In a PICKLE: A gold standard entity and relation corpus for the molecular plant sciences

Lotreck, Serena[1,2,*], Segura Abá, Kenia[3.6], Lehti-Shiu, Melissa[1], Seeger, Abigail[1,4+], Brown, Brianna N.I.[1], Ranaweera, Thilanka[1,6], Schumacher, Ally[1], Ghassemi, Mohammad[5], Shiu, Shin-Han[1,2,3,6.*]

[1] Department of Plant Biology, Michigan State University, East Lansing, Michigan, 48824

[2] Department of Computational Mathematics, Science & Engineering, Michigan State University, East Lansing, Michigan, 48824

[3] Program in Genetics and Genome Sciences, Michigan State University, East Lansing, Michigan, 48824

[4] Department of Statistics, University of Michigan, Ann Arbor, Michigan, 48109

[5] Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan, 48824

[6] DOE-Great Lake Bioenergy Research Center, Michigan State University; East Lansing, Michigan, 48824, USA

ORCIDs: 0000-0001-7282-6272, 0000-0003-0329-289X, 0000-0003-1985-2687, 0009-0004-0149-6084, 0000-0002-2623-5583, 0000-0002-2413-1537, 0000-0001-5135-8588, 0000-0001-6470-235X

+ Changed institutions over the course of this work

*Corresponding author:

Email: lotrecks@msu.edu, shius@msu.edu

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

© The Author(s) 2023. Published by Oxford University Press on behalf of the Annals of Botany Company.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Natural language processing (NLP) techniques can enhance our ability to interpret plant science literature. Many state-of-the-art algorithms for NLP tasks require high-quality labeled data in the target domain, in which entities like genes and proteins, as well as the relationships between entities are labeled according to a set of annotation guidelines. While there exist such datasets for other domains, these resources need development in the plant sciences. Here, we present the Plant ScIenCe KnowLedgE Graph (PICKLE) corpus, a collection of 250 plant science abstracts annotated with entities and relations, along with its annotation guidelines. The annotation guidelines were refined by iterative rounds of overlapping annotations, in which inter-annotator agreement was leveraged to improve the guidelines. To demonstrate PICKLE's utility, we evaluated the performance of pretrained models from other domains and trained a new, PICKLE-based model for entity and relation extraction. The PICKLE-trained models exhibit the second-highest in-domain entity performance of all models evaluated, as well as a relation extraction performance that is on par with other models. Additionally, we found that computer science-domain models outperformed models trained on a biomedical corpus (GENIA) in entity extraction, which was unexpected given the intuition that biomedical literature is more similar to PICKLE than computer science. Upon further exploration, we established that the inclusion of new types on which the models were not trained substantially impacts performance. The PICKLE corpus is therefore an important contribution to training resources for entity and relation extraction in the plant sciences.

Keywords

Natural language processing, corpus, named entity recognition, relation extraction, plant science, plant biology

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

1. Introduction

Information overload (where there is too much available information for an individual to effectively process) is a universal problem in the sciences because of the rapid increase in new publications per year. For example, there has been an average 15% increase in new publications per year for the search term “plant science” on the search engine PubMed . In Semantic Scholar (Fricke, 2018), a search for “plant science” in December 2022 returned upwards of 6.5 million results. The information in these articles represents the collective intellectual output from plant science researchers over the past century. However, even the most diligent scientists cannot read all relevant papers; for this reason, a majority of the literature is left unused by a vast majority of the research community. This reality introduces a second-order problem: high-quality research demands hypotheses based on the wealth of existing knowledge found in scientific literature, but current researchers predominantly rely on time- and labor-intensive manual methods to manage, understand, and integrate the knowledge into downstream scientific workflows (Landhuis, 2016).

The field of natural language processing (NLP) focuses on the analysis of human language text, and in recent years has made strides towards robust automated information retrieval from unstructured text, including the scientific literature. Two particularly relevant NLP approaches that have gained traction in the biological sciences are Named Entity Recognition (NER) and Relation Extraction (RE), which are used to extract entities and their relationships, respectively, from text (Nicholson and Greene, 2020); examples of entities in the biological sciences include: genes, proteins, and pathways while examples of relationships in the biological sciences include: activates, produces, and interacts. For instance, NER and RE algorithms can be used to distill complex statements such as “These results demonstrate that JA-induced AabHLH1 positively regulates artemisinin biosynthesis… (PMID: 30719762)” into a semantic triple: (“JA-induced AabHLH1”, “activates”, “artemisinin biosynthesis”). Once entities and relations are obtained, downstream tasks such as knowledge graph construction can be completed. In knowledge graph construction, entities and their relations are formed into a graph, where entities are the nodes of the graph, and relations form the edges between them. The graph can be thought of as a set of semantic triples. The emergent graph describes the knowledge landscape and can be used to infer possible relationships between entities that may not yet be experimentally validated in the literature. Consequently, the graph can be used to identify prospective directions of research by providing novel biological hypotheses based on previous experimental observations (Nicholson and Greene, 2020). For example, if the graph predicted an “activates” relation between two proteins known to be involved in a biological phenomenon, and that relationship was not previously experimentally proven, this would be a novel hypothesis about how those two proteins give rise to the observed phenomenon.

Hypothesis generation from knowledge graphs has been successful in the biomedical sciences, where graph predictions have been used to identify hypotheses about drug-drug interactions and drug-disease relations (Celebi *et al.*, 2019; Karim *et al.*, 2019; Mohamed, Nováček and Nounu, 2019; Bougiatiotis *et al.*, 2020; Dai *et al.*, 2020; Mohamed, Nounu and Nováček, 2020); these automated systems for hypothesis generation can be useful for similar applications in the plant sciences. However, graph generation methods rely on large labeled training datasets (i.e. corpora) with manually marked entities and relations; these data can be used to teach algorithms what text spans (i.e. contiguous segments of text) should be identified as entities and connected by relations. Generating annotated datasets is a labor-intensive process requiring expert annotation

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

following a set of annotation guidelines. In the biomedical domain, an increasing number of high-quality corpora have been developed in the last twenty years. An example is the GENIA corpus, which is labeled with entities and static relations (as opposed to causal relations)[1]. GENIA has 2,000 abstracts labeled with entities, and 1,600 instances of static relations labeled within the same documents (Kim *et al.*, 2003; Pyysalo *et al.*, 2009). Causal relations are important in hypothesis generation, as we are often interested in the mechanisms that give rise to certain observations. Importantly, GENIA does not include annotations of causal relationships, unlike the ChemProt corpus[2], which contains 2,482 abstracts annotated with both entities as well as causal relations like "upregulates" or "downregulates", clustered into overarching relation categories that are used for annotation. Another example is the CRAFT corpus, which includes 97 full-text articles labeled with entities only (Bada *et al.*, 2012). The biomedical domain also benefits from many NER and RE pre-trained models based on corpora like GENIA and ChemProt, which allow models to be directly applied to new data without requiring investment of computational resources and time into training models from scratch (Wadden *et al.*, 2019; Zhong and Chen, 2020).

Currently, there exist multiple pioneering annotation efforts in the plant sciences. Three of the existing plant science corpora contain documents that focus on a single species: OrzyaGP is a dataset of titles and abstracts from papers about rice (*Oryza sativa*), containing only entity annotations that were automatically derived (rather than manually annotated) (Larmande, Do and Wang, 2019); SeeDev focuses on *Arabidopsis thaliana*, with entity, relation, and event annotations (Chaix *et al.*, 2016); and the corpus described in Singh et al. 2021 consists of 34 full-text articles about potato (*Solanum tuberosum L.*), annotated with entities and relations (Singh *et al.*, 2021). Another subset of the existing plant-related corpora focuses on medicinal plants and their relationships to human biology: Choi et al. 2016 annotates medicinal compounds derived from plants (Choi *et al.*, 2016), while Cho et al. 2022 and Kim et al. 2019 focus on the effect that plants have on human disease phenotypes (Kim, Choi and Lee, 2019; Cho *et al.*, 2022). Finally, there are two non-public corpora, annotated with entities but not relations, that focus specifically on the horticultural domain (Liu *et al.*, 2020) or agriculture (S., Lex and Lalitha Devi, 2016). Across the existing efforts for plant biology entity and relation annotation there are four areas that can be further improved, and each of the aforementioned corpora can be improved at least one areas: (1) the documents in the corpus are limited to a specific species (Chaix *et al.*, 2016; Larmande, Do and Wang, 2019; Singh *et al.*, 2021); (2) the thematic focus of the corpus is on medicinal plants and human biology, rather than plant science (Choi *et al.*, 2016; Kim, Choi and Lee, 2019; Cho *et al.*, 2022); (3) there are no relation annotations in the dataset (S., Lex and Lalitha Devi, 2016; Larmande, Do and Wang, 2019; Liu *et al.*, 2020); or (4) the dataset is not publicly available (S., Lex and Lalitha Devi, 2016; Liu *et al.*, 2020).

[1] An example of a static relation is in the statement "The definition of the AHA-like motifs present in AtHB1, AtHB7, AtHB12 and AtHB13 TFs… (PMID: 24531799)"; this statement becomes four semantic triples, of the form ("AHA-like motifs", "is-in", "AtHB1 TFs"), one for each protein. In a static relationship, change is not implied: the AHA-like motif is always part of those proteins, and is not taking any action that affects another entity. In contrast, a causal relation describes entities that take action on or towards another entity, like in the previous example of ("JA-induced AabHLH1", "activates", "artemisinin biosynthesis").

[2] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Our work seeks to further improve upon existing corpus-building efforts in plant science by creating a publicly available, high-quality annotated dataset that includes multiple plant species, as well as annotations based on a set of ontologies that adapt previous efforts to allow the corpus to scale to additional levels of biological organization (e.g. molecular biology as well as ecology), and incorporate a range of biological relationships. We only include molecular biology documents in this initial corpus for two main reasons: (1) we wanted to test our annotation guidelines in a narrower and more controlled setting, without introducing the added complexity of cross-discipline annotations, and (2) our annotators are specialists in molecular biology, not ecology. However, we intentionally chose relation types that would generalize well to other levels of biological organization. We apply the guidelines developed in this study to create the Plant scienCe KnowLedgE graph (PICKLE) corpus for the molecular plant sciences. In addition to providing a generalizable corpus for the plant sciences, we address the question of how well existing models perform in the plant science domain by evaluating pre-trained NLP models for downstream NER and RE tasks, and determine how effective the PICKLE corpus is as a training corpus by training a PICKLE model.

