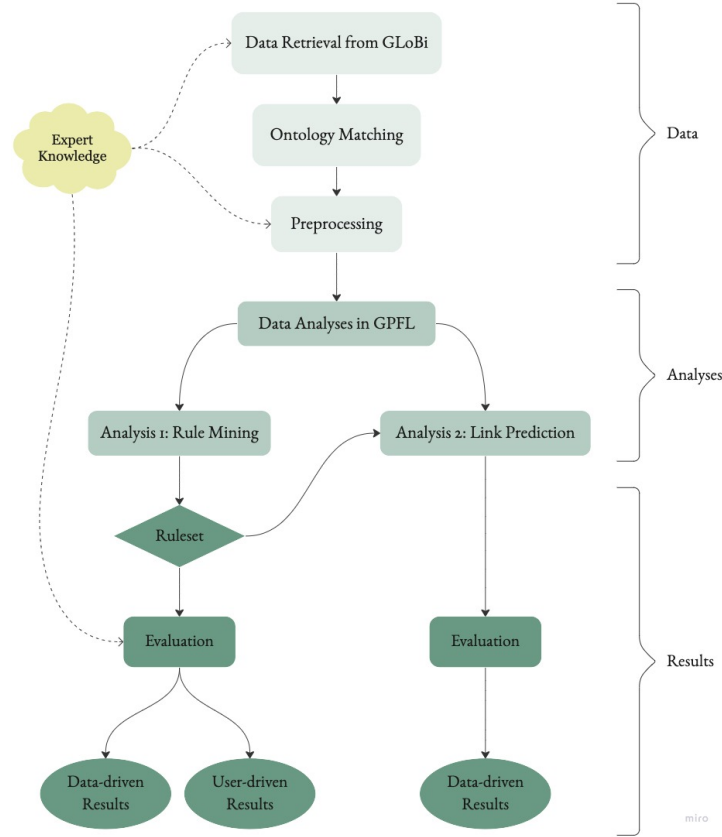
1. Mined rules can act as motifs or interesting patterns for species interaction understanding.
2. The task of link prediction can be used as a tool to predict and understand invasive species behavior.

The flowchart in figure 1 shows all the steps of the research process. The Global Biotic Interactions (GloBI) dataset will be used to retrieve species interactions and taxonomic data [17]. Species from all taxonomic levels will be included (e.g. fungi, animals, and plants). Additionally, all types of interactions that exist between species will be included (e.g. pollination networks, parasitoid web, and seed dispersal networks). Taxonomic information will act as proxies for species phylogenetic relationships to enhance predictive performance. Ontology matching will be used for all species interactions and taxonomic data. This will ensure that all taxa have one unique GBIF identifier (id).

Analyses for rule mining and link prediction will be done with Graph Path Feature Learning (GPFL), a probabilistic logical rule learner [16].

Ensuring that our results deliver their full potential of helping understand species interaction networks, domain expert knowledge will be used at different stages of the research process. We first use domain expert knowledge for making decisions in the data retrieval phase. At the end of the research process, domain expert knowledge is used to evaluate the interestingness of our rules by use of a survey.

Fig. 1: A flowchart that shows the steps of the research process.



## 2 Background

In order to provide a clear context for the current research, this chapter will provide a brief overview of the relevant theory and key concepts in the analysis of ecological networks.

### 2.1 Theoretical Background

#### *Species Interaction Networks*

A species interaction network can represent different types of interactions. These include mutualistic, antagonistic, strong, weak, symmetric versus asymmetric and direct versus indirect [18]. Interactions between species are either trophic (the transfer of energy from the bodies of two organisms), also called a food chain, or symbiotic (interaction between two organisms). Generally an interaction between two species means that *at least* one of the species is affected by the presence of another [18]. Many interaction types exist between species (e.g.

kills/is killed by, eats/is eaten by, has host/host of, pollinates/pollinated by, and epiphyte/epiphyte of) [19]. Ecological interactions can also have an effect on each other [20]. For example, a pathogen might not be able to infect a healthy host, unless the host's immune system is already being compromised by another infection.

### ***Invasive Species***

Invasive species are species that have been redistributed outside their native geographic ranges as a result of human mediated translocation [10]. Invasive species are a global threat to human livelihoods and biodiversity. Increasing globalization promotes the movement of invasive species and environmental changes such as climate change promote the establishment of the invasive species. Since 1500AD, they have contributed to the extinction of more plants and animals than any process other than habitat loss [21]. This is why understanding and predicting the behavior of invasive species is crucial for mitigating their negative impacts on native species in the ecosystems, biodiversity, and human health.

### ***Motifs***

When analysing interactions between nodes in species interaction networks, certain structures and groupings of interactions can be found to emerge around them. Species interaction networks can be broken up into smaller sub graphs of  $n$  species, these are called modules or motifs [22]. Motifs in graphs are repeated sub units of the same structure [22]. Motifs can find species behaving in certain patterns together with other species. Different kinds of motifs can emerge. For example, a linear chain can exist between species A, B and C ( $A \rightarrow B \rightarrow C$ ) or species A and B can have a shared resource C ( $A \rightarrow C \leftarrow B$ ). The distributions of motifs are representative for the type of network. Motifs are considered to be the basic building blocks of complex networks and communities, as they represent typical relationships between species [22–24]. In this research we use logical rule mining to identify patterns of species interactions that occur more frequently. This way, the mined rules can be used for the discovery of motifs in species interaction networks.

### ***Species Classification***

Phylogenetics is the study of the evolutionary history and relationships among and within groups of organisms. Phylogenetics can improve our understanding of grouping structures and the mechanisms underlying species' behavior [25, 26]. Detailed phylogenetic information is not always available and can be difficult to obtain for all interacting species. Taxonomic information gives an easy opportunity to group species based on their evolutionary relatedness. It is for these reasons that in the current research, taxonomic information will be added for all species, acting as a proxy for inferring phylogenetic relationships.

Taxonomy in biology, is the naming, defining and classifying groups of organisms based on shared characteristics [27]. Organisms are divided into taxa and are given a taxonomic rank. The taxonomic rank is the relative level of a group of organisms in an hereditary or ancestral hierarchy. The primary ranks

in which species can be subdivided are domain, kingdom, phylum, class, order, family, genus, and species. This structure creates an overall hierarchy for categorizing life. Domain and kingdom are at the top of the hierarchy. Moving down, the grouping characteristics become more granular. For example, humans are classified as the species *Homo sapiens*. *Homo sapiens* belong to the genus called *Homo*. This genus also includes other species like *Homo erectus* and *Homo neanderthalensis*. The family this genus belongs to is called Hominidae and includes all great apes [28].

## 2.2 Background of the Problem Domain

### Graph Theory

The mathematical formalism of graph theory provides a robust framework to handle and interpret interactions between species [29]. A Knowledge Graph (KG) is a graph-based way of representing knowledge. KGs represent domains that involve interactions between entities, such as social relationships, biological interactions and bibliographical citations [16]. A KG,  $G = (E, R, T)$ , is a directed multi-graph that contains ground atoms (facts) in the form of triples  $T$ . In these graphs,  $r_i \in R$  are the types of relations and  $e_j, e_k \in E$  are the entities. The type of relationship is called the predicate. The entities are called constants and consist of the subject and object. The predicate expresses a binary relation between the subject and the object [16].

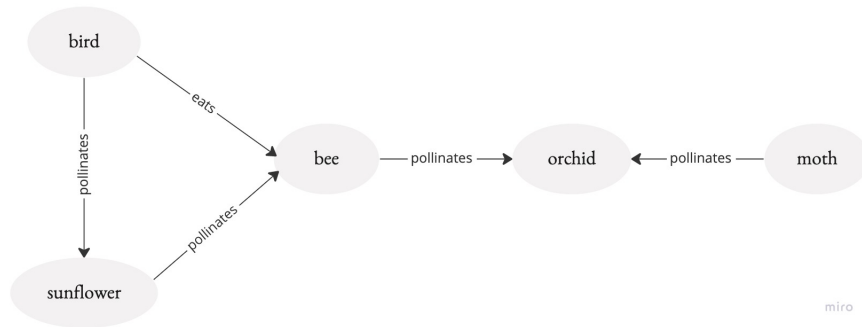
When using a network approach in ecology, species are depicted as nodes and interactions are depicted as links. An example is shown in figure 2. This figure shows an interaction between a bee species and a orchid species. We can write the interaction in figure 2 in logic terms as: *pollinate(bee, orchid)*.

Here, *bee* is the subject, *pollinates* is the predicate, and *orchid* is the object. Figure 3 shows multiple species interactions that together form a small KG.

Fig. 2: A graphical representation of a species interaction.



Fig. 3: Small knowledge graph of species interactions.



**Rule Learning**

Rule-based reasoning over KGs is gaining popularity as it is shown to be interpretable, inductive and transferable [16]. Rule learning methods mine first-order logic rules that are based on information present in the KG.

A Path or Ground Rule is a sequence of ground atoms extracted from the KG. The first atom is called the head atom of the rule and the rest of the atoms are called the body atoms. The count of the body atoms is the length of the rule. A ground rule is written as follows:

$$r_0(e_0, e_1), r_1(e_1, e_2), \dots, r_n(e_n, e_{n+1})$$

A ground rule rewritten into a first-order logical rule is written as:

$$r_0(e_0, e_1) \leftarrow r_1(e_1, e_2), \dots, r_n(e_n, e_{n+1})$$

The head atom is the logical consequence if all the body atoms can be found in the KG. We can differentiate between abstract and instantiated rules. Abstract rules are rules that do not contain constants extracted from the KG and instead refer to variables. Instantiated rules contain constants extracted from the KG. Instantiated rules express and explain concepts in more detail than abstract rules.