2. Methods

2.1 Initial creation of ontologies and annotation guidelines

An ontology is a structured set of hierarchical classes (which can be used as annotation types for a given corpus) and their interrelationships. In creating our ontologies and guidelines, we sought to combine the most useful elements of several different annotation paradigms and approaches in order to maximize the flexibility and utility of our resulting approach to annotation. An ontology is a structured set of hierarchical relationships between annotation types. PICKLE's annotation guidelines are based on two ontologies: one for entities and one for relations (**Supplemental document 1**, **Supplemental document 2**). The PICKLE entity ontology is derived from the GENIA project term ontology (Kim *et al.*, no date), and consists of three sub-hierarchies: Chemicals, Organisms, and Anatomy. Before writing the annotation guidelines, ontology terms from the original GENIA version that were not applicable to plants were modified. The PICKLE relations ontology is based on the GENIA project static relation (Pyysalo *et al.*, 2009) ontology as well as the BioNLP13 Gene Regulation Network Task relations (Bossy *et al.*, 2015). Initially, all relations from two of the three parts of the GENIA relation ontology were included in the static relations ontology (those under *relation* and *has-variant*) in addition to a new relation, *is-in*, which refers to the physical location of an entity (e.g. "Protein 1 *is-in* the nucleus").

The PICKLE annotation guidelines are in two parts, one for entities and one for relations (both included in **Supplemental document 3**). Each section contains general rules for annotation, as well as examples to illustrate rules that may not be intuitive without an example. The entity annotation guidelines are derived from GENIA's term annotation guideline as described in (Kim *et al.*, 2003), and consist of two sections: (i) the semantic annotation guidelines, which define the terms that should be annotated using the entity ontology and (ii) the syntactic guidelines, which define where to draw the boundaries of labeled entities based on parts-of-speech (i.e. the different grammatical roles that words play in a given sentence). Examples are provided from plant science-specific abstracts to illustrate the implementation of specific aspects of the annotation guidelines, inspired by the examples provided in the CRAFT annotation guidelines (Bada and Eckert, no date).

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

The relation annotation guidelines are a direct corollary of ontologies on which they are based, and consist of a general annotation procedure, as well as examples of specific rules. The general annotation procedure for relations is to look for relationships between entities labeled within a single sentence (relations typically do not cross sentence boundaries for downstream NLP applications). The relationship identified must be explicitly expressed within the sentence. Due to the limitations of downstream algorithms like those used here (DyGIE++ (Wadden *et al.*, 2019), see *Evaluation of entity and relation extraction in the plant sciences*), entities must be continuous (e.g. not have any interrupting words in the middle), and relations can only be annotated within sentences.

2.2 Document selection

At each iteration of the annotation guidelines refinement process described in the next section, a different set of 5 - 10 documents was chosen from PubMed search results. For most of the rounds of annotation, the documents were selected from the combined output of two PubMed search results for the terms “gibberellic acid” and “jasmonic acid”. We chose the terms “gibberellic acid” and “jasmonic acid” for document selection for this initial corpus because (1) they are present nearly ubiquitously across plant species (Ruan *et al.*, 2019; Hedden, 2020), and (2) because they are central to the concept of the growth-defense tradeoff, which is an important topic in plant science (Huot *et al.*, 2014). The text of each resulting abstract was extracted into a separate document; this yielded a total of 8673 documents. However, after the first three rounds of annotation, we hypothesized that we might see an improvement in annotation if we used only abstracts from the top five ranked plant biology journals, as we expected better clarity of writing due to higher quality control. In those rounds, we chose the documents from the PubMed searches for the same two terms, but this time, refined the searches by only including abstracts from the journals *Plant Cell*, *Plant Physiology*, *Nature Plants*, *New Phytologist*, and *Molecular Plant*. Document sets 4 & 5 were drawn from this search.

In both cases, documents were selected from the search results based on the statistical characteristics of the abstracts. More specifically, the Doc2Vec algorithm (Dai, Olah and Le, 2015) was used to generate vector representations of the documents; the document vectors were clustered using mean-shift clustering (Derpanis, 2005), and then sampled sequentially through the clusters (without replacement) until reaching between 5 and 10 documents per annotation set. We found that the Doc2Vec embeddings of abstracts from the two PubMed searches were represented by one main cluster, with only a handful of smaller clusters containing one or two documents each; the overwhelming membership of the main cluster means that documents were being selected randomly from the PubMed search results. The same document selection process was used to choose the remaining 195 documents that were annotated after the refinement process described below, resulting in a total of 250 documents in the corpus.

2.3 Annotation guideline and ontology refinement process

The final annotation guidelines and ontologies were reached through an iterative process of improvement. Iterative improvement allowed us to identify what elements of the ontologies and guidelines were the most useful to us, as well as any elements that might have been confusing or otherwise limit the quality of the final annotations. We performed the iterative refinement process

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

sequentially for entities and relations, first generating a final set of guidelines and ontologies for entities along with a gold-standard entity set, and then using that entity gold standard to perform the iterative process for relation annotation. Annotators were five doctoral students in computational plant biology, one PhD scientist in molecular plant biology, and one data scientist. There were six iterations of annotation and improvement for both entities and relations, using the same sets of documents for both. Annotations for both entities and relations were completed using the brat annotation tool (Stenetorp *et al.*, no date). Instructions for locally installing and using the brat tool were provided to annotators (supplementary file **Supplemental document 4**).

To evaluate the effectiveness of the guidelines at leading to high-quality annotations during each annotation, we calculated an inter-annotator agreement (IAA) score after annotators finished their versions for each round (see Discussion for further commentary on the rationale behind this approach). IAA was calculated using the bratiaa package[3] with scispacy (Neumann *et al.*, 2019) as the tokenizer in order to accurately break up scientific terminology into words. To calculate IAA, we used the F1 score, which is equivalent to the commonly used Cohen's kappa when the metrics are used on natural language tasks where there are a large number of text spans that are never labeled by annotators (Hripcsak, 2005; Boguslav and Cohen, no date). Given the equivalency between Cohen's kappa and F1 in this annotation scenario, we chose to use the F1 metric to calculate an IAA score for double-blind annotations; we refer to this metric as *double-blind IAA*. In double-blind IAA, all annotators are compared against one another in pairs, and the F1 scores of all pairs are averaged to get an overall agreement score for the round. For each round of annotation, we used double-blind IAA as well as qualitative feedback from annotators and observations of specific disagreements to refine the ontologies and guidelines; the updated versions of the ontologies and guidelines that resulted from this quantitative and qualitative feedback were then passed on to the next iteration of annotation.

We began with the entity annotation guidelines. Sets of 5-10 documents were provided to a group of 2-5 annotators, along with the entity annotation guidelines. Based on numbers from the literature (see Discussion for further details), we set a target double-blind IAA of 0.6, and performed the iterative improvement process over unique document sets until that target was reached. Subsequently, the unification process described in the next section was used to create a final gold standard entity set. The same sets of 5-10 abstracts with the final entity annotations were returned to the annotators for relation annotation. We chose to give annotators the gold standard entities while annotating relations so that IAA for relation annotations measured only their ability to identify relations; this prevented confounding accurate relation annotations with being able to correctly identify entities. We performed the same process of refinement based on IAA and annotator feedback for relation annotations. Since to the best of our knowledge there are no publicly-available packages to compute IAA for non-text bound annotations (like relations) in brat, we wrote a script to compute F1 (`relationIAA.py`). This included two approaches: (1) Strict F1 considers the two entities that are linked by the relation, as well as the type and the direction of the relation; in order for two relations to agree, they must connect the same entities, point in the same direction, and have the same type; (2) Loose F1 only checks that the same entities are connected, irrespective of type and direction. By comparing strict and loose IAA scores, we could determine what specifics of

[3] https://github.com/kldtz/bratiaa

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

the relation annotation guidelines needed to change. If the loose agreement was substantially higher than the strict, the problem lay with confusion about types and directions, whereas if they were similar, the problem was knowing which entities should be connected by relations. Rather than setting a target IAA for relations, we decided to complete the iterative process for the sets of documents included in the gold standard entity annotations, as this obviated the need to generate extra entity annotations if the relation IAA didn't reach the target over the course of annotating the existing document sets. Once all 6 iterations were complete for both entities and relations, we performed one more round of overlapping annotation with 50 documents to get a more robust evaluation of the IAA for both entities and relations. Finally, the relation annotations were unified and additional documents were annotated as described below to create the final, gold standard corpus.

2.4 Creation of gold standard corpus

Once the iterative improvement process was completed for entities, annotations from all the annotators were unified. Identical annotations from each annotator were automatically merged, leaving conflicts to be manually resolved. The author of the guidelines reviewed these conflicts and made executive decisions about the correct annotation for each instance, using the principle of greedy annotation (annotating the longest span in cases of doubt, see **Supplemental document 4** for more information) to consistently determine span boundaries. Annotations using ontology types that were eventually removed during the refinement process were also changed to the appropriate type from the final ontologies. All 55 of the initial documents that were annotated with different versions of the entity and relation ontologies were randomized and reviewed again in order to ensure that no bias during unification was introduced. After finalizing entity annotations and returning them to annotators to complete the same refinement process for relations, relation annotations were unified using the same method.

To quantitatively justify a threshold for how many documents should be annotated for the final corpus, we performed an analysis of the impact of training corpus size on model performance for NER and RE tasks using the ChemProt[4] (biomedical), GENIA (biomedical, (Kim *et al.*, 2003)), SciERC (computer science, (Luan *et al.*, 2018)), and BioInfer PPI subset (biomedical, (Pyysalo *et al.*, 2007)) corpora[5]. Specifically, we asked after how many training documents the model performance based on these corpora would stop improving. A test set and development set of 50 documents each were randomly selected from each of the four corpora. We selected a baseline training corpus of ten initial documents, and then iteratively added ten documents and trained and evaluated a model until the number of documents in the training corpus was equal to 500 (or 400 in the case of SciERC, as the entire corpus contains 500 documents). We generated confidence intervals, considering the types of entities and relations when we performed evaluation (see *Evaluation of entity and relation extraction in the plant sciences* below for details). For the performance of the model on the development set, we pulled the development set performance reported during model training (which does not include a confidence interval). It is important to note that the DyGIE++ model fails

[4] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vi/track-5/

[5] https://huggingface.co/datasets/bigbio/bioinfer; at the time of submission, the full BioInfer dataset was not publicly available

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

unpredictably during application when trained on small datasets. While the internal performance on the development set calculated during training follows the expected trend, some models in our analysis result in F1=0 on the test set. Zero performance does not seem to be a result of a given instance in the training set, as once an instance is added to the training set, it appears in all subsequent training sets. However, given that the overall trend is clear, we have moved forward with this analysis.

Beyond the original 55 documents, unification was also undertaken for the 50 overlapping documents from the final assessment of IAA, and as a result of the corpus size analysis, 145 additional documents were annotated singly (each document received one set of annotations from a single annotator) to bring the corpus to a total of 250 documents. After annotation, all 250 documents were reviewed for consistency in entity and relation annotation (without randomization like we have done for the original 55 documents, as the ontologies did not change across document groups). Throughout all parts of the review process, further decisions were made about the specificity of the annotation guidelines. These additions, which include both changes to written rules as well as additional examples, are included in the version of the annotation guidelines (see **Supplemental document 3**) , and the most recent version of the annotations themselves are found in the data distributed with this manuscript (see *Data*). Note that annotation statistics reported in **Table S2** are the number of entity and relation annotations present *before* conversion to the input format required for the downstream algorithms; this conversion resulted in the loss of some entities and relations due to tokenization mismatches between the character-based indices for annotations in brat[6] and the token-based indices for annotations in the DyGIE++ data format[7]. After conversion, there were 6,164 entity annotations and 2,094 relation annotations.

2.5 Evaluation of entity and relation extraction in the plant sciences

To examine the performance of NLP models from other domains in the plant sciences, we used the model architecture DyGIE++ for NER and RE. DyGIE++ (Wadden *et al.*, 2019) is a joint NER and RE algorithm that offers an option to improve predictions by also integrating coreference resolution, where different mentions of the same real world object are mapped back to a single entity (see (Wadden *et al.*, 2019) for architecture details). We contributed code to the DyGIE++ repository[8] to convert brat annotations to the required input format for DyGIE++. DyGIE++ provides off-the-shelf models trained on ACE05 (a general domain corpus) (Walker, 2006), SciERC, ChemProt, and entity-only extraction on GENIA, and uses coreference resolution in all models except for the model for ChemProt. There are also "lightweight" models for GENIA and SciERC which don't use coreference resolution; we include both sets of models in our initial evaluation of model performance. In addition to pretrained models, we trained two DyGIE++ models, one on the publicly available portion of the SeeDev corpus (Chaix *et al.*, 2016) and one on the PICKLE corpus. The exact hyperparameter configurations used for each model, including those used by DyGIE++ to train the pretrained models, are available in our repository (see *Code Availability* for details). All parameter

[6] https://brat.nlplab.org/standoff.html

[7] https://github.com/dwadden/dygiepp/blob/master/doc/data.md

[8] https://github.com/dwadden/dygiepp

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

tuning occurs automatically within the DyGIE++ implementation. We evaluated the models' performance on the test set of the PICKLE corpus, as well as the test set from the dataset on which they were trained (with the exception of ACE05, as the dataset is not publicly available). We used the same train/development/test splits provided by DyGIE++ for SciERC, GENIA, and ChemProt. SeeDev is still maintained as a closed-source test set, so we only had access to the train and development sets, which is a total of 58 documents. We combined the SeeDev train and dev sets into one dataset, and for both SeeDev and PICKLE we performed our own train/development/test splits, using 20% of the dataset as the test set and 12% of the dataset as the development set.

We used an external script to evaluate model performance, as we wanted to (a) bootstrap confidence intervals around performance estimates and (b) be able to evaluate model performance both without consideration to types and considering types[9]. One major weakness of pretrained models is that the model can only predict the same entity and relation types on which it was trained. However, not all downstream applications require types; many only need the specific text identified as entities and the presence or absence of the connections between them. Many commonly used domain-agnostic NER and RE approaches such as Stanford OpenIE (Angeli, Johnson Premkumar and Manning, 2015) are unable to assign types to identified entities and relations. Therefore, the evaluation of model performance while ignoring their predicted types is an important part of our evaluation paradigm. Additionally, we evaluated all models on their original test sets both with and without relations. When considering types in evaluation, a predicted entity is a true positive if the boundaries of the span (which characters are included in the annotation) and the entity type match. For relations, the entity span boundaries must match, as well as the type and directionality of the predicted relation. When types are ignored, only entity span boundaries must match, and for relations, only the span boundaries of both entities must match (type and direction are both ignored). For all performance values, we calculated a bootstrapped 95% confidence interval (CI) using 500 samples, where we sampled the test set with replacement 500 times and calculated performance metrics for each sample, and used the percentile method to determine the bounds of the 95% CI.

2.6 Analysis of the impact of the original training domain on performance in plant science

To further explore discrepancies in performance of NER on the PICKLE test set between pretrained models from different domains, we took a multi-pronged approach. First, we looked at the differences between the models' training dataset vocabulary and the PICKLE corpus. The vocabulary of a model is extremely important in determining model performance, as the model is unable to differentiate words that do not appear in its vocabulary. All else being equal, if two models have more similar vocabularies, they should perform similarly on the same data. We obtained the tokens from each corpus, and used the nltk package[10] to get uni-, bi- and trigrams[11]. For each of

[9] When direction and type were ignored when evaluating relations, one predicted relation could match multiple gold standard relations, if, for example, there are two relations of different types with opposite directions connecting the same two entities. In this case, both of the gold standard relations are counted as correctly identified (true positives).

[10] https://www.nltk.org/

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

those three gram lengths, for each of the SciERC, ChemProt, SeeDev and GENIA corpora, we identified words that were in the PICKLE corpus but not in the other corpus, normalizing by total number of words in the PICKLE corpus, which calculates the fraction of the PICKLE vocabulary that is *not* covered by the comparison corpus. However, for more intuitive visualization, we subtracted the fraction from 1, in order to visualize the fraction of the PICKLE vocabulary that *is* covered by the comparison corpus, as we expect this value to correlate with model performance. The final calculation follows the equation:

$$1 - \frac{number\ of\ n{-}grams\ in\ PICKLE\ that\ are\ not\ in\ comparison\ corpus}{number\ of\ n{-}grams\ in\ PICKLE}\ for\ n\ = 1, 2, 3$$

To further explore the unexpected low performance of the GENIA models, we examined the impact of the addition of new types in the test set, compared to the types that each model was trained to predict. To test the hypothesis that the performance of the models would be negatively impacted by being used to predict entities that didn't correspond to one of the type categories on which they were trained, for each model, we filtered the PICKLE test set to remove any types that the model was not trained to predict, and re-evaluated each model. To perform filtering, we mapped the types from the SeeDev, GENIA, and ChemProt models to the types from the PICKLE dataset. Any PICKLE types that did not map to SeeDev/GENIA/ChemProt types were removed from the test set used to evaluate each model. We excluded the SciERC models from this analysis, as the computer science entity types are too semantically different from the biological types of the other models to be mapped back to the PICKLE dataset. The SeeDev and GENIA datasets result in a many-to-one mapping, where some types are more specific than those in the PICKLE dataset, and therefore map to the same PICKLE type; for example, "cell_type" and "cell_line" in the GENIA dataset both map to "Cell" in the PICKLE dataset. In the SeeDev dataset, there is one type, Environmental factor, that doesn't map to any PICKLE type; however, there are only 2 individual predictions (0.1% of the total predictions) made by SeeDev in this category, so we were not concerned by not having a mapping. In comparison to SeeDev and GENIA, the ChemProt dataset has a much broader annotation schema than PICKLE does, so there is a one-to-many mapping, where each ChemProt type encompasses multiple PICKLE types. In contrast to a many-to-one situation, where it's clear which types correspond to the broader categories of the PICKLE dataset, to get a ground truth mapping in a one-to-many scenario, we would need to perform a full re-annotation of the ChemProt dataset. As a complete re-annotation was not achievable with our resources, we used the ChemProt model's predictions on the full PICKLE test set as a proxy to determine which PICKLE types were being represented by the categories in PICKLE. We looked at true positive predictions from the ChemProt model, and determined that the CHEMICAL type encompassed organic compounds, inorganic compounds, elements, and plant hormones in the PICKLE ontology, and that the GENE type included both DNA and Proteins.

[11] A unigram is a single word, a bigram is two words treated as a single unit, a trigram is three words. For example, "for_example" would be a bigram from this sentence.

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

As SciERC was excluded from the previous analysis, we also looked into the effect that SciERC's original types had on prediction capability. We looked at true positive predictions from the full SciERC model, and quantified the percentage of true positive predictions accounted for by each of SciERC's prediction types. We then took the two largest categories, Method and OtherScientificTerm, and quantified the proportion of PICKLE's types present in each category's predictions by dividing the number of gold standard annotations with a given PICKLE type that were predicted as either Method or OtherScientificTerm by the total number of true positive predictions with the Method or OtherScientificTerm label.

2.7 Code availability

All original code and training configurations for this manuscript can be found in our GitHub repository (https://github.com/ShiuLab/pickle-corpus-code), with the exception of the code contributed to DyGIE++, which can be found in the DyGIE++ repository (https://github.com/dwadden/dygiepp). The previously published model code, as well as the pretrained models themselves, referenced in this manuscript come from the DyGIE++ repository. Instructions for how to use these two repositories together to reproduce the results in this paper is detailed in the README in our repository.

3. Results

3.1 Annotation guidelines and ontology refinement process

In **Figure 1** we show the 21 terminal entity types across three sub-hierarchies (**Figure 1A**), and five relation types across two sub-hierarchies for the final PICKLE ontologies (**Figure 1B**). The entity ontology is split into the three sub-hierarchies Chemicals, Organisms, and Anatomy, each of which contains a set of specific sub-categories. The relations ontology is split into two sub-hierarchies, Static Relations and Causal Relations. These ontologies can be found in their entirety with type definitions in the supplementary files **Supplemental document 1** and **Supplemental document 2**.