The example knowledge graph shown in figure 3 will be used to explain the concepts and show how different type of rules can be created. A closed path that can be derived from figure 3 is:

$$eats(bird, bee), pollinates(bird, sunflower), pollinates(bee, sunflower)$$

This path can be abstracted into a rule:

$$eats(X, Y), pollinates(X, A), pollinates(Y, A)$$

This rule can explain the relationship *eats* when written in logic terms as:

$$eats(X, Y) \leftarrow pollinates(X, A), pollinates(Y, A)$$

The body of the rule is the premises and the head of the rule is the consequence. This is an example of an abstract rule as it does not contain any constants. An instantiated rule can be derived from the following path:

$$eats(birds, bees), pollinates(bee, orchid), pollinates(moths, orchid)$$

We can derive an instantiated rule specifying the correlation pattern between *birds* and *moths* as:

$$eats(birds, Y) \leftarrow pollinates(Y, A), pollinates(moths, A)$$

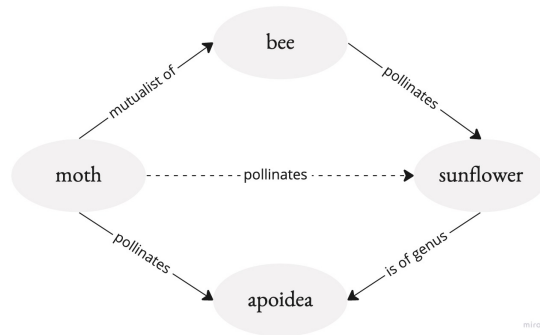
### Link Prediction

Link prediction or Knowledge Graph Completion is the task of constructing missing triples based on known triples from the KG. By altering test triples, a query will take the form  $r_t(e_i, ?)$  or  $r_t(?, e_i)$ . In both type of queries, the type of relation  $r_t$  is known and one of the constants is missing. The missing constant is expected to be replaced by entities that are suggested based on the learned rules [16]. A query where the subject is missing is called a head query. A query where the object is missing is called a tail query. The queries can be composed for all species and relationship types. Figure 4 shows a graphical example for both type of queries.

The predictions for these queries make are listed and ranked based on their confidence. Figure 5 shows an example of a link prediction task. In this figure, a small KG shows known species interactions that infer an interaction between a moth and a sunflower.

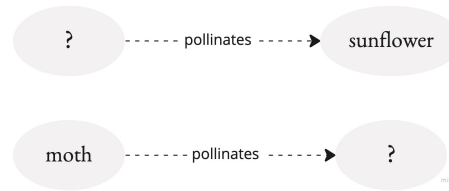
To understand and predict how a species behaves in a network, we can compose queries to infer what interactions a species is most likely to form with other species. For example, if we want to understand and predict what a pandas is likely to eat in a network, we can use a tail query such as  $eats(Panda, ?)$ . The predictions indicate what species are most likely to eat the panda. Verifications for the predictions are given by showing the rules that are used for predictions. The same can be done the other way around. Identifying which species will eat the pandas can be done by composing a head query such as:  $eat(?, Panda)$ . These predictions will indicate what species are most likely to eat a panda.

Fig. 5: Using rules to predict a link between moth and sunflower.



way, researchers can gain insights into the mechanisms driving the spread of invasive species and enhance their ability to prevent future spread.

Fig. 4: Example head query (top) and tail query (bottom).



These examples show the potential of using link prediction in invasive biology by obtaining insight in an invasive species or a species that potentially will invade a new area. Queries can be composed to understand and predict how the species will likely behave with native species in the network. The verifications for the predictions can tell us how and why the interaction between the invasive species and the native species are formed. This



**GPFL**

Graph Path Feature Learning (GPFL) is a probabilistic logical rule learner that is optimized to mine instantiated first-order rules containing constants extracted from the KG [16]. GPFL mines more quality rules compared to the gold standard rule learner AnyBURL [16, 30]. GPFL generates both abstract rules and instantiated rules and uses a two-stage rule generation mechanism. A rule produced by GPFL has some general components:

$$r_t(X, Y) \leftarrow r_1(X, V_0), r_2(V_0, V_1), \dots, r_n(V_n, V_{n+1})$$

Lowercase letters refer to constants and uppercase letters refer to variables. The  $X$  and  $Y$  are variables in the head atoms.  $X$  is the original variable as the body atoms originate from it and  $Y$  is the free variable.  $V_0$  and  $V_i$  are the variables in the body atoms of the rule.  $V_0$  is a connecting variable as it connects adjacent atoms,  $V_n + 1$  is a non-connecting variable. A rule is closed if  $Y$  is also the tail variable and open if the  $Y$  does not occur in the body atoms of the rule. GPFL produces four types of rules that are displayed below.

**Template:**  $r_t(X, Y) \leftarrow r_1(X, V_0), \dots, r_n(V_n, V_{n+1})$

**HAR:**  $r_t(X, e_k) \leftarrow r_1(X, V_0), \dots, r_n(V_n, V_{n+1})$

**BAR:**  $r_t(X, e_i) \leftarrow r_1(X, V_0), \dots, r_n(V_n, e_j)$

**CAR:**  $r_t(X, Y) \leftarrow r_1(X, V_0), \dots, r_n(V_n, Y)$

A template is an open acyclic abstract rule. Templates are not included in the rule set because they are too general. A Head Anchored Rule (HAR) specialises the template by substituting the free variable  $Y$  with a constant. HAR rules can be used to identify entities that have a relation type  $r_t$  with entity  $e_k$  by a pattern. A Both Anchored Rule (BAR) specialises the HAR by replacing the tail variable with a constant. BAR rules can show patterns that involve the correlation between  $e_i$  and  $e_j$ . A Closed Abstract Rules (CAR) are closed cyclic rules of predicates connecting entity pairs. CAR rules are often small in size but give good base predictive performance [16].

Rule generation in GPFL is divided into two phases. The first phase is the generalization for abstract rules. The second phase is the specialization for instantiated rules. The algorithm for the generation phase is shown in figure 6a. Here, a set of frequent templates are created by saturating the template space. This template space is usually much smaller in size than when we use the complete rule space. Therefore, template saturation is easier to converge.

The specialization phase is shown in figure 6b. Here, instantiated rules are derived and evaluated by using the template groundings from the generation phase. Rules that are derived from the same template are similar in their structure. Therefore, instead of grounding all individual rules, only the templates are grounded. The templates groundings are used to evaluate the instantiated rules.

All rules generated by GPFL are saved in the output file *rules.txt*. In this file, all rules are scored with five metrics, support (*supp*), body groundings (*BG*),

confidence (*conf*), head coverage (*HC*) and validation precision (*VP*). A description and calculation for each metric is given below.

1. *Supp* is the number of correct predictions the rule suggests over the training data.
2. *BG* is the number of possible groundings of the body atoms of the rule (total predictions).
3. *Conf* is calculated by dividing the *supp* by the *BG*. This measure is used as quality measure throughout the analyses. We handle a smooth confidence (*SMC*) with  $\eta = 5$ . The *SMC* for a rule  $l$  is calculated by:

$$SMC(l) = \frac{Supp}{\eta + BG} \quad (1)$$

4. *HC* is calculated by dividing the support by the number of positive instances of the head atom of rule  $l$ :

$$HC(l) = \frac{Supp}{|rt|} \quad (2)$$

Here,  $rt$  is the head atom of rule  $l$  and  $|rt|$  is the number of positive instances of the head atom  $rt$ .

5. *VP* is the precision ( $P$ ) of a rule over the validation set and is calculated by dividing the *supp* of rule  $l$  in the validation test by the *BG* of rule  $l$  over the validation set. The (test) precision rule  $l$  is calculated by:

$$P_{test}(l) = \frac{Supp_{test}}{BG_{test}} \quad (3)$$

The validation precision of a rule is calculated in the same way:

$$P_{valid}(l) = \frac{Supp_{valid}}{BG_{valid}} \quad (4)$$

Because instantiated rules are more specific, there is a bigger chance of overfitting the training data. A validation method is used to remove the overfitting rules. This makes for better predictive performance [16]. Overfitting rules generally have high quality measures on the training set but have low test precision. Therefore, a rule is considered to overfit if the test precision of the rule is smaller than 10% of the quality of the rule.

### 3 Related Work

Networks approaches have been used in disease ecology. It has shown to be useful in predicting a multitude of infectious diseases in livestock and wildlife [31], and predicting the risks and dynamics of both dengue [32] and Chagas disease [32]. There are existing methods for predicting species interactions. These include predicting species interaction on basis of functional trait matching [33], position in the trophic niche [34], phylogenetic distance [27] and interaction frequency [35].

These methods are commonly based around a single mechanism at a single scale. Species interaction networks are the product of mechanisms across different scales including ecological processes, evolutionary history, individual behavior and group dynamics [2, 15]. When examining these networks on the basis of one mechanisms or scale, the complexity of the networks remains hidden. Additionally, these methods do not give estimates of interaction strength while we know that species interactions networks occur probabilistically.

Logical rule learners can help identify complex interactions between species that occur across multiple mechanisms and scales. Applications where first-order logic rules have been used have been shown to be successful are question answering [36] fact checking [37] and knowledge graph completion [30]. Rule learners can be categorized by the types of rules that they produce. ScaLeKB [38] and QuickFOIL [39] produce Horn rules to deduce unknown facts; RuDiK [24] offers negative rules, these can be used to identify contradictions in the data. RuLES [40] discovers non-monotonic rules for exception handling; AMIE+ [41] and AnyBURL [30] contain instantiated rules that include constants to enrich the expressivity of the rule space. The rule learner used in the current research, GPFL, focuses on mining probabilistic positive Horn rules. These rules are optimized to discover instantiated rules [16].

Wilcke and colleagues (2019) used rule mining for domain understanding in archaeology. In their work, an end-to-end pipeline for user-centric pattern mining on knowledge graphs in the humanities is developed, implemented and evaluated [42]. Results are evaluated both data-driven and user-driven. To evaluate the *interestingness* of the produced rules, a survey was used in which domain-experts could judge the rules. *Interestingness* of the rule was based on three metrics; plausibility, relevancy, and newness. To asses *interestingness* of the rules in the current research, the same approach and metrics will be used.

Fig. 6: Algorithms for the rule generation and generalization phase.