The ontologies and annotation guidelines are the final result of a process of iterative improvement that was performed interactively with the annotators (**Figure 2**). We performed this process first for entity annotation, and once a final gold-standard version of the entity annotations was established, repeated the process for relation annotation. After each round of annotations, we calculated the IAA to quantify disagreement between annotators, and incorporated qualitative observations about disagreements and annotator feedback to improve the guidelines. Changes to the ontologies and the guidelines were made to improve annotators' interpretation of the guidelines. An example of changes to the ontologies is that the most detailed level of organization in the Chemicals sub-hierarchy, which consisted of further differentiation of the *DNA*, *RNA,* and *Protein* types, was removed to improve inter-annotator agreement (**Supplemental document 5**, **Figure S1A**). Similarly, all static relations from the GENIA static relations ontology besides *is-in* were removed. We also simplified the BioNLP13 Gene Regulation Network ontology from eight terminal terms to four (**Figure S1B**). The second type of changes, to the guidelines themselves, were made as ambiguous situations cropped up that could be resolved through generalizable changes in the guidelines. These

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

changes included adding general rules, like the policy of greedy annotation, or adding examples clarifying an existing rule.

3.2 Determination of corpus size and annotation quality

An analysis of the impact of corpus size on the performance of downstream models with the ChemProt, GENIA, SciERC, and BioInfer PPI datasets demonstrated that no statistically significant (non-overlapping 95% confidence intervals) increase in performance for entity or relation extraction was obtained beyond a corpus size of 150 documents for all datasets (**Figure 3**). In consideration of this result, as well as looking at the sizes of other similar corpora, we chose to target the annotation of 250 abstracts from journal articles from PubMed searches for "jasmonic acid" and "gibberellic acid"; **Table S1** contains statistics about these abstracts.

Since human language is inherently ambiguous, there is often no one "right" way to annotate a given document, unlike other types of annotation (for example, in annotating pictures of cats and dogs, it's clear whether the right answer is "cat" or "dog"). Instead of relying on comparison with a correct answer to determine whether our annotation guidelines and resulting annotations are high-quality, we rely on the concept of the *reliability* of annotation. This is the idea that, if we gave the same document and annotation guidelines to different people, they would be able to reliably identify most of the same entities and relations. We quantify this reliability by using IAA to compare annotators against one another (Hripcsak, 2005). We set our sights on an entity IAA of 0.6, as a reasonable compromise between reliability, generalizability, and labor.

Figure 4 shows the improvement in IAA over the course of the refinement process for both entities (**Figure 4A**) and relations (**Figure 4B**), as well as a final IAA with a larger document size for the final guidelines (Round 7 in **Figure 4**). Entity annotation improved in a statistically significant manner over the rounds of annotation, while relation annotation significantly improved when the type and direction of the relation were considered (strict tolerance). For the final round of annotation, a double-blind IAA of 0.65 was reached for entities, and for relations an IAA of 0.49 (strict tolerance) and 0.63 (loose tolerance) were reached. After the final round of entity and relation annotation improvement, the final corpus was generated by unifying all the annotators' annotations from all document sets. The final corpus contains 6,245 entity and 2,149 relation annotations across 250 documents (**Tables S1** & **S2**). When we extended the previously discussed analysis of the impact of corpus size on model performance to include the completed PICKLE corpus, we found that the performance in downstream models follows the same trend as the other corpora, with a leveling-out of performance observed at a training corpus size of 150 documents (**Figure 3**).

3.3 Evaluation of entity and relation extraction model performance on the PICKLE corpus

We evaluated how well existing, pre-trained models, and new models (trained on our PICKLE and the SeeDev corpus), perform in the plant sciences (out-of-domain), as well as on their original domains (in-domain). Because all models are based on DYGIE++, they are named after the corpora used to train the models. Analyzing both in- and out-of-domain performances helps us determine viable future directions for NER and RE in the plant sciences, answering the question of how much we can utilize existing resources from other domains.

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

First we examined the performance of pre-trained models based on the DyGIE++ model architecture (see **Methods**) to jointly predict entities and relations from the PICKLE corpus (out-of-domain), as well as on the original test sets from the domain on which the models were trained (in-domain). We first focused on entity performance. The models were evaluated first without consideration of the predicted entity types. While the models' performance on their original domains were relatively similar for entity extraction (**Figure 5B**; **Table S5**), they varied greatly in their performance on the PICKLE corpus (**Figure 5A**; **Table S3**). The ChemProt and the SciERC models performed similarly well (**Figure 5A**; **Table S3**). Surprisingly, the SciERC model, which is based on a corpus specialized to the computer science domain, outperformed most other pre-trained models on PICKLE including the SeeDev model trained with the SeeDev corpus containing plant literature. Model performance does not seem correlated with corpus sizes. While the poor performing SeeDev model has a small training set (38), the GENIA model has more documents in its training set (1,568) than those in the SciERC model (350) by a large margin. We explore this phenomenon further in the next section. We subsequently trained DyGIE++ models on SeeDev and PICKLE, and compared them to the performance of the pre-trained models on the PICKLE test set. The model trained on the PICKLE training dataset provided significantly better performance on the PICKLE test dataset (**Figure 5A**). The performance comparison highlights the low generalizability of pretrained models from other domains and the need for domain-specific models. Consistent with this, the pre-trained models also performed better on its own test set in-domain (non-overlapping 95% confidence intervals, **Figure 5B**).

After evaluating model performance on the NER task, we next focused on RE by evaluating models first without consideration of the predicted entity types. Pre-trained model performance on the relation extraction task (**Figure 5C**) was much lower than the corresponding entity extraction performance for all models on all datasets (**Figure 5A**), regardless of original domain similarity to the PICKLE corpus (**Figure 5B**; **Table S2**). Only the PICKLE model was able to achieve a significantly higher level of performance on relation extraction for its own test set (F1=0.4, non-overlapping 95% confidence intervals), as the performance of all other models on the PICKLE test set hovered below an F1 of 0.1 (**Figure 5B**; **Tables S2** and **S4**). When comparing performances of the non-PICKLE models on the original domain test set for relation extraction, all models performed significantly better except the SeeDev model (non-overlapping 95% confidence intervals, **Figure 5D**; **Tables S5 and S6**). We expected the SeeDev model's performance on relation extraction to be poor, as SeeDev is a relatively small dataset (58 available annotated documents across all three splits) with a large number of relation types that were derived from event annotations[12] (as opposed to being originally annotated as binary relations). On the other hand, the version of the ChemProt dataset utilized by DyGIE++ has 2,432 documents and fairly well-defined binary relation types, so we expected to see relatively good performance. The PICKLE and SciERC datasets split the middle ground of the other models' performance on RE on their original domain (**Figure 5D**), which we expected due to the types of relation included in these datasets. Both PICKLE and SciERC only have a handful of relation types; as opposed to a dataset with more fine-grained relation types, each relation type in PICKLE or SciERC is broader and encompasses more semantic variation. Additionally, both PICKLE and SciERC have smaller amounts of documents than the ChemProt corpus. When we evaluate the models'

[12] An event is an n-*ary* relation (as opposed to a bin*ary* relation), where multiple entities are connected by relationships to form the event.

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

performance on their original domain datasets with respect to types, we see a reduction in performance for almost all models in entity extraction (except for ChemProt), while the opposite is true for relation extraction, where only the ChemProt model experiences a substantial decrease in performance when types are considered.

3.4 Analysis of the impact of original training domain on NER performance in plant science

The observation that SciERC models outperformed GENIA, ChemProt and SeeDev models on the PICKLE corpus for NER was unexpected because we expect model performance to correlate with training dataset similarity, and the GENIA, ChemProt and SeeDev corpora are all more intuitively similar to PICKLE than the SciERC corpus. To assess whether our intuition about the similarity of the other datasets to PICKLE was flawed, we looked at the overlap in whole-word vocabulary between corpora. All else being equal, models trained on a dataset with more similar vocabulary to PICKLE should have higher performance on PICKLE, as a model can't distinguish between words that weren't represented in its training vocabulary. Because the GENIA and SeeDev models performed worse on the PICKLE dataset than the ChenProt and SciERC datasets did, (**Figure 5**), we expected to see that the GENIA and SeeDev models' vocabularies contained a lower proportion of PICKLE's vocabulary than SciERC or ChemProt, and for ChemProt to have a similar proportion as SciERC, even though this defies the intuitive expectation of the similarity between these corpora, as this could explain the differences in model performance. However, when we look at the proportion of PICKLE's vocabulary that overlaps with other corpora (**Figure S4**), we see that only SeeDev has lower vocabulary coverage than SciERC, which is most likely due to its size. While our analysis doesn't specifically address the impact of size on vocabulary coverage, PICKLE contains twice the number of tokens than SeeDev, so much of PICKLE's vocabulary would likely not be present in SeeDev as a matter of course. We see what we would have intuitively expected before seeing model performance on the PICKLE dataset for the rest of the datasets, with GENIA and ChemProt having better vocabulary coverage of PICKLE than SciERC. Therefore, while vocabulary coverage alone can explain SeeDev's poor performance, it can't explain why SciERC performs so well, or why GENIA performs so poorly.