(a) Rule Generation phase

(b) Generalization phase

Algorithm 1: Rule Generation for a Target Predicate	Algorithm 2: Generalization
<b>Input</b> : $\mathcal{G}, I, sat, bs, len$ <b>Output</b> : learned rule set $F$	<b>Input</b> : $\mathcal{G}, I, sat, bs, len$ <b>Output</b> : rule frequency map $M$
1 Initialize empty set $F$ ; 2 $M \leftarrow \text{Generalization}(\mathcal{G}, I, sat, bs, len)$ ; 3 $L \leftarrow \text{Sort}(M)$ ; 4 <b>for</b> $l \in L$ <b>do</b> 5 $G \leftarrow \text{Ground}(\mathcal{G}, l)$ ; 6 <b>if</b> $l$ is a CAR <b>then</b> 7 $\text{Score}(l, G)$ ; 8 $F \leftarrow F \cup l$ ; 9 <b>else</b> 10 $S \leftarrow \text{Specialization}(l, G, I)$ ; 11 <b>for</b> $s \in S$ <b>do</b> 12 $\text{Score}(s, G)$ ; 13 $F \leftarrow F \cup s$ ; 14 <b>end</b> 15 <b>end</b> 16 <b>if</b> $\text{Constraints}()$ <b>then</b> 17 <b>Break</b> ; 18 <b>end</b> 19 <b>end</b> 20 $F \leftarrow \text{Quality}(F)$ ; 21 <b>return</b> $F$ ;	1 Initialize empty set $T$ and map $M$ ; 2 $c, sat' \leftarrow 0$ ; 3 <b>do</b> 4 $i \leftarrow$ randomly sample an instance from $I$ ; 5 $P \leftarrow \text{PathSampler}(\mathcal{G}, i, len)$ ; 6 <b>for</b> $p \in P$ <b>do</b> 7 $c \leftarrow c + 1$ ; 8 $t \leftarrow \text{Abstraction}(p)$ ; 9 $T \leftarrow T \cup t$ ; 10    Update $M$ with $p$ ; 11 <b>if</b> $\text{mod}(c, bs) = 0$ <b>then</b> 12 $sat' \leftarrow \frac{ M.keys \cap T }{ T }$ ; 13 $T \leftarrow \emptyset$ ; 14 <b>if</b> $sat' > sat$ <b>then</b> 15 <b>Break</b> ; 16 <b>end</b> 17 <b>end</b> 18 <b>end</b> 19 <b>while</b> $sat' < sat$ ; 20 <b>Return</b> $M$ ;

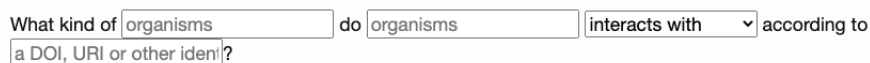
## 4 Materials & Methods

### 4.1 GloBI

GloBI<sup>3</sup> is a database of species interactions based on literature. GloBI provides open access to finding species interaction data (e.g. parasite-host, pathogen-host, pollinator-plant, and predator-prey) by combining existing open datasets using open source software. Currently, GloBI includes 8,933,091 references obtained from 328 data sources. In total, 15,128,089 interaction records were discovered, covering 875,369 taxa [43] (retrieved on 28 October 2022). The OBO Relations Ontology [19] is used for interaction terms.

A interaction row in the GloBI dataset has a total of 92 columns with information regarding the interaction between two taxa. The first taxon, or object in the interaction is called the *source taxon*, the second taxon, or subject is called the *target taxon*. The information in the columns regards the study, specimen, taxon, and location concepts. The taxonomic coverage of the GloBI dataset varies widely between datasets [43]. The density of interaction studies is highest in the Gulf of Mexico, the North Sea and Weddel Sea. North America and Europe provide the most coverage in the density of data sources [17]. The data available on GloBI can be downloaded and browsed through via their website. Figure 7 and 8 show the user interface on the website. A target and source taxon can be typed in along with an interaction term. Additionally, the source of the interaction can be identified. All possible interactions are listed.

Fig. 7: GloBI interface for querying.



What kind of  do   according to

### 4.2 Data preprocessing

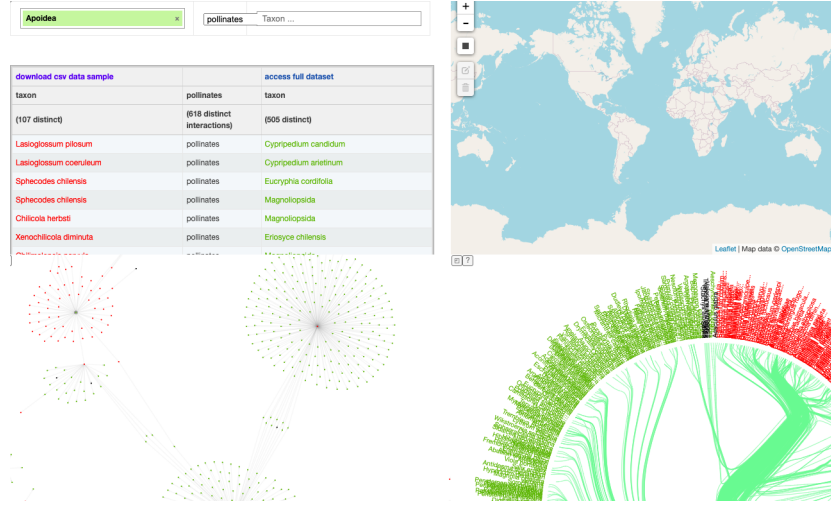
#### 4.2.1 Retrieval & matching

##### *Access to GloBI database*

Species interaction data from GloBI can be accessed by use of the GloBI search or browse pages, REST-y GloBI Web API, R, Cypher, SPARQL or python [43]. Due to encountering technical problems by using querying through Cypher, SPARQL and R, the decision was made to download the entire, raw *interactions.csv* file. The data was retrieved with commandline code and the preprocessing was done with Python.

<sup>3</sup> <https://www.globalbioticinteractions.org/>

Fig. 8: GloBI interface for querying.



### *Species names and unique identifiers*

The main challenge with using the raw interactions data set was to get a unique identifier for all taxa and their higher taxonomic ranks. Taxonomic identification is an important problem for the integration and comparison of different datasets across all fields of biology, ranging from genomics to ecology. Scientific species names are imperfect identifiers for organisms. Even though species names are formalized and validated conforming to strict codes of nomenclature, taxonomists can use the same name to different taxonomic views of the characteristics of a taxon [44]. As the meaning or context of the names are not recorded within the names, datasets can not be reliably integrated on the basis of scientific names.

It is important to take this problem into account when applying our rule learner as it affect our predictions and conclusions. It is for these reasons that the decision was made to retrieve one unique identifiers from a single species ontology.

### *Ontology Matching*

Ontologies are controlled vocabularies in specific areas that define and identify different concepts and the labels (terms) to refer to them. The terms used are defined and maintained by groups of experts of the specific domain of the terminology. Biological or ecological ontologies can be used to define and identify species and the relationships between them.

The GloBI dataset gives a set of different available species identifiers from different ontologies including EOL, WoRMs, ITIS, NCBI and GBIF [17]. A species identifier often gives the name of the used ontology along with a unique code for the taxon (e.g. Giant Panda has code: GBIF:9387176 in the GBIF ontology and ITIS:621845 in the ITIS ontology). All datasets that are included in GloBI have a certain way of identifying the species encountered in their research. Because

many naming databases exist, one species has different identifiers (e.g. homo sapiens is GBIF:2436436, ITIS:180092 and EOL:327955). This is also known as the heterogeneity problem [45].

Ontology matching is the process of determining correspondences between concepts in ontologies [46]. We will use this to find all available identifiers for a given species name using Nomer [47]. From these different identifiers, we can choose which one to use. This will be done for all taxa and their higher taxonomic ranks in order to deal with the heterogeneity problem.

### ***Ontology matching with Nomer***

Nomer<sup>4</sup> can be used to show, match, append, and replace information by mapping species identifiers and names to other species identifiers and names [47]. If we have a certain species name or identifier, we can use Nomer to find properties of the species in different available ontologies. Nomer can match species names with 38 different ontologies (e.g. DiscoverLife, WORMS, GulfBase, Inaturalist, ITIS, Plazi and GBIF). It expects tab separated input in form of [term id] \t [term name]. Columns to be used for identifier/name selection are defined in 'nomer.schema.' properties.

From the 38 available ontologies in Nomer, a choice had to be made which one to choose. Ontologies all have characteristics that can make them more or less suited as they are a context-dependent projection of reality. Biological/ecological ontologies differ in their taxonomic and spatial coverage of species included. As they all differ and come with bias, the decision was made to use the ontology with which the least data was lost. Using the GBIF ontology [28] resulted in the maximum number of remaining interaction rows and was therefore chosen. Many species are identified with more than one GBIF identifier, in these cases, the first identifier was chosen.

### ***Retrieving and matching the interaction data***

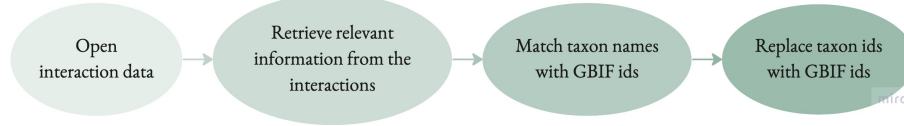
As the matching of millions of names to their GBIF identifier requires a lot of memory, a remote SSH was used to retrieve all data and do the matching with Nomer [47]. Before matching, the GloBI's stable versioned integrated interaction data published on Zenodo was downloaded to serve as our dataset [43]. The code to retrieve this data can be found in Appendix 1.A.

The entire dataset contained 15,128,089 interaction rows. Of the total 92 columns in GloBI, seven columns needed to be included in our research, five of these are; *sourceTaxonId*, *sourceTaxonRank*, *interactionTypeName*, *targetTaxonId* and *targetTaxonRank*. For matching the names with identifiers we also need to include *sourceTaxonName* and *targetTaxonName*. There are two reasons for choosing *InteractionTypeName* instead of *InteractionTypeId*. Firstly, this enhances interpretability of the results. Secondly, most interaction types are based on the OBO Relations Ontology [19] but a few are not. Those give as *interactionTypeId* a 'no:match'. Interactions that are not based on the OBO Relations Ontology are: *endoparasitoidOf*, *livesOn*, *livesNear*, *guestOf*, *livesInsideOf*,

<sup>4</sup> <https://github.com/globalbioticinteractions/nomer>.