To test the hypothesis that GENIA's unexpectedly poor performance is due to the reduced proportion of types on which it was trained, we performed an analysis that examined the effect of adding types to the evaluation set that were not originally included in model training. We performed a type mapping between the types on which the biological models (ChemProt, GENIA, and SeeDev) were trained, and removed all type annotations from the PICKLE gold standard that did not map to the types that each biological model could predict (**Figure S5**, see Methods). The purpose of removing extraneous types was to determine if model performance would improve upon only being presented with types that were previously seen during training. Upon evaluating the models on their corresponding model-filtered dataset, we see that the GENIA models' performance drastically increases, and that the ChemProt model, while the confidence interval of its F1 score does still overlap with the F1 score on the full test set, does show signs of improvement (**Figure 6**, **Table S7**). In contrast, the SeeDev model does not exhibit a significant change in performance (**Figure 6**, **Table S7**). However, the SeeDev model contains many of the types in PICKLE, accounting for 81% of the original PICKLE test set, which explains why filtering did not impact performance. We further explore how filtering impacted performance by looking at changes in precision and recall (**Figure 7**, **Table S7**). For the SeeDev model, we see that precision and recall do not exhibit major changes, which is

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

consistent with the hypothesis that the high level of original dataset coverage by the SeeDev types is responsible for the maintenance of performance on filtering. In contrast, for both the GENIA and ChemProt models, we see that precision doesn't change substantially (**Figure 7A**, **Table S7**), but that recall does (**Figure 7B**, **Table S7**). As both GENIA and ChemProt's types don't offer high coverage of the PICKLE types like SeeDev does (23% and 54% of the test set accounted for by types in the original PICKLE test set, respectively), filtering helps improve the recall by removing types that the model was unable to predict. These results support the hypothesis that it is the addition of new types that drive the unexpectedly low performance of the GENIA models.

While the SciERC model could not be included in the previous analysis, as its types were too different to be semantically mapped back to the PICKLE categories, investigating the proportion of SciERC's predictions that pertain to each of its categories also helps us understand why the model performs so well in such a distinct domain. The SciERC model predicts 6 types: Method, Task, Material, OtherScientificTerm, Metric, and Generic. We hypothesized that the presence of the OtherScientificTerm category was allowing the model to predict a wide range of entities in the PICKLE dataset. When we look at the predicted types of true positives for the full SciERC model, we in fact see that 49% of true positive predictions of the SciERC model pertain to the OtherScientificTerm prediction category, which corresponds to our expectation that OtherScientificTerm is functioning as a catch-all category. Surprisingly, we also see that 28% of the true positive predictions pertain to the Method category. When we look at the distribution of gold standard types in these two categories, we see that both Method and OtherScientificTerm function as general catch-all categories for the types in the PICKLE dataset, as they both contain predictions corresponding to many of the PICKLE types (**Figure S6**).

4. Discussion

The PICKLE corpus provides publicly available annotation guidelines and a corpus for both entity and relation annotation in the plant sciences, using best practices adapted from other domains. We have demonstrated that our corpus size is large enough to train robust models for NER and RE in the plant sciences that have performance equivalent to or better than the state of the art in other domains, and that our corpus quality is equivalent to gold standard corpora in other domains.

We found that our initial IAA goal of 0.6 was a reasonable expectation for achievable IAA because, at the 6th set of documents for entity annotation, we reached an IAA of ~0.6, to stop the refinement cycle because, based on the feedback from the annotators concerning observations about the remaining disagreements. Further clarifications in the guidelines would have resulted in "overfitting" the instructions to our particular set of documents, i.e., defining narrow rules that would only ever apply to a few specific documents and are not generalizable. The IAA improved over the course of annotation for both entities and relations, but relation annotation only improved in a statistically significant manner when types were considered. This reflects the fact that while annotators became more consistent in choosing types for relations over time, the ability to identify whether or not any relation should be annotated between two entities did not follow the same improvement trajectory. We contextualized our relation IAA by comparing it to the SeeDev corpus IAA. The ChemProt and BioInfer corpora do not report an IAA score for their relation annotations; however, the SeeDev corpus does report an IAA score for events (which are decomposed to create

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

their binary relations) that compares the initial annotations for the finalized gold standard (as opposed to annotators' agreement with one another), which is an F1 of 0.55-0.65 when types are considered, and 0.60-0.72 when types are ignored. We also performed a literature search for additional biomedical corpora with comparable schema, and while we did not find corpora with comparable entity schema that reported IAA scores, we found that we can compare our relation IAA to the ChEBI corpus' (Shardlow *et al.*, no date). Across all categories of relations, ChEBI's annotators averaged an IAA of 0.63 (no confidence interval reported), while our final strict IAA was 0.49, with a confidence interval from 0.18 to 0.80, which substantially overlaps the mean of the ChEBI average relation IAA. Using SeeDev and ChEBI as a comparison, the relation IAA achieved on PICKLE is in line with other published corpora with similar levels of schema type specificity.

Part of the PICKLE corpus's utility lies in its application to the evaluation of downstream NLP methods like NER and RE. Using PICKLE, we have demonstrated that existing off-the-shelf NLP models for NER and RE are insufficient for high-quality NER and RE in the plant sciences. Additionally, PICKLE's utility as a training dataset has been established, as models trained on the PICKLE dataset achieve comparable performance to that of models from other domains on their original domain. Our analysis of model performance also yielded the interesting result that the similarity of a pretraining domain to the target domain does not necessarily correlate with the performance of the pretrained model on the target domain. The performance discrepancy can be attributed to size and vocabulary similarity in the case of SeeDev, and is explained by the introduction of new entity types for the GENIA and ChemProt models. We have additionally presented evidence that the SciERC model performs well in the plant science domain as a result of having two ontological terms, Method and OtherScientificTerm, that are successful in predicting a catch-all cross section of the PICKLE types. However, it remains unclear exactly why the Method category is able to function as a catch-all term. We leave to future work a deeper exploration of the relationship between domain similarity and transfer model performance.

The initial PICKLE corpus discussed in this work was built using abstracts related to the search terms "jasmonic acid" and "gibberellic acid". We chose these hormones to anchor our corpus because they are present nearly ubiquitously across land plants, and are central players in the plant growth-defense tradeoff (Huot *et al.*, 2014). Starting with a narrower subject breadth reduced variability introduced by broader topics to help the annotators better learn and develop the guidelines, and simultaneously, choosing such ubiquitous plant hormones kept the corpus relevant to a meaningful portion of plant science research. However, this choice means that the PICKLE corpus is not necessarily generalizable to all of plant science in its current form. Thus, the evaluation and extension of PICKLE's generalizability should be a future focus.

One important area for future work on the PICKLE corpus is to add coreference annotations, which link different mentions of entities that represent the same underlying real-world object. One of the advantages of the DyGIE++ models compared to similar architectures is that coreference resolution can be incorporated to improve model performance on entity and relation extraction; since we did not include coreference annotations in our initial work on the PICKLE corpus, we were unable to take advantage of this feature. Adding coreference resolution could also improve model performance further; we hypothesize that the largest benefit would be seen in relation extraction, which is most important for downstream applications like building knowledge graphs. To build knowledge graphs with the model trained on the PICKLE corpus, the model would be applied to a

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

large body of abstracts from the plant science domain to obtain relationship triples, which form a rudimentary knowledge graph. Coreference resolution is also important in graph building, as it prevents the proliferation of syntactically unique nodes that all represent the same semantic object.

5. Conclusion

The PICKLE corpus is an open-source, high-quality natural language corpus of 250 abstracts in the plant sciences, with gold-standard manual annotations of plant science entities and relations. We have also provided annotation guidelines and instructions that can be used for further annotation efforts to expand this corpus, creating a large set of training documents that can be used in downstream NLP tasks such as training plant science-specific NER and RE models. We have demonstrated the downstream utility of the PICKLE corpus by training a joint NER and RE model with the DyGIE++ model architecture, and have shown that the PICKLE model outperforms available DyGIE++ pre-trained models from other domains as well as a model trained on the SeeDev corpus from the plant science domain on plant science data. Additionally, PICKLE's performance is comparable to the performance of the other models tested in their original target domains, indicating that the PICKLE corpus is of sufficient quality to be useful in information extraction tasks in the plant science domain. Finally, we have performed an analysis of the effect of introducing new types to a pretrained model in a given domain, and have demonstrated that the introduction of new types has an outsized effect on model performance. In summary, the guidelines, corpus, and proof-of-concept studies serve as the foundation for future efforts in developing NLP toolkits for application in the plant science domain. With the advances in developing language models, future work based on our study will facilitate the construction of domain-specific models for automated knowledge extraction from plant science literature.

6. Data availability

There are two components to the annotation data for the PICKLE corpus: the brat-formatted data, and the DyGIE++ (jsonl)-formatted data. The brat-formatted data is the set of `.txt`/`.ann` files required to render the annotations in the brat annotation tool, contained in a `.zip` file. The second component is a set of `.jsonl` files, containing the annotations in the format required for input into the DyGIE++ pipeline. Both the brat- and jsonl-formatted components are available on Zenodo (https://zenodo.org/records/10076664), and the `.jsonl` format of the dataset is also available as a Huggingface Dataset (https://huggingface.co/datasets/slotreck/pickle).

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Figure Legends

Figure 1. Entity and relation ontologies for the PICKLE corpus. (A) The entity ontology, based on the GENIA corpus term ontology. There are three sub-hierarchies: Chemicals (blue), Organisms (purple), and Anatomy (green). (B) The relations ontology, based on the GENIA static relations ontology and the BioNLP13 Gene Network Regulation task relations. There are two sub-hierarchies: Causal (red) and Static Relations (yellow).

Figure 2. Iterative improvement of annotation guidelines and ontologies. Once the guidelines and ontologies were written or revised (A), they were given to annotators along with a set of abstracts for annotation (B). The inter-annotator agreement (IAA) was then calculated for each document and pair of annotators (C). The resulting fine-grained IAA, along with qualitative observations of disagreements and feedback from the annotators, was used to improve the ontologies and annotation guidelines. Once the target IAA of 0.6 was reached (D), the cycle was stopped, and annotations were unified by the author of the guidelines in order to create the final gold-standard set of annotations. This process was performed twice; once for entity annotations, and then, once there was a gold entity standard, again for relations, using the gold-standard entity annotations as a starting point.

Figure 3. Analysis of impact of training corpus size on model performance. Entity (A) and relation (B) performance for models trained on datasets varying in size from 10 documents to 500 documents. The top of each plot contains the bootstrapped performance estimates on the test set, while the bottom of each plot contains the best validation score pulled directly from the model during training. All performances were evaluated with consideration of entity and relation types. Points are staggered around their x-axis value for better clarity.