*ectoParasitoid*, *farms*, *inhabits*, *hasDispersalVector*, and *livesUnder*. The code to retrieve GBIF identifiers for all taxa and their taxonomic level can be found in Appendix 1.B. Figure 9 shows the steps for retrieving the data.

Fig. 9: Steps for retrieving the data.



The entire dataset was opened and the seven columns were retrieved. Then, Nomer used the *sourceTaxonName* to match and replace the *sourceTaxonId* with a GBIF identifier. Then, the same is done for the target taxon, resulting in a dataset only containing species with a GBIF identifier. The remaining dataset contained 5,264,588 interaction rows.

#### 4.2.2 Augmentation

As discussed in *Species Classification* in section 2.1, the current research uses taxonomic information as a proxy for phylogenetic relationship. The taxonomic information we aimed to include in our model are the identifiers of the higher taxonomic ranks each taxa belongs to. The ranks that were chosen too include are genus, family and order. The choice was made exclude ranks higher than order, as these ranks become too general.

GBIF identifiers were again used to make sure every taxa was identified with one unique genus- family- and order identifier. The code to retrieve these identifiers is shown in Appendix 1.B. This code shows how the genus identifiers for all source taxa was retrieved. The *sourceTaxonName* was again used to match and replace the *sourceTaxonId* with a GBIF identifier. Then, the name of the genus (specified in *genus.properties*) was used to match and append a GBIF identifier for the genus. The order and family identifiers were retrieved in the same way and these were all also retrieved for the *target taxon*.

#### 4.2.3 Removal of higher level interactions.

Within the GloBI dataset, interactions occur between taxa from all taxonomic levels. This means a kingdom can be found interacting with a genus (e.g. Animalia eats Panda) or a species with a phylum (e.g. *Phormia regina* (black blow fly) eats Chordata) [43]. These examples show that interactions like these might be too general to teach us anything interesting<sup>5</sup>.

<sup>5</sup> The reason for the existence of interactions between these higher ranks can be due to the granularity of taxonomic resolution of the specific datasets or due to the specific circumstance in which the interaction claim was documented.

It is for these reasons that the decision was made to exclude all interactions that have *at least* one taxon with a taxonomic rank higher than species. These interaction rows were deleted from the data. Interactions involving taxa with ranks lower than species were kept, as they give us information on a fine-grained level. Deleting these interactions left us with a dataset of 3,120,566 interaction rows. Finally, all duplicate interaction rows were deleted and interaction rows where either one of the three essential columns, *sourceTaxonId*, *interactionTypeName*, and *targetTaxonId*, had a missing value. This left us with a remaining dataset of 986,219 interaction rows. Initially, some taxa were classified with multiple taxonomic ranks (i.e. GBIF:6293214 is classified with rank subspecies, species and variety). All duplicate rows were removed and the first row was kept (i.e. GBIF:6293214 was thus ranked as subspecies).

### 4.3 Dataset statistics

#### 4.3.1 Basic statistics

Many rows of the interaction data are lost due ontology matching and the data preprocessing procedure. The initial total file consisted of 15,128,089 interaction rows. After ontology matching, 5,264,588 interaction rows remained. After the preprocessing procedure, this number was reduced to 986,219 interaction rows. The additional taxonomic information adds a total of 507,175 rows to the data, resulting in a complete dataset of 1,493,394 interaction rows.

The file was split into train, test and validation sets with a split ratio of 0.8:0.1:0.1. After splitting, the training data consisted of 1,296,137 interaction rows, the test set of 98,639 and the validation of 98,618. The statistics for the dataset before and after splitting are shown in table 1.

The degree of 6.85 is calculated by dividing the total number of interactions by the total number of nodes in our network. This is the mean number of interactions per node. The density of 37334.85 is calculated by dividing the total number of relationships by the number of relationship types. This is the mean number of relationships per relationship types.

The amount of nodes shown in table 1 includes both the *taxonId*'s and the genus, family and order identifiers that are added for taxonomic information. Of these nodes, 170,541 are the distinct taxa of rank species ( $n = 161,034$ ), subspecies ( $n = 3,528$ ), varieties ( $n = 1,205$ ), and forms ( $n = 31$ ). A total of 4,743 taxa have an unknown taxon rank. The other nodes ( $n = 46,519$ ) are the genus, family, and order identifiers. Not for all taxa this information was available, eight taxa have a missing genus identifier, 101 have a missing family identifiers, and 1,797 have a missing order identifier.

Table 1: General statistics of the dataset before and after splitting.

Metric	All data	Train data
Relationships	1,493,394	1,296,137
Density	37334.85	32403.43
Degree	6.85	5.94
Nodes	218,094	218,094
Relationship Types	40	40



A total of 40 relationship types exist in our data, three of those are used to specify taxonomic information (i.e. *isOfGenus*, *isOfFamily*, and *isOfOrder*) and 37 are the biological relationship. Table 2 shows the number of interactions in the dataset per relationship type. The number of occurrences vary widely between the relationship types. The relationship type with the highest occurrence is *eats* and occurs in almost a third of all interactions (31.85%). The one with the lowest occurrence is *inhabits* (<0.0001%).

Table 2: Number of interactions per relationship type (before splitting).

Relationship type	n	%	Relationship type	n	%
eats	311,165	31.55	coRoostsWith	778	0.07
interactsWith	190,279	19.29	coOccursWith	764	0.07
hasHost	155,189	15.74	hasHabitat	543	0.06
parasiteOf	102,777	10.42	parasitoidOf	439	0.04
preysOn	90,463	9.17	livesOn	311	0.03
visitsFlowersOf	48,414	4.91	commensalistOf	299	0.03
pathogenOf	16,272	1.65	kills	272	0.03
pollinates	14,621	1.48	createsHabitatFor	259	0.03
ectoparasiteOf	10,285	1.04	livesInsideOf	235	0.02
adjacentTo	9,622	0.96	guestOf	53	<0.001
visits	8,150	0.83	livesNear	49	<0.001
endoparasiteOf	6,298	0.64	hyperparasiteOf	45	<0.001
mutualistOf	5,972	0.61	ectoParasitoid	44	<0.001
ecologicallyRelatedTo	3,270	0.33	laysEggsOn	17	<0.001
hasVector	2,838	0.29	livesUnder	15	<0.001
hasDispersalVector	2,052	0.21	kleptoparasiteOf	13	<0.001
symbiontOf	1,817	0.18	providesNutrientsFor	6	<0.0001
epiphyteOf	1,667	0.17	inhabits	3	<0.0001
endoparasitoidOf	923	0.09			

#### 4.3.2 Taxonomic coverage

To see how the species in our dataset are divided over the taxonomic levels, the taxonomic coverage is shown in table 3. The seven kingdoms are shown with the number of taxa in the dataset that belong to each of them, along with the relative proportion to the total taxa. The kingdom could not be retrieved for all taxa, therefore 27,060 taxa are unaccounted for in the table. The total amount of species, subspecies, forms and varieties per kingdom available on GBIF serves as a baseline for comparison. The most notable difference in taxonomic coverage is that our dataset seems to have much lower proportion of taxa of kingdom Fungi than baseline, 20.73% and 35.71% respectively and a much bigger proportion of kingdom Plantae than baseline, 18.65% and 6.95% respectively. We also see that 'other' kingdoms contain 3.98% of the total taxa whereas in baseline this is only 0.02%. In our dataset, 'other' kingdoms include Metazoa (n = 4222),

Viridiplantae (n=1509), Protista (n = 38), Protoctista (n = 35), Archaeplastida (n = 28), Monera (n = 1). In the GBIF data, 'other' kingdoms include incertae sedis (n=412) and (viruses = 11).

To compare the ranks of the taxa that are included in our dataset to baseline, table 4 shows the number of taxa that belong to rank species, subspecies, variety and form. The numbers are given for the current dataset and the GBIF dataset. The table shows that almost all taxa in our dataset are of rank species. Additionally, our dataset covers less taxa of rank subspecies and variety than baseline (more notably for variety). The 10 most occurring taxa are shown in table 5.

Table 3: Taxonomic coverage of the data compared to baseline (GBIF).

Kingdom	Dataset (n)	%	GBIF (n)	%
Animalia	79694	53.85	1506009	54.24
Fungi	30682	20.73	991355	35.71
Plantae	27593	18.65	192930	6.95
Bacteria	2342	1.58	59380	2.14
Chromista	1041	0.70	22057	0.80
Protozoa	731	0.48	3665	0.13
Archaea	8	<0.01	621	0.02
Other	5890	3.98	423	0.02

Table 4: Ranks of taxa included compared to baseline (GBIF).

Taxon rank	Dataset (n)	%	GBIF (n)	%
species	161034	97.13	2453731	88.38
subspecies	3528	2.13	163831	5.90
variety	1205	0.73	141122	5.08
form	31	0.20	17756	0.64

Table 5: The 10 most occurring species in dataset.