Figure 4. Double-blind inter-annotator agreement. IAA for entity (A) and relation (B) annotations. The dark blue dotted line is the double-blind IAA calculated with a strict tolerance, and the light blue dash-dot line in (B) is the double-blind IAA calculated with a loose tolerance. To compute IAA, all possible pairs of annotators were compared against one another, and their IAA scores averaged; error bars represent the standard error. The error bars for Round 1 and Round 7 don't overlap for both entities and strict tolerance relations, indicating a significant improvement in IAA from the beginning to the end of the refinement process. In the boxes under each round, a is the number of annotators that participated in the round, and n is the number of documents that were annotated.

Figure 5. Model performance on entity and relation extraction. F1 for (A) entity extraction on the PICKLE test set, (B) entity extraction on the models' original domain test sets, (C) relation extraction on the PICKLE test set, and (D) relation extraction on the models' original domain test sets. When applied to the original domain's test set, evaluations were performed with (purple triangles) and without (orange circles) consideration of the predicted entity and relation types. Models with no

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

values indicate that there was no result available for that model; either the dataset was not publicly available (ACE05), or the model didn't predict relations (GENIA, GENIA lightweight).

Figure 6. Performance of all models on the model-filtered PICKLE datasets. F1 scores for models on the original PICKLE test set (pink circles) and on the PICKLE test set filtered to the types that each model predicts (blue triangles).

Figure 7. Changes in precision and recall after dataset filtering. (A) Precision and (B) recall on the full (red circles) and model-filtered (blue triangles) PICKLE test sets.

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Author contributions

SL and SHS developed the project idea. SL designed the ontologies & annotation guidelines and wrote code to collect abstracts, unify annotations, apply and evaluate models, and create figures, and manually reviewed & unified all abstracts and wrote the initial draft and figure legends. KSA contributed to analyses of unexpected model performance. KSA, MLS, AS, BB, TR, and AS annotated abstracts and provided feedback for improvements to annotation guidelines. MG provided ideas for several analyses. SHS and MG oversaw project progress and provided feedback on the design of study. All authors participated in the drafting and revision of the manuscript.

Acknowledgements

We would like to thank Harry Shomer for the PICKLE acronym, and for his help with annotation. This work was supported by the National Science Foundation Research Traineeship Program (DGE-1828149) as a fellowship to SGL, BNIB, and KSA, the U.S. Department of Energy Great Lakes Bioenergy Research Center (BER DE-SC0018409 to SHS), and additional National Science Foundation grants (DGE-1828149, IOS-2218206

to SHS; IOS-2107215 and MCB-2210431 to MDL and SHS).

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

References

Angeli, G., Johnson Premkumar, M.J. and Manning, C.D. (2015) 'Leveraging Linguistic Structure For Open Domain Information Extraction', in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Beijing, China: Association for Computational Linguistics, pp. 344–354. Available at: https://doi.org/10.3115/v1/P15-1034.

Bada, M. *et al.* (2012) 'Concept annotation in the CRAFT corpus', *BMC Bioinformatics*, 13(1), p. 161. Available at: https://doi.org/10.1186/1471-2105-13-161.

Bada, M. and Eckert, M. (no date) 'CRAFT concept annotation guidelines', p. 47.

Boguslav, M. and Cohen, K.B. (no date) 'Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing'.

Bossy, R. *et al.* (2015) 'Overview of the gene regulation network and the bacteria biotope tasks in BioNLP'13 shared task', *BMC Bioinformatics*, 16(10), p. S1. Available at: https://doi.org/10.1186/1471-2105-16-S10-S1.

Bougiatiotis, K. *et al.* (2020) 'Drug-Drug Interaction Prediction on a Biomedical Literature Knowledge Graph', in M. Michalowski and R. Moskovitch (eds) *Artificial Intelligence in Medicine*. Cham: Springer International Publishing (Lecture Notes in Computer Science), pp. 122–132. Available at: https://doi.org/10.1007/978-3-030-59137-3_12.

Celebi, R. *et al.* (2019) 'Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings', *BMC Bioinformatics*, 20(1), p. 726. Available at: https://doi.org/10.1186/s12859-019-3284-5.

Chaix, E. *et al.* (2016) 'Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016.', in *Proceedings of the 4th BioNLP Shared Task Workshop*. *Proceedings of the 4th BioNLP Shared Task Workshop*, Berlin, Germany: Association for Computational Linguistics, pp. 1–11. Available at: https://doi.org/10.18653/v1/W16-3001.

Cho, H. *et al.* (2022) 'Plant phenotype relationship corpus for biomedical relationships between plants and phenotypes', *Scientific Data*, 9(1), p. 235. Available at: https://doi.org/10.1038/s41597-022-01350-1.

Choi, W. *et al.* (2016) 'A corpus for plant-chemical relationships in the biomedical domain', *BMC Bioinformatics*, 17(1), p. 386. Available at: https://doi.org/10.1186/s12859-016-1249-5.

Dai, A.M., Olah, C. and Le, Q.V. (2015) 'Document Embedding with Paragraph Vectors', *arXiv:1507.07998 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1507.07998 (Accessed: 4 March 2021).

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Dai, Y. *et al.* (2020) 'Drug-Drug Interaction Prediction with Wasserstein Adversarial Autoencoder-based Knowledge Graph Embeddings', *arXiv:2004.07341 [cs, stat]* [Preprint]. Available at: http://arxiv.org/abs/2004.07341 (Accessed: 12 April 2021).

Derpanis, K. (2005) 'Mean Shift Clustering'. Available at: http://www.cs.yorku.ca/~kosta/CompVis_Notes/mean_shift.pdf (Accessed: 2 December 2022).

Fricke, S. (2018) 'Semantic Scholar', *Journal of the Medical Library Association : JMLA*, 106(1), pp. 145–147. Available at: https://doi.org/10.5195/jmla.2018.280.

Hedden, P. (2020) 'The Current Status of Research on Gibberellin Biosynthesis', *Plant and Cell Physiology*, 61(11), pp. 1832–1849. Available at: https://doi.org/10.1093/pcp/pcaa092.

Hripcsak, G. (2005) 'Agreement, the F-Measure, and Reliability in Information Retrieval', *Journal of the American Medical Informatics Association*, 12(3), pp. 296–298. Available at: https://doi.org/10.1197/jamia.M1733.

Huot, B. *et al.* (2014) 'Growth-defense tradeoffs in plants: a balancing act to optimize fitness', *Molecular Plant*, 7(8), pp. 1267–1287. Available at: https://doi.org/10.1093/mp/ssu049.

Karim, M.R. *et al.* (2019) 'Drug-Drug Interaction Prediction Based on Knowledge Graph Embeddings and Convolutional-LSTM Network', *arXiv:1908.01288 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1908.01288 (Accessed: 12 April 2021).

Kim, B., Choi, W. and Lee, H. (2019) 'A corpus of plant–disease relations in the biomedical domain', *PLOS ONE*. Edited by P. Pławiak, 14(8), p. e0221582. Available at: https://doi.org/10.1371/journal.pone.0221582.

Kim, J.-D. *et al.* (2003) 'GENIA corpus--a semantically annotated corpus for bio-textmining', *Bioinformatics*, 19(Suppl 1), pp. i180–i182. Available at: https://doi.org/10.1093/bioinformatics/btg1023.

Kim, J.-D. *et al.* (no date) 'GENIA Ontology', p. 10.

Landhuis, E. (2016) 'Scientific literature: Information overload', *Nature*, 535(7612), pp. 457–458. Available at: https://doi.org/10.1038/nj7612-457a.

Larmande, P., Do, H. and Wang, Y. (2019) 'OryzaGP: rice gene and protein dataset for named-entity recognition', *Genomics & Informatics*, 17(2), p. e17. Available at: https://doi.org/10.5808/GI.2019.17.2.e17.

Liu, Z. *et al.* (2020) 'Named Entity Recognition for the Horticultural Domain', *Journal of Physics: Conference Series*, 1631(1), p. 012016. Available at: https://doi.org/10.1088/1742-6596/1631/1/012016.

Luan, Y. *et al.* (2018) 'Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction', *arXiv:1808.09602 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1808.09602 (Accessed: 26 October 2020).

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Mohamed, S.K., Nounu, A. and Nováček, V. (2020) 'Biological applications of knowledge graph embedding models', *Briefings in Bioinformatics*, p. bbaa012. Available at: https://doi.org/10.1093/bib/bbaa012.

Mohamed, S.K., Nováček, V. and Nounu, A. (2019) 'Discovering Protein Drug Targets Using Knowledge Graph Embeddings', *Bioinformatics*. Edited by L. Cowen, p. btz600. Available at: https://doi.org/10.1093/bioinformatics/btz600.

Neumann, M. *et al.* (2019) 'ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing', *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327. Available at: https://doi.org/10.18653/v1/W19-5034.

Nicholson, D.N. and Greene, C.S. (2020) 'Constructing knowledge graphs and their biomedical applications', *Computational and Structural Biotechnology Journal*, 18, pp. 1414–1428. Available at: https://doi.org/10.1016/j.csbj.2020.05.017.

Pyysalo, S. *et al.* (2007) 'BioInfer: a corpus for information extraction in the biomedical domain', *BMC Bioinformatics*, 8(1), p. 50. Available at: https://doi.org/10.1186/1471-2105-8-50.

Pyysalo, S. *et al.* (2009) 'Static relations: a piece in the biomedical information extraction puzzle', in *Proceedings of the Workshop on BioNLP - BioNLP '09*. *the Workshop*, Boulder, Colorado: Association for Computational Linguistics, p. 1. Available at: https://doi.org/10.3115/1572364.1572366.

Ruan, J. *et al.* (2019) 'Jasmonic Acid Signaling Pathway in Plants', *International Journal of Molecular Sciences*, 20(10), p. 2479. Available at: https://doi.org/10.3390/ijms20102479.

S., M.C., Lex, E. and Lalitha Devi, S. (2016) 'Named Entity Recognition for the Agricultural Domain', *Research in Computing Science*, 117(1), pp. 121–132. Available at: https://doi.org/10.13053/rcs-117-1-10.

Shardlow, M. *et al.* (no date) 'A New Corpus to Support Text Mining for the Curation of Metabolites in the ChEBI Database'.

Singh, G. *et al.* (2021) 'Extracting knowledge networks from plant scientific literature: potato tuber flesh color as an exemplary trait', *BMC Plant Biology*, 21(1), p. 198. Available at: https://doi.org/10.1186/s12870-021-02943-5.

Stenetorp, P. *et al.* (no date) 'brat: a Web-based Tool for NLP-Assisted Text Annotation', p. 6.

Wadden, D. *et al.* (2019) 'Entity, Relation, and Event Extraction with Contextualized Span Representations', *arXiv:1909.03546 [cs]* [Preprint]. Available at: http://arxiv.org/abs/1909.03546 (Accessed: 15 March 2021).

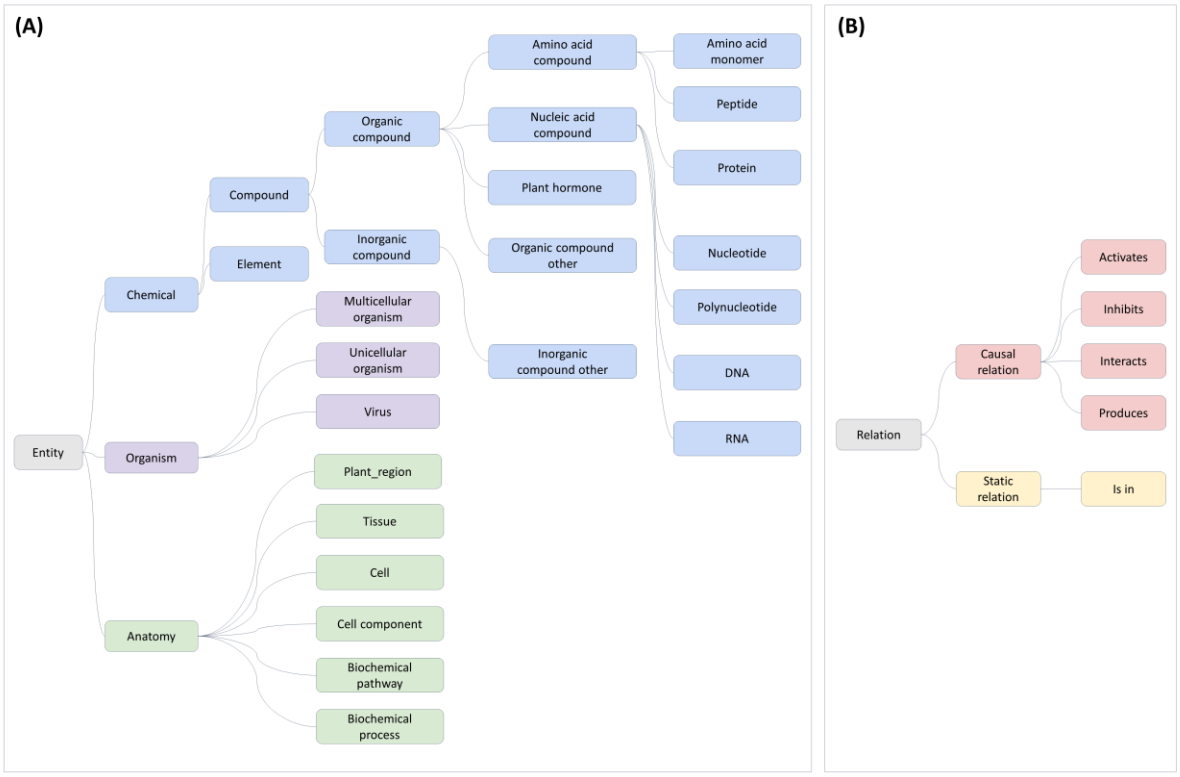
Walker, C. (2006) 'ACE 2005 Multilingual Training Corpus LDC2006T06'. Philadelphia: Linguistic Data Consortium.

Zhong, Z. and Chen, D. (2020) 'A Frustratingly Easy Approach for Joint Entity and Relation Extraction', *arXiv:2010.12812 [cs]* [Preprint]. Available at: http://arxiv.org/abs/2010.12812 (Accessed: 5 March 2021).