TaxonId	n	Scientific Name	Kingdom
GBIF:1341976	4505	<i>Apis mellifera</i>	Animalia
GBIF:2436436	4366	<i>Homo sapiens</i>	Animalia
GBIF:9079676	2398	<i>Pinus sylvestris</i>	Plantae
GBIF:7591641	1902	<i>Fraxinus excelsior</i>	Plantae
GBIF:5219243	1604	<i>Vulpes vulpes</i>	Animalia
GBIF:2685796	1350	<i>Pseudotsuga menziesii</i>	Plantae
GBIF:2441022	1324	<i>Bos taurus</i>	Animalia
GBIF:9221087	1324	<i>Achillea millefolium</i>	Plantae
GBIF:3189846	1303	<i>Acer platanoides</i>	Plantae
GBIF:5219173	1248	<i>Canis lupus</i>	Animalia

## 4.4 Experiments

### 4.4.1 Quantitative analysis.

#### *Experimental set-up*

GPFL takes a datafile containing triples as input. Triples consist of 3 columns containing the object, predicate and subject, separated by tabs. For the model, two types of triples will be combined. Firstly, the interaction triples. These triples are formed by the *sourceTaxonId*, the *interactionTypeName*, and the *targetTaxonId*. The second type of triple includes taxonomic information for all taxa. The first two columns of these triples are formed by the *taxonId* for all taxa and the relation to higher taxonomic ranks, *isOfGenus*, *isOfFamily* and *isOfOrder*. The third column consists of the identifiers for these higher ranks.

Both type of triples together form the complete dataset, containing 1,493,394 interaction rows. The complete recording (i.e. interaction and taxonomic information) of example interaction *Homo Sapiens* eats *Bos Taurus* in our data looks as follows:

<b>taxonId</b>	<b>interaction</b>	<b>taxonId</b>
GBIF:2436436	eats	GBIF:2441022
GBIF:2436436	isOfGenus	GBIF:2436435
GBIF:2436436	isOfFamily	GBIF:5483
GBIF:2436436	isOfOrder	GBIF:798
GBIF:2441022	isOfGenus	GBIF:2441017
GBIF:2441022	isOfFamily	GBIF:9614
GBIF:2441022	isOfOrder	GBIF:731

The data was analysed using GPFL with the configuration settings shown in table 6a. The maximum length of both instantiated rules (*ins depth*) and closed abstract rules (*car depth*) was specified. A threshold was set for the confidence (*conf*), number of correct predictions (*supp*), head coverage (*head coverage*) and template saturation (*saturation*). Lastly, a number was specified for the size of the batch over which the saturation is evaluated (*batch size*) and for the running threads (*thread number*).

A smooth confidence with  $\eta = 5$  is used as the quality measure. Gu et al., 2020 show that smooth confidence produces better results than the other quality measures as it has a smaller overfitting proportion while maintaining a larger rule space [16]. For the smooth confidence,  $\eta$  is an offset used to cope with the bias that assigns high confidence to rules that make only a few predictions.

To limit the runtime, some additional settings were used, these are shown in table 6b. A maximum is set for the number of groundings for evaluating a rule during learning (*learn groundings*), and for evaluating a rule during application, (*apply groundings*). The *suggestion cap* sets the maximum number of predictions a rule can make during application and the *ins rule cap* sets the maximum number of instantiated rules that can be derived from a template. A maximum runtime in seconds is specified for the essential rule generation procedure (*essential time*), the generalization procedure (*gen time*), and the specialization procedure (*spec time*). Lastly, 10 random walkers are used to sample paths.

Table 6: Configuration settings for GPFL.

(a) Required configurations settings. (b) Time and space constraints.

Option	Setting	Option	Setting
support	3	random walkers	10
head coverage	0.001	ins rule cap	10000000
conf	0.001	suggestion cap	10000000
ins depth	3	gen time	120
car depth	3	essential time	60
saturation	0.99	spec time	240
batch size	100000	learn groundings	5000
thread number	6	apply groundings	3000
overfitting factor	0.1		
quality measure	smoothedConf		

### Results

All experiments are conducted on a computer with 8 cores and 32GB RAM. GPFL produced a total of 2,536,397 rules. An overfitting analyse was conducted to improve rule performance [16]. After overfitting rules were deleted, 130,686 rules remained with 31 CAR, 130,529 BAR and 126 HAR rules. The amount of rules per rule type is given in table 7. This table also shows the amount of rules per rule length along with the relative frequencies. In the configuration settings, the maximum length for both instantiated rules and closed abstract rules was set at 3. GPFL produced rules of length 1 and 2 for all three types of rules. Only CAR rules of length 3 are produced. The biggest proportion of the HAR and BAR rules have a length of 1.

Table 7: Length of rule per type of rule.

Type of rule	Total rules	len = 1		len = 2		len = 3	
		n	%	n	%	n	%
HAR	126	117	92.86	9	7.14	-	-
BAR	130,529	98,803	75.69	31726	24.31	-	-
CAR	31	5	16.13	3	9.68	23	74.19

Two relationship types didn't produce any rules, these are *inhabits* and *laysEggsOn*. Due to deletion of overfitting rules, the relationship types *kleptoparasiteFf*, *guestOf*, *livesNear*, *livesUnder* and *providesNutrientsFor* also had no remaining rules. Table 9 shows for every remaining relationship type, the number of rules it appears in the head atom of, along with the relative frequencies. The relationship type that occurs the most in the rules is *visitsFlowersOf* (n = 23,080). The relationship type that occurs the least is *livesInsideOf* (n = 1).

Table 8: High quality (*HQ*) rules and extremely high quality (*EHQ*) rules per rule type.

Type of rule	<i>HQ</i> rules		<i>EHQ</i> rules	
	n	%	n	%
HAR	100	-	-	-
BAR	127,492	96.67	720	0.55
CAR	20	64.52	4	12.90

We follow the method used in [41] and [48] to classify rules into: *high quality rules* (*HQ*), rules with a confidence  $\geq 0.1$  and head coverage  $\geq 0.01$  and *extremely high quality rules* (*EHQ*), rules with a confidence  $\geq 0.7$ . Table 8 shows per type of rule, the number of *HQ* and *EHQ* rules. We can see that the largest part of the produced

rules consist of high quality rules. Table 10 shows the global average precision and quality over the top rules with and without validation (see *GPFL* in section 2.2 for calculation of these metrics). We can see that the quality for all top-k scores is much higher than its precision. The results for using *GPFL* for link prediction are as shown in table 11. The numbers in table 10 and 11 are the average over all relationship types and rule types.

Table 9: The number of times a relationship type appears in the head atom of a rule.

Relationship type	n	%	Relationship type	n	%
visitsFlowersOf	23,080	17.66	hasDispersalVector	897	0.69
pathogenOf	19,790	15.14	ectoparasiteOf	686	0.52
eats	18,069	13.83	visits	544	0.42
preysOn	17,621	13.48	createsHabitatFor	455	0.35
parasiteOf	10,585	8.10	coOccursWith	315	0.24
epiphyteOf	9,123	6.98	kills	218	0.17
ecologicallyRelatedTo	7,578	5.80	parasitoidOf	92	0.07
hasHost	5,485	4.20	hyperparasiteOf	59	0.05
mutualistOf	5,422	4.15	adjacentTo	47	0.04
pollinates	4,346	3.33	endoparasitoidOf	11	<0.01
endoparasiteOf	1,895	1.45	commensalistOf	5	<0.01
interactsWith	1,293	0.99	hasHabitat	4	<0.01
coRoostsWith	1,214	0.93	ectoParasitoid	3	<0.01
hasVector	930	0.71	livesOn	2	<0.01
symbiontOf	916	0.70	livesInsideOf	1	<0.001

Table 10: Average precision and quality over top-k rules.

Top-k	Precision	Quality
5	0.134	0.605
10	0.132	0.572
20	0.136	0.535
50	0.123	0.532
100	0.117	0.522

Table 11: Evaluation of results for link prediction.

	Head	Tail	All
hits@1	0.103	0.010	0.101
hits@3	0.153	0.140	0.147
hits@10	0.220	0.196	0.208
MRR	0.136	0.127	0.132

One tail query and one head query are shown to present the nature of these predictions in table 12. Table 12a shows the predictions ranked on their confidence for head query *pollinates*(?, *GBIF:7605771*). Table 12b shows the predictions for tail query *parasitoidOf*(*GBIF:1272291*, ?). The prediction with the highest confidence in table 12a is *pollinates*(*GBIF:1341976*, *GBIF:7605771*). The verifications for the top prediction in 12a are shown in table 13. This table shows five rules that are used for this top prediction.

Table 12: Two examples of predictions for queries and their confidence (c).

(a) Top 10 Predictions for head query <i>pollinates</i> (?, <i>GBIF:7605771</i> ).			(b) Top 10 Predictions for tail query <i>parasitoidOf</i> ( <i>GBIF:1272291</i> , ?).		
?	<i>pollinates</i> → <i>GBIF:7605771</i>	c	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → ?	c
<i>GBIF:1341976</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.97	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:1174839</i>	.64
<i>GBIF:1340298</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:9589642</i>	.46
<i>GBIF:5135513</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:1690553</i>	.46
<i>GBIF:5766556</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:1362399</i>	.37
<i>GBIF:1535529</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:1257950</i>	.37
<i>GBIF:4452928</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:7546024</i>	.37
<i>GBIF:11196431</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:5170999</i>	.37
<i>GBIF:1340513</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:5170802</i>	.37
<i>GBIF:4429760</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:5172698</i>	.37
<i>GBIF:1929659</i>	<i>pollinates</i> → <i>GBIF:7605771</i>	.33	<i>GBIF:1272291</i>	<i>parasitoidOf</i> → <i>GBIF:5169894</i>	.37

Table 13: Verifications for head query *pollinates*(?, *GBIF:7605771*).

<i>pollinates</i> ( <i>GBIF:1341976</i> , <i>GBIF:7605771</i> )	
Rule	Conf
BAR, <i>pollinates</i> ( <i>GBIF:1341976</i> , Y) ← <i>pollinates</i> (V1, Y), <i>eats</i> (V1, <i>GBIF:3021496</i> )	0.969
BAR, <i>pollinates</i> ( <i>GBIF:1341976</i> , Y) ← <i>pollinates</i> (V1, Y), <i>eats</i> (V1, <i>GBIF:2988188</i> )	0.961
BAR, <i>pollinates</i> ( <i>GBIF:1341976</i> , Y) ← <i>pollinates</i> (V1, Y), <i>eats</i> (V1, <i>GBIF:8017574</i> )	0.958
BAR, <i>pollinates</i> ( <i>GBIF:1341976</i> , Y) ← <i>pollinates</i> (V1, Y), <i>eats</i> (V1, <i>GBIF:3034827</i> )	0.956
BAR, <i>pollinates</i> ( <i>GBIF:1341976</i> , Y) ← <i>pollinates</i> (V1, Y), <i>eats</i> (V1, <i>GBIF:5375920</i> )	0.930

#### 4.4.2 Qualitative analysis

##### *Purpose of the study*

A survey was made with the purpose of evaluating the *interestingness* of the mined rules by domain experts. To evaluate the *interestingness* of the rules,

three metrics were used; plausibility, relevancy, and newness. Plausibility questions whether a rule can be true in the context of the data. Relevancy questions whether a rule could be relevant and support research. Newness questions whether a rule is new/unknown.