Accepted Manuscript

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

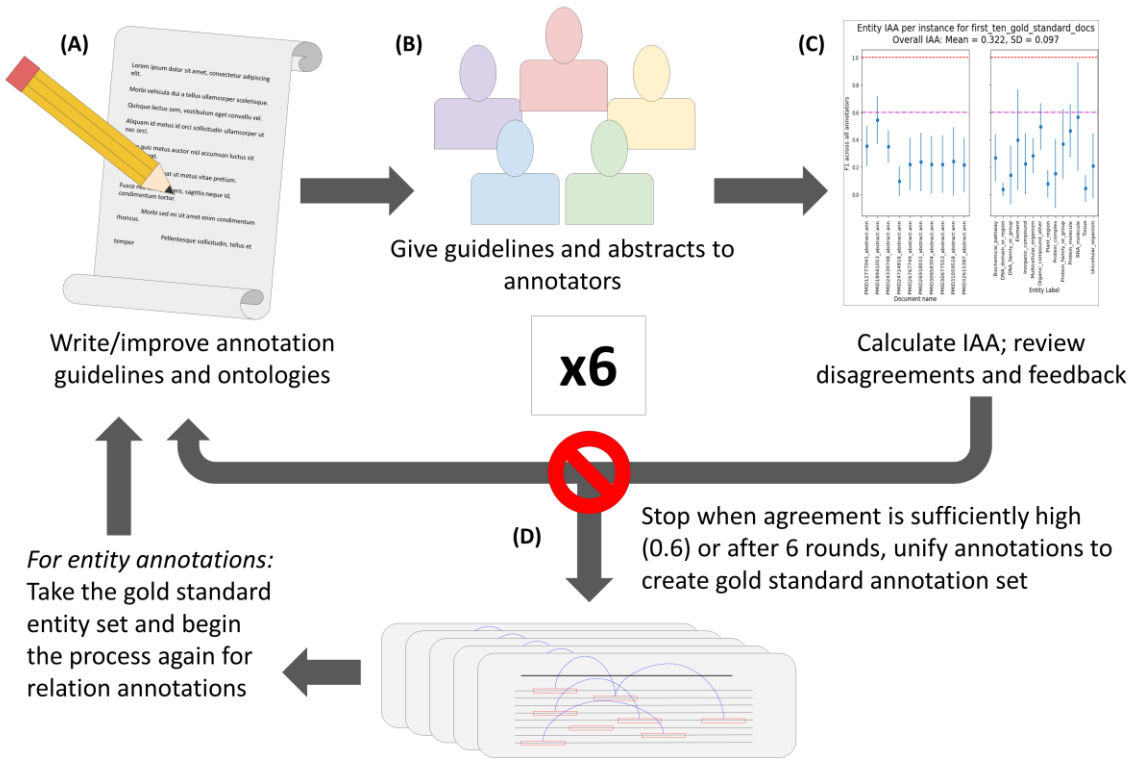
Figure 1



Accepted M

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

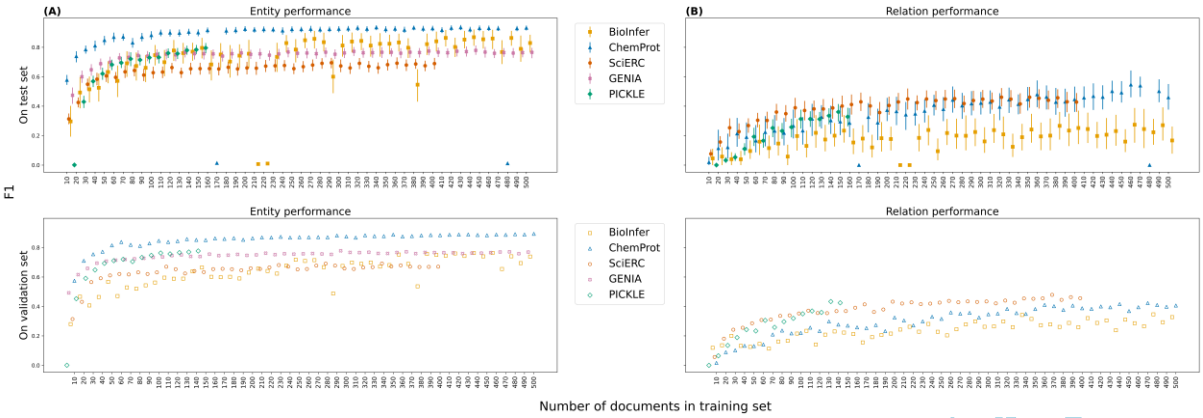
Figure 2



Accepted M

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

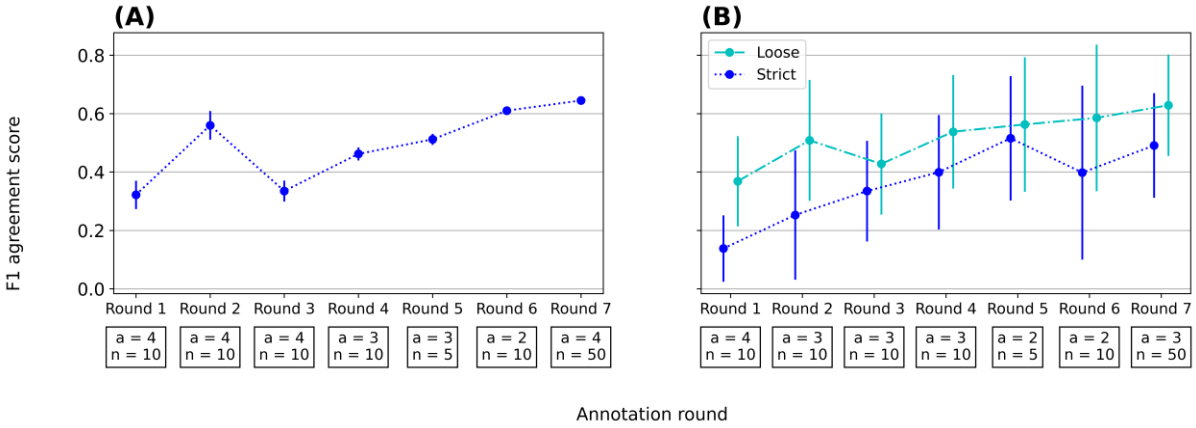
Figure 3



Accepted Manuscrip

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

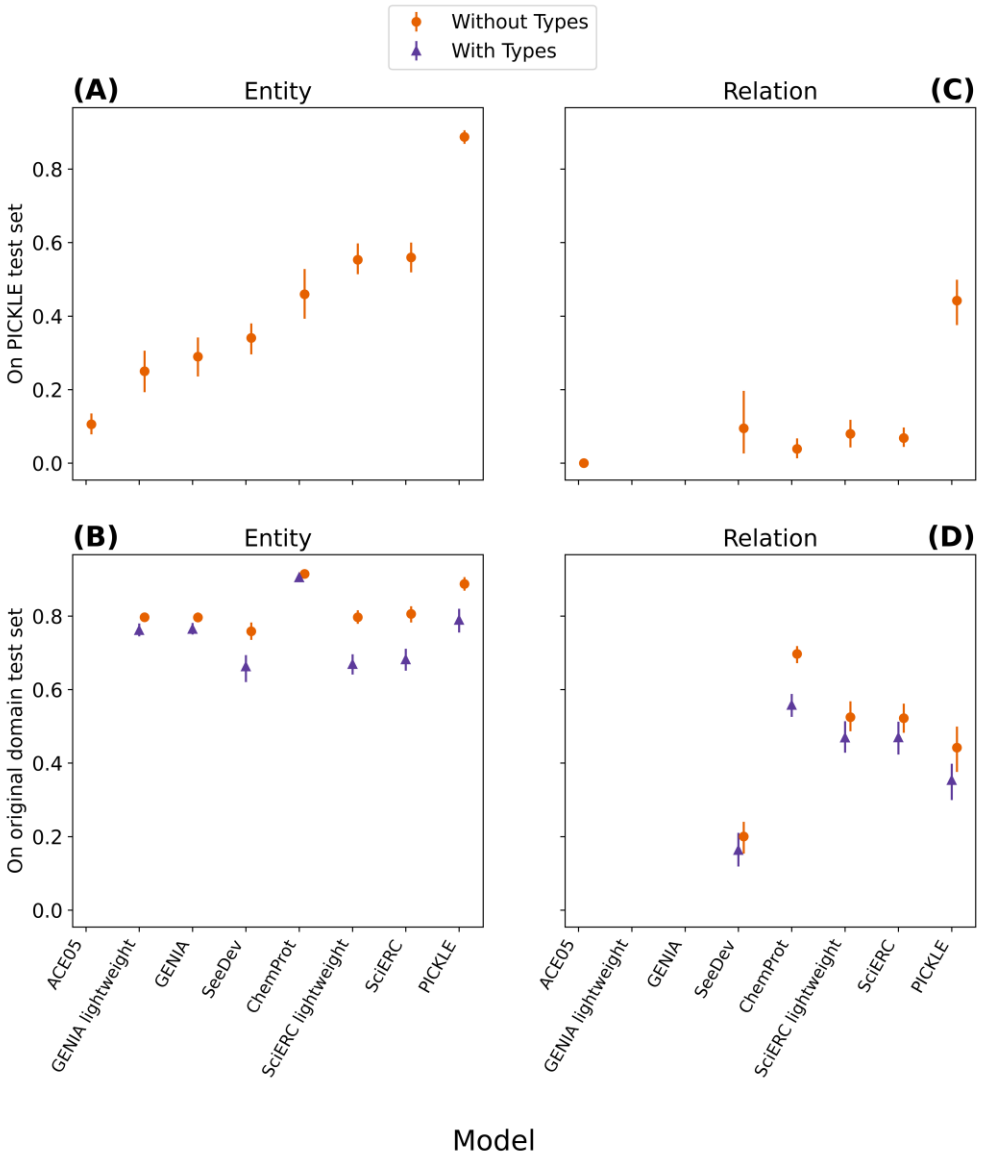
Figure 4



Accepted Manuscript

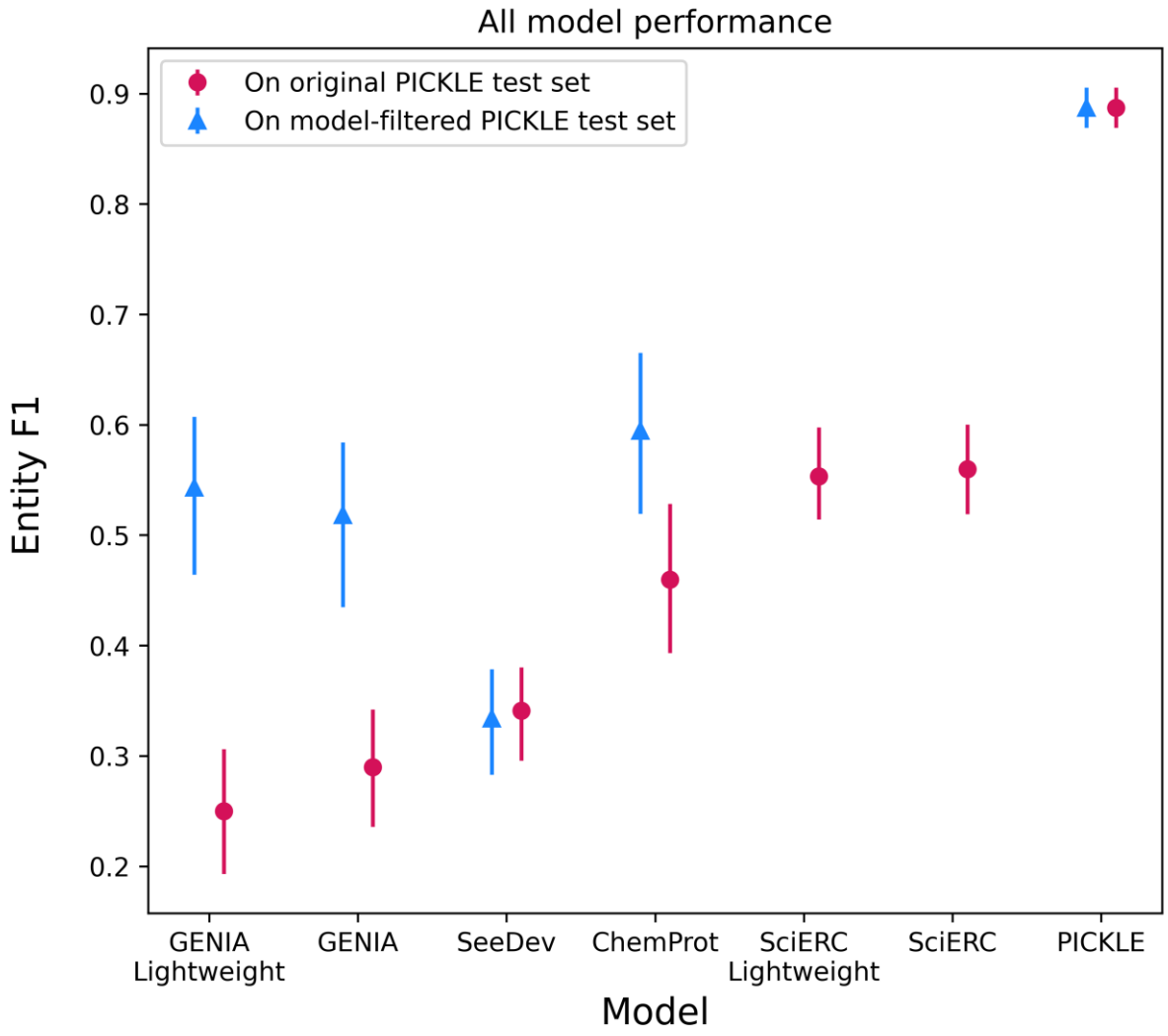
Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Figure 5



Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

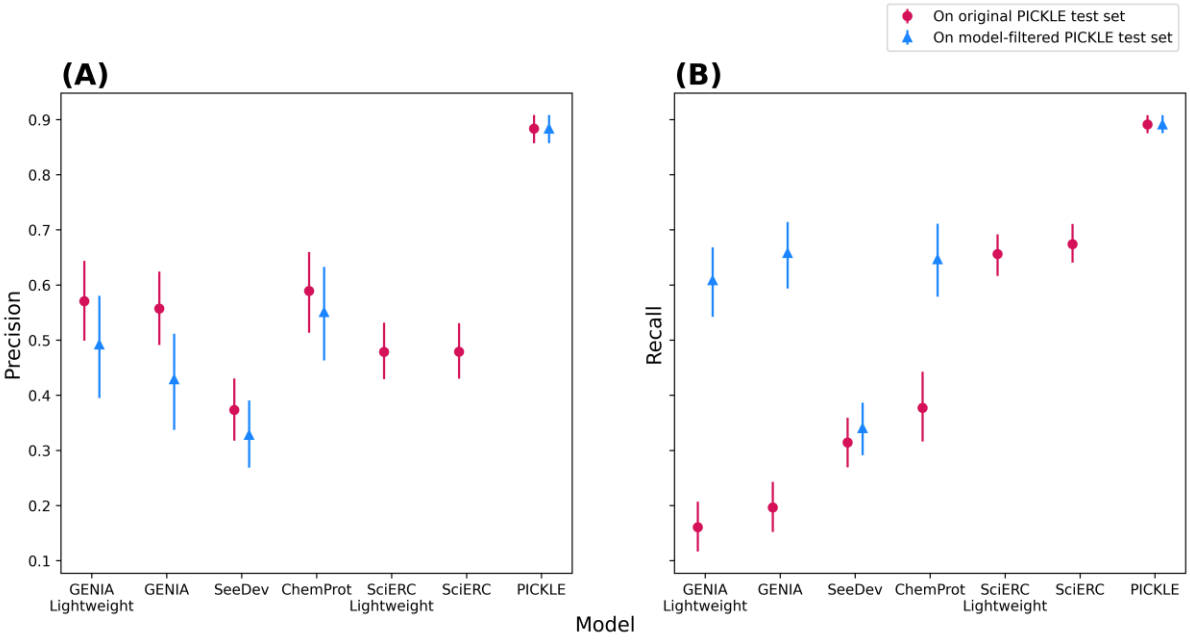
Figure 6



Accepted

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023

Figure 7



Accepted Manu

Downloaded from https://academic.oup.com/insilicoplants/advance-article/doi/10.1093/insilicoplants/diad021/7413143 by Michigan State University user on 14 November 2023