### Rule Selection

From all HAR, CAR and BAR rules, a number of rules were selected to use for the domain expert survey. Of the CAR rules, five were selected, one of len = 1, two of len = 2, and three of len = 3. The selected CAR rules are shown in table 14. Of the HAR and BAR rules, three were selected, two of len = 1 and one of len = 2. The selected BAR and HAR rules are shown in table 15 and 16. The choice was made not to select the rules randomly but to select the rules by use of selection criteria. This was done in order to have a rule set that is representative of the entire rule set. These criteria were:

1. Per type of rule, the goal was that the selected rules have different relationship types in the head atom.
2. The body atoms should also contain different relationship styles so the rules presented are not too similar to each other.
3. Relationship types with high occurrences are present in the rules selected (e.g. *eats*, *visitsFlowersOf*, *preysOn* *hasHost*) as well as relationship types with a lower occurrence (e.g. *laysEggsOn*, *parasitoidOf*).
4. At least one rule should be included containing as relationship type *isOfGenus*, *isOfFamily*, and *isOfOrder*. Only BAR rules produced rules with taxonomic information. This is rule b2 in table 15.
5. For all rule types, all different lengths of rules should be included. The amount of rules per rule length are based on the relative frequencies shown in table 7.
6. A last selection requirement is that all rules should be either high quality of extremely high quality rules.

Table 14: The selection of CAR rules with their confidence (conf) and support (supp).

id	CAR rule	conf	supp
c1	$\text{eats}(X,Y) \leftarrow \text{hasDispersalVector}(Y,X)$	0.71	1163
c2	$\text{parasitoidOf}(X,Y) \leftarrow \text{parasitoidOf}(X,V1), \text{parasitoidOf}(V1,Y)$	0.56	53
c3	$\text{epiphyteOf}(X,Y) \leftarrow \text{epiphyteOf}(X,V1), \text{laysEggsOn}(V2,V1), \text{eats}(V2,Y)$	0.50	72
c4	$\text{parasitoidOf}(X,Y) \leftarrow \text{parasitoidOf}(X,V1), \text{preysOn}(V2,V1), \text{parasitoidOf}(V2,Y)$	0.65	50
c5	$\text{eats}(X,Y) \leftarrow \text{hasDispersalVector}(V1,X), \text{hasDispersalVector}(V1,V2), \text{mutual-istOf}(V2,Y)$	0.54	2220

### Participants

*Profile.* In total eight participants filled in the survey. Participants are all experts in the field of biology or ecology. As all species available in the GloBI data were

included, researchers and specialists from all domains could fill in the survey. Fields in which the experts were occupied are marine biology (n=1), plant-frugivore interactions (n=1), plant insect interactions (n=1), plant-pollinator biogeography (n=1), insects (n=2), insect-plant interactions (n=1) general ecology and entomology (n=2).

### Design

The survey was made with 'Qualtrics Online Survey Software'. At the start of the survey, a short introduction on the content was given along with the consent agreement request. Then two personal questions were asked to determine the area and level of expertise. The first personal question was an open question asking the participants expertise within the field of Biology/Ecology. The second asked the level of expertise in this field, both with a likert scale of 1-5. After the introduction, consent, and personal information, the survey consists of 11 question blocks with introductions in between. In every question block, a rule was given with a graphical representation of the rule like shown in figure 10. The rule was translated into natural language to be more understandable.

Table 15: The selection of BAR rules with their confidence (conf) and support (supp).

id	BAR rule	conf	supp
b1	visitsFlowersOf(GBIF:1544431,Y) $\leftarrow$ visitsFlowersOf(GBIF:1533312,Y)	0.84	67
b2	eats(GBIF:2498089,Y) $\leftarrow$ isOfFamily(Y,GBIF:6750)	0.69	35
b3	pathogenOf(GBIF:3228247,Y) $\leftarrow$ pathogenOf(V1,Y), hasHost(V1,GBIF:2492081)	0.94	116

Table 16: The selection of HAR rules with their confidence (conf) and support (supp).

id	HAR rule	conf	supp
h1	epiphyteOf(X,GBIF:7591641) $\leftarrow$ epiphyteOf(X,V1)	0.69	235
h2	hasHost(GBIF:4549759,Y) $\leftarrow$ epiphyteOf(V1,Y)	0.43	10
h3	epiphyteOf(X,GBIF:7591641) $\leftarrow$ epiphyteOf(X,V1), preysOn(V2,V1)	0.70	235

The questions were the same for all rules and are shown in figure 11. Per rule, the level of expertise on the subject of the rule is asked and could be answered on a scale from 1-5. Then three statements were given on the rule regarding plausibility, relevancy and newness.

1. This rule is plausible.
2. Biological/Ecological research can benefit from rules like this.
3. This rule was unknown to me.

For each of these statements, a five-point Likert scale was offered as answer template ranging from strongly disagree to strongly agree. A higher score indicates a higher *interestingness*. Additionally, if desired, participants could enter general remarks about the rule.



Fig. 11: The questions to answer for every rule.

1.1. My expertise on the subject of this rule:

1	2	3	4	5
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

1.2. This rule is plausible.

Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

1.3. Biological/Ecological research can benefit from rules like this.

Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

1.4. This rule was unknown to me.

Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

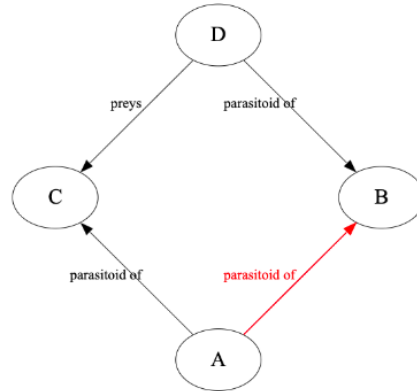
1.5. General remarks about this rule:

Before every set of questions, an introduction was given about the type of rule that was going to be shown and an example of how a rule can be formed. This introduction is shown in figure 12. At first, the BAR and HAR rules were considered a good type of rule to start the questionnaire with. As the rules refer to specific animals they seemed better interpretable than the other rule types referring to variables. Due to comments of participants being unfamiliar with the specific species of the BAR and HAR rules, the order of questions was switched to prevent early termination of the survey. This was done when the survey was already out. Only people that did not yet open the survey thus started the questionnaire with a different order of questions. The choice was made to start the questionnaire with CAR rules. As these rule are abstract and can involve all kingdoms and types of species, they seemed more broadly interpretable for participants of all different expertises.

Initially the choice was made to include five rules for all rule types. Due to comments regarding the time it took to fill in the the survey and many early

Fig. 10: A rule in the questionnaire

**IF** A is a parasitoid of C,  
**AND** D preys on C,  
**AND** D is a parasitoid of B,  
↓  
**THEN** A is parasitoid B.  
Confidence = 0.65




terminations after the BAR rules, the choice was made to delete two BAR and two HAR rules. Finally, the survey consisted of questionblocks one to five on CAR rules, questionblocks six to eight on HAR rules and questionblocks nine to eleven on BAR rules. At the end of the survey, the participants were thanked for their time and contribution.

### Procedure

The link for the survey was distributed by use of email. Due to collaboration with Naturalis, we could distribute the survey more easily under their researchers. Participants who wanted to fill in the survey could access it with an anonymous link. The participants did not need anything other than a computer or mobile phone to complete the survey. The introductory page explained the details about the survey, and reported that filling in the survey lasted about 20 minutes. The link was available from 12 december to 25 january. After that, the link was closed and the results were collected and analysed.

Fig. 12: An introduction for the rules was given.



VRUJE

UNIVERSITEIT

AMSTERDAM

**Introduction.** An example: If we know that the bee is of genus Apoidea, and species of the Apoidea pollinate the buttercup flower. Then we can infer a rule: a bee pollinates the buttercup flower with argumentation: the species is of genus Apoidea, and species of Apoidea pollinate the buttercup flower.

If we make it a more general rule: a species pollinates a plant, if the species is of a specific genus and that specific genus pollinates that plant. This can be written as: A pollinates B, if A is of genus C and C pollinates B.

Rules can be abstract (only variables A-D are used to refer to species) or have specific species to which they apply.

To be able to get a clearer image of the rules, all rules will be presented with a complementary graphical illustration.

So a rule consists of:

- An inferred interaction (colored red in the graph).
- An argumentation for this inferred interaction (colored black in the graph).
- Letter variables (A-D) that indicate species from the species interaction network.

The confidence is given per rule. Rules with a confidence between 0.1 and 0.7 are 'high quality rules' and rules with a confidence between 0.7 and 1.0 are 'extremely high quality rules'.

---

Question 1 to 5 will consist of abstract rules.

### Analyse

In total, eight people participated in our survey. Two of those eight participants filled out the entire survey, the rest filled out only parts of the survey. This resulted in a total of four complete responses and one partial response (3/5 rules) for the CAR rules, three complete and one partial response (1/3 rules) for the HAR rules, and five complete responses for the BAR rules. The difference in which parts were filled out is due to firstly, changing the order of the questions when some participants already filled in and secondly, premature endings of the survey.

Mean scores are calculated for all rule types and metrics to indicate performances. The Krippendorff's alpha ( $\alpha_K$ ) is used for inter-rater reliability. This statistic is used because it can handle various numbers of raters and incomplete data. Also, Krippendorff's alpha can be used on ordinal data.

### Results

The mean scores for plausibility, relevancy and novelty scores for all rule types are shown in table 17. The mean for all metrics on all rules types is 3.06. Out of the three metrics novelty scores highest with a mean of 3.71. In terms of the different rule types, scores for rules of type BAR score highest with a mean of 3.71.

A statistical test (Kruskal-Wallis) shows a significant difference between the mean scores of the different rule types ( $p < 0.05$ ). Using Dunn's test to see which pairs of rule types are different, no significant differences are found between rule types. The biggest difference is found between HAR and BAR rules ( $p = 0.07$ ). The same statistical tests for the difference between scores of the different metrics, shows that novelty scores are significantly higher than relevancy, ( $p < 0.0001$ ) and plausibility scores ( $p < 0.001$ ).

Table 17: Per rule type, the mean score and standard deviation for plausability, relevancy and novelty metrics.

Type rule	Plausability	Relevancy	Novelty	All
CAR	3.52 (1.16)	2.82 (1.11)	3.52 (1.27)	2.96 (1.24)
HAR	2.3 (0.95)	2.3 (1.06)	3.9 (1.37)	2.83 (1.34)
BAR	3.07 (1.16)	3.2 (1.15)	3.87 (0.92)	<b>3.38</b> (1.13)
All	2.65 (1.14)	2.83 (1.14)	<b>3.71</b> (1.18)	3.06 (1.24)

Table 18: Krippendorff alpha for all rule types and metrics.

		$\alpha_K$
Rule type	CAR	0.004
	HAR	<b>0.306</b>
	BAR	-0.022
Metric	Plausability	<b>0.011</b>
	Relevancy	-0.011
	Novelty	-0.172
	Overall	0.105

The inter-rater agreement per rule type and metric is shown in table 18. The overall inter-rater agreement is  $\alpha_K = 0.105$ . This score indicates that there is a slight agreement between the participants on the questions of the survey. Table 18 also shows negative  $\alpha_K$  numbers. A negative Krippendorff's Alpha is an indication of a complex coding process and the need for further investigation and improvement. These suggest that the observed agreement between annotators is poor and less

than would be expected by chance. Of the rule types, the highest agreement is found for HAR rules with  $\alpha_K = 0.306$ , indicating fair agreement. Of the metrics, the highest agreement is found for plausibility with  $\alpha_K = 0.011$ , indicating poor performance.

The general remarks left by some of the respondents can give us insight on the results of the survey. Regarding understanding the rules and the knowledge on the subject of the rules, remarks were made on the uncertainty how to interpret

a rule and unfamiliarity of certain aspects of a rule. This unfamiliarity regarded either the species or the relationship type.

In regard of the judgement on the quality of the rules, remarks were made that some rules generalised too quickly or were based on incomplete data. A participant made a remark that a rule could not be true because the regions the species occur in do not overlap and another could not be true, giving a falsification and proposing a correct rule. A remark was also made about the overlapping semantics of two relationship types.

These remarks demonstrate that the participants had different thoughts and concerns about the quality of the rules and that some participants did not seem to entirely understand the meaning of each of the rules.

## 5 Conclusions

The current research investigated the use of rule mining and link prediction on a large species interaction network. The mined rules were used to describe general patterns or motifs. These rules can help us gain a deeper understanding of the relationships between species and reveal the underlying structure of species interaction networks. We explored the potential of using the tool of link prediction to predict and understand how invasive species behavior. Link prediction can help us identify the links or pathways through which invasive species are most likely to spread and which links are most likely to form between invasive species and native species in a network. Both quantitative and qualitative research was conducted.

Rule mining produced a total of 2,536,397 rules. After doing an overfitting analyses, 130,52 rules were left. This leaves us with only 5% of the initial rule set. This shows that initially, almost all rules overfitted the training data. Following the method used in [41] and [48], 97.7% of the remaining rules are of high quality and 0.6% of extremely high quality. Results show that the rules do well at explaining the interactions in the training data, with an average quality of 0.61 and 0.52 over the top-5 and top-100 patterns, respectively. The rules have low generalizability, with an average precision of 0.13 and 0.12 over the top-5 and top-100 patterns, respectively. This indicates that the rules are overfitting to the training data.

The domain experts survey show that on average, the respondents judge the interestingness of the rule with an overall score of 3.06 out of 5. The average score for plausability is 2.65 out of 5, relevancy scores 2.83 out of 5 and novelty scores 3.71 out of 5. Higher scores indicates a higher *interestingness*. All the three metrics score higher than the midpoint of the scale (2.5). This means that on average, the domain-experts had a above-average judgement of the *interestingness* of the rules. The inter-rater agreement with an  $\alpha_K$  of 0.105 indicates a poor agreement between the participants of the survey. This is likely due to the difference expertises the respondents are employed at and a low response rate.

Results for link prediction show that link prediction can be a transparent and interpretable way of predicting invasive species behavior. In 10% of the cases, the

first prediction of our model is the correct answer to the query, in 15% of the cases the correct answer is in the top 3 predictions and in 21% of the cases the correct answer is in the top 10 predictions.

Rule mining and link prediction are promising tools for understanding species interaction networks and to understand and predict invasive species behavior. Further research is needed to enhance explanatory and predictive performance.

## 6 Discussion and Future Work

As the largest proportion of time spent on this research was on retrieving and preprocessing to get the right data due to, many questions worth asking remain unanswered. Furthermore, drawing any conclusions on the performance of our model should be done with caution due to the different steps and caveats of this research.

The main obstacle that was faced during this research was the substantial loss of data. This loss was largely due to ontology matching to get GBIF identifiers for all taxa, and deleting interactions with a rank higher than species. Of the total 15,128,089 interaction records available on GloBI, covering 875,36 taxa, only 986,219 interaction records covering 170,541 taxa remained. To obtain insight into general patterns and processes between species, we want to be representative of all interactions and species that occur. This lack of coverage can explain the low rule precision of the rules produced. As the results for link prediction are based on the mined rules, having a better species coverage would improve both the quality and precision of the mined rules and the link prediction results.

Additionally, although GloBI covers many species, the sampling density across space, time, and taxonomic ranks is highly variable. Therefore, results found only apply to the species interaction data that is found in specific areas of the world between the species that occur in these areas. As the current research uses data from all taxonomic levels, many observations are needed to produce generalizable results.

A way to obtain a better coverage of the species interaction data is to find ways of retrieving unique identifiers for all taxa that result in deleting a smaller proportion of data than it did in the current research. Additionally, investigating the data coverage per taxonomic level of different available datasets could help determine what scope to use for modelling.

To evaluate the performance of rule mining, taxa from all taxonomic levels are considered. It is possible that the rules for certain taxonomic levels (i.e. a specific kingdom, phylum, order, or family) produced better than the rules for other levels. Analyses could be done for different taxonomic levels in order to get a clearer image on the performance of rule mining and link prediction on different taxonomic levels.

To retrieve one specific identifier for all taxa, the decision was made to use the GBIF ontology. This has implications for the characteristics of the remaining data. By choosing one ontology, interactions where at least one species does not have an available identifier in the chosen ontology got deleted. This means that

the remaining data is highly dependent on the ontology used. All ontologies, have different taxonomic, temporal and spatial coverage of species data. The choices of choosing one ontology over the other should be taken into account when analysing and interpreting the results.

To get insight in the different ontologies, an analysis on the choice of ontology can be done. Different results for rule mining and link prediction will likely be found when matching species identifiers and names with different ontologies. Doing such an analyse could give an insight on how different ontologies perform on different taxonomic levels.

The GBIF parser to match names with, was inconsistent in matching the names. This was probably due parsing failures. This issue was not investigated in the current research but should be noted. When matching the data, the amount of rows in the dataset had a different amount of interaction rows per run. Between some runs the difference was 400.000 interaction rows. This is a large difference and means that not all data available is obtained when not performing accurately.

Regarding the evaluation of the performance for both rule mining and link prediction, three types of rules were considered, BAR, HAR, and CAR rules. The rules produced for our model consist almost entirely of BAR rules (99.88%). So when we look at the top-k performance and quality of the rules, approximately 99.88% are rules of type BAR. The numbers therefore do not say much about the performance of HAR and CAR rules. This can be combatted by changing the configuration settings for GPFL so a more even distribution of the rule types can be found.

Another thing worth noting is the semantics of the names for the interaction types. The data used in the current research contains 37 types of biological interactions. Some of these are more specific (e.g. eats, has host, and preys on) and some are more broadly interpretable (e.g. ecologically related to, visits, interactions with, and co-occurs with). Interactions between species such as *interacts With* or *co occurs With* are used to document interaction claims that can be either intentionally or unintentionally ambiguous. These interactions are used uniformly throughout the analyses but do not always have a uniform meaning. Species found interacting with the same interaction type, do not need to have the same interaction in real life. So, the interpretation of an interaction claim can vary depending who documents the interaction and who interprets the interaction.

The results for link prediction show that link prediction can be a transparent and interpretable way of predicting and understanding invasive species behavior. An interesting application would be that if the species that exist in a certain area (e.g. country, city, geographic area) are known, we can use this information in our model to answer questions concerning the consequences of invasive species in this area. Questions such as: What happens when we put a specific type of bird species in the Northern part of the Netherlands?. We can use link prediction to answer (tail) queries such as: what species will the bird species eat, kill or prey on? Or to answer (head) queries such as: by what species will this bird be eaten? The rules that GPFL produced will only contain the species that exist

in this area, and therefore solves the problem that species can be predicted to interact while not living in the same area.

Another way of dealing with the problem that species that do not occur in the same area can not be found interacting is to add location data. The choice can be to consider biological area's, longitude/latitude or country. This depends on the type of research and the data available. The location data can serve as a restriction that interactions can only be predicted between species that occur together in the same area.

Despite our greatest efforts, a relatively low number of participants of the survey is found. This low number of participants also almost all prematurely ended the survey, resulting in low response rates. This should be taken into account when looking at the results of the survey. The overall judgement of interestingness of domain experts on the different rules is cautiously positive with a score of 3.06 with novelty scoring highest. A score for novelty that is higher than that of plausibility and relevancy implies that when the experts did not judge the rules plausible or relevant or when they did not understand aspects of the rules, they judged them as novel. This results in a high novelty score. A poor inter-rater agreement was found. It is likely that the large differences between raters is caused by the different expertises, familiarities and experiences of the domain experts. As the field of biology and ecology stretches out to many different expertises, the large variability in the answers of the domain experts is not surprising.

## 7 Acknowledgments

I would like to acknowledge and thank to my supervisor Lise Stork for helping me write this thesis by her guiding role. I would also like to thank Stefan Slobach for taking the time to attend my presentation, providing me with valuable comments and insight, and for being the second reader of this thesis. I would also like to thank Vincent Merckx for collaborating with me and by giving his advice at different stages of the research. Finally, I would like to give special thanks to Jorrit Poelen who guided me through GloBI and the enormous amounts of data.

## References

1. Darwin, C. *The origin of species* (Routledge London, 1859).
2. Delmas, E. *et al.* Analysing ecological networks of species interactions. *Biological Reviews* **94**, 16–36 (2019).
3. Christin, S., Hervet, É. & Lecomte, N. Applications for deep learning in ecology. *Methods in Ecology and Evolution* **10**, 1632–1644 (2019).
4. Bohan, D. A. *et al.* Next-generation global biomonitoring: large-scale, automated reconstruction of ecological networks. *Trends in Ecology & Evolution* **32**, 477–487 (2017).

5. Borrett, S. R., Moody, J. & Edelman, A. The rise of network ecology: maps of the topic diversity and scientific collaboration. *Ecological Modelling* **293**, 111–127 (2014).
6. Proulx, S. R., Promislow, D. E. & Phillips, P. C. Network thinking in ecology and evolution. *Trends in ecology & evolution* **20**, 345–353 (2005).
7. Poisot, T., Stouffer, D. B. & Kéfi, S. Describe, understand and predict. *Functional Ecology* **30**, 1878–1882 (2016).
8. Albrecht *et al.* Plant and animal functional diversity drive mutualistic network assembly across an elevational gradient. *Nature communications* **9**, 1–10 (2018).
9. Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C. & Dieckmann, U. in *Systems analysis approach for complex global challenges* 209–248 (Springer, 2018).
10. Cassey, P. *et al.* *Invasion biology: searching for predictions and prevention, and avoiding lost causes* (CAB International Wallingford, 2018).
11. Early, R. *et al.* Global threats from invasive alien species in the twenty-first century and national response capacities. *Nature communications* **7**, 1–9 (2016).
12. Hortal, J. *et al.* Seven shortfalls that beset large-scale knowledge of biodiversity. *Annual Review of Ecology, Evolution, and Systematics* **46**, 523–549 (2015).
13. Poisot, T. *et al.* Global knowledge gaps in species interaction networks data. *Journal of Biogeography* **48**, 1552–1563 (2021).
14. Poisot, T., Stouffer, D. B. & Gravel, D. Beyond species: why ecological interaction networks vary through space and time. *Oikos* **124**, 243–251 (2015).
15. Strydom, T. *et al.* A roadmap towards predicting species interaction networks (across space and time). *Philosophical Transactions of the Royal Society B* **376**, 20210063 (2021).
16. Gu, Y., Guan, Y. & Missier, P. Towards learning instantiated logical rules from knowledge graphs. *arXiv preprint arXiv:2003.06071* (2020).
17. Poelen, J. H., Simons, J. D. & Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological Informatics* **24**, 148–159 (2014).
18. Morales-Castilla, I., Matias, M. G., Gravel, D. & Araújo, M. B. Inferring biotic interactions from proxies. *Trends in ecology & evolution* **30**, 347–356 (2015).
19. Smith, B. *et al.* Relations in biomedical ontologies. *Genome biology* **6**, 1–15 (2005).
20. Golubski, A. J. & Abrams, P. A. Modifying modifiers: what happens when interspecific interactions interact? *Journal of Animal Ecology* **80**, 1097–1108 (2011).
21. Bellard, C., Cassey, P. & Blackburn, T. M. Alien species as a driver of recent extinctions. *Biology letters* **12**, 20150623 (2016).



22. Milo, R. *et al.* Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
23. Bascompte, J. & Melián, C. J. Simple trophic modules for complex food webs. *Ecology* **86**, 2868–2873 (2005).
24. Melián, C. J., Bascompte, J., Jordano, P. & Krivan, V. Diversity in a complex ecological network with two interaction types. *Oikos* **118**, 122–130 (2009).
25. Cavender-Bares, J., Kozak, K. H., Fine, P. V. & Kembel, S. W. The merging of community ecology and phylogenetic biology. *Ecology letters* **12**, 693–715 (2009).
26. Mouquet, N. *et al.* Ecophylogenetics: advances and perspectives. *Biological reviews* **87**, 769–785 (2012).
27. Elmasri, M., Farrell, M. J., Davies, T. J. & Stephens, D. A. A hierarchical Bayesian model for predicting ecological interactions using scaled evolutionary relationships. *The Annals of Applied Statistics* **14**, 221–240 (2020).
28. GBIF. *GBIF Occurrence Download* 2023. [doi.org/10.15468/dl.kxn79b](https://doi.org/10.15468/dl.kxn79b).
29. Dale, M. & Fortin, M.-J. From graphs to spatial graphs. *Annual Review of Ecology, Evolution, and Systematics*, 21–38 (2010).
30. Meilicke, C., Chekol, M. W., Fink, M. & Stuckenschmidt, H. Reinforced anytime bottom up rule learning for knowledge graph completion. *arXiv preprint arXiv:2004.04412* (2020).
31. Craft, M. E. Infectious disease transmission and contact networks in wildlife and livestock. *Philosophical Transactions of the Royal Society B: Biological Sciences* **370**, 20140107 (2015).
32. Zhao, N. *et al.* Machine learning and dengue forecasting: Comparing random forests and artificial neural networks for predicting dengue burden at national and sub-national scales in Colombia. *PLoS neglected tropical diseases* **14**, e0008056 (2020).
33. Bartomeus, I. *et al.* A common framework for identifying linkage rules across different types of interactions. *Functional Ecology* **30**, 1894–1903 (2016).
34. Gravel, D., Poisot, T., Albouy, C., Velez, L. & Mouillot, D. Inferring food web structure from predator–prey body size relationships. *Methods in Ecology and Evolution* **4**, 1083–1090 (2013).
35. Vázquez, D. P., Morris, W. F. & Jordano, P. Interaction frequency as a surrogate for the total effect of animal mutualists on plants. *Ecology letters* **8**, 1088–1094 (2005).
36. Zhang, Y., Dai, H., Kozareva, Z., Smola, A. J. & Song, L. *Variational reasoning for question answering with knowledge graph* in *Thirty-second AAAI conference on artificial intelligence* (2018).
37. Gad-Elrab, M. H., Stepanova, D., Urbani, J. & Weikum, G. *Exfakt: A framework for explaining facts over knowledge graphs and text* in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (2019), 87–95.

38. Chen, Y., Wang, D. Z. & Goldberg, S. ScaLeKB: scalable learning and inference over large knowledge bases. *The VLDB Journal* **25**, 893–918 (2016).
39. Zeng, Q., Patel, J. M. & Page, D. Quickfoil: Scalable inductive logic programming. *Proceedings of the VLDB Endowment* **8**, 197–208 (2014).
40. Ho, V. T., Stepanova, D., Gad-Elrab, M. H., Kharlamov, E. & Weikum, G. *Rule learning from knowledge graphs guided by embedding models* in *International Semantic Web Conference* (2018), 72–90.
41. Galárraga, L., Teflioudi, C., Hose, K. & Suchanek, F. M. Fast rule mining in ontological knowledge bases with AMIE ++. *The VLDB Journal* **24**, 707–730 (2015).
42. Wilcke, W., de Boer, V., de Kleijn, M., van Harmelen, F. & Scholten, H. J. User-centric pattern mining on knowledge graphs: An archaeological case study. *Journal of Web Semantics* **59**, 100486 (2019).
43. *Global Biotic Interactions: Interpreted Data Products (0.4) [interactions.csv]* <https://doi.org/10.5281/zenodo.6604060>. "(accessed: 02.11.2022)".
44. Kennedy, J. B., Kukla, R. & Paterson, T. *Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration* in *International Workshop on Data Integration in the Life Sciences* (2005), 80–95.
45. Bodenreider, O. & Stevens, R. Bio-ontologies: current trends and future directions. *Briefings in bioinformatics* **7**, 256–274 (2006).
46. Lambrix, P. & Ivanova, V. A unified approach for debugging is-a structure and mappings in networked taxonomies. *Journal of Biomedical Semantics* **4**, 1–19 (2013).
47. José Augusto Salim, J. P. *nomer: 0.4.8* <https://doi.org/10.5281/zenodo.7458675>. "(accessed: 2.11.2022)".
48. Omran, P. G., Wang, K. & Wang, Z. *Scalable Rule Learning via Learning Representation*. in *IJCAI* (2018), 2149–2155.

## Appendix 1.A Commandline for retrieving the dataset

```
curl https://zenodo.org/record/6604060/files/interactions.csv.gz
> interactions.csv.gz
```

## Appendix 1.B Commandline for matching and retrieval of interaction data

```
cat interactions.csv.gz |
| gunzip
| mlr --icsv --otsv-lite cut -f sourceTaxonId,sourceTaxonName,
sourceTaxonRank,interactionTypeName,targetTaxonId,targetTaxonName,
targetTaxonRank
| nomer replace --properties replace-source.properties gbif
| nomer replace --properties replace-target.properties gbif
```

```
| grep "GBIF:.*GBIF:.*"  
> GBIF-interactions.tsv
```

### Appendix 1.C Commandline for retrieving taxonomic information

```
cat interactions.csv.gz  
| gunzip  
| mlr --icsv --otsvlite cut -f sourceTaxonId,sourceTaxonName  
| nomer replace --properties replace.properties gbif  
| nomer append --properties genus.properties gbif  
> GBIF-SourceGenus.tsv
```