AAbAAC: An Annotated Corpus for Autoimmunity Information Extraction

Fabien Maury[1,2], Solène Grosdidier[3], Maud de Dieuleveult[1,*], Adrien Coulet[2,*]

[1]Inserm, Université Paris Cité, U1163 Institut Imagine, Paris, France,
[2]Inria, Inserm, Université Paris Cité, U1346 HeKA, Paris, France,
[3]Freelance researcher, The Hague, Netherlands,

[*]These authors contributed equally

Correspondence: fabien.maury@inserm.fr

Abstract

Despite advances in information extraction driven by deep learning and large language models, performance gaps remain in highly specialized biomedical fields, where domain-specific complexity poses challenges for generalist models. In this work, we focus on the domain of autoimmunity, where the main entities of interest are autoimmune diseases, autoantibodies (*i.e.*, molecules that may mark or cause these diseases), their molecular targets, their location in the body, and their associated clinical signs. Herein, we present AAbAAC (AutoAntibodies and Autoimmunity Annotated Corpus), a corpus of 115 abstracts selected from PubMed, where we manually annotated entities and their relationships. First, AAbAAC was used to evaluate several methods on the task of named entity recognition (NER), and secondly, to fine-tune NER models. Our study demonstrates the utility of AAbAAC for information extraction in the domain of autoimmunity, showing expected improvement in NER performance after fine-tuning. This illustrates the value of small-scale annotation efforts for specialized domains and contributes to the computational study of autoimmunity. The AAbAAC corpus is available at https://github.com/f-maury/AAbAAC.

1 Introduction

Autoimmune diseases arise from dysfunction of the adaptive immune system, whereby molecules of the self are targeted by self-antibodies, named autoantibodies. Those are generally detected in blood or cerebrospinal fluid, and widely used as a biomarker of autoimmune diseases, critically guiding differential diagnosis. Autoimmune diseases collectively affect an estimated 3–5% of the global population (Ahsan, 2023; Shapira et al., 2010). Yet, when considered individually, many of these diseases are rare (Hayter and Cook, 2012). In part due to the rarity of these pathologies, there is no single centralized resource including all relevant information about autoimmunity. Indeed, relevant bits of domain expertise are scattered across various sources, including many recent research papers. For this reason, information extraction is a relevant way to study autoimmune diseases, by automating the processing of large numbers of documents into structured, interoperable, and machine-readable knowledge bases.

Lately, BERT-based models have consistently achieved state-of-the-art performance in information extraction tasks, particularly when adapted through fine-tuning on domain-specific documents matching the target domain. While many domain-specific models exist, high-quality annotated data to adapt general models to highly specialized or niche domains are also lacking in many cases. To our knowledge, this is the case in human autoimmunity, where no manually annotated corpus dedicated to autoimmune diseases and autoantibodies exists to evaluate or fine-tune models specialized for the recognition of autoantibodies, autoimmune diseases, and their relationships. To fill this gap, we propose AAbAAC, a corpus of PubMed abstracts annotated for entities and relations related to autoimmunity, and demonstrate its value for named entity recognition (NER).

Section 2 reviews related works. Section 3 describes the methodology used to create the corpus, and Section 4 presents the resulting corpus. Section 5 introduces the experimental setup for named entity recognition (NER) using AAbAAC, and Section 6 reports the associated results. Finally, the paper concludes with perspectives for future work enabled by this new corpus.

2 Related works

2.1 Existing resources for autoimmunity

Some general and widely used knowledge bases and resources exist for broad biomedical

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 791–800

use, such as the Human Phenotype Ontology (HPO) (Gargano et al., 2024), Orphanet (Rath et al., 2012), or the UMLS (Bodenreider, 2004). Those include information about autoimmune diseases and autoantibodies, but are not sufficiently exhaustive, detailed, nor up-to-date for real-world studies. Indeed, autoimmunity is a specialized and evolving topic that is not the focus of generalist resources. They can nonetheless be used to generate an initial list of autoantibody names, that can then be used for information extraction.

More specialized resources have also been developed that include information about antibodies, autoantibodies, or related elements. For example, this is the case of IEDB (Vita et al., 2025), a database about epitopes, *i.e.*, regions of antigens bound by antibodies, or AAgAtlas (Wang et al., 2017), which focuses on autoantigens. Each of these resources focuses on a topic that is relevant to the study of autoimmunity, but none of them compile all the relevant knowledge in the field of autoimmunity for clinicians or researchers to use. However, for information extraction purposes, these resources could be used to build lists of autoantibody names from their antigen names.

Some annotated text corpora already exist for biomedical information extraction, but to the best of our knowledge, none is dedicated to human autoimmunity and autoimmune diseases. For example, the MedMentions corpus (Mohan and Li, 2019) tackles the biomedical domain in a general manner, with annotations of UMLS terms of most Semantic Types. As a result, it may include the most common autoimmune diseases and autoantibodies, but it does not include uncommon ones that do not correspond to UMLS terms, nor does it include relationships between the entities. The ABAG corpus (Dinh et al., 2022) focuses on antibodies and antigens, but is not specific to autoimmunity and does not provide annotations of relationships between entities. Therefore, existing resources are of limited use for autoimmunity-related information extraction, since isolating in these resources the annotations specific to autoimmunity would be challenging and result in limited coverage of autoimmunity-related concepts.

2.2 NER for autoimmunity

Many works make use of dictionaries and rule-based methods to conduct NER with decent performance. Such dictionaries can be built from either ontologies or databases. For example, a recent work by Remaki et al. (2025) used such a combination, including terms from SNOMED CT (Wang et al., 2002) and rules, to extract biology exams and drugs in relations to immune-mediated inflammatory diseases.

Another work by Subramanian and Ganapathiraju (2017) extracts antibody names using simple regular expressions based on the fact that antibodies are often written as "X antibody", or "antibody to X", where "X" is the target of the antibody. This method does not require a dictionary of possible target names, but fails at covering the various possible ways an antibody can be mentioned in text. Moreover, in the context of autoimmunity, such a method would struggle to differentiate autoantibodies from regular antibodies.

The development of multi-head attention and transformer-based architectures of language models enables efficient information extraction from long texts with more flexibility and context awareness than rule-based approaches. For NER tasks, encoder models such as BERT-based models (Devlin et al., 2019) perform competitively while being lighter and less computationally expensive to work with than large language models (LLM) (Naguib et al., 2024). The original BERT model has been adapted to many domains, including the biomedical domain. However, we did not find any model specialized for the recognition of autoimmunity-related entity types.

3 Corpus creation methods

3.1 Text selection

To select a set of relevant texts to be annotated, we first built a dictionary of autoantibody names and synonyms from HPO terms, which we used to query a PubMed API (Sayers, 2018) for titles and abstracts of scientific papers. Indeed, HPO includes terms describing positivity to some autoantibody detection tests (terms under *HP:0030057*). Most of them are of the form: "*anti-X antibody positivity*", where "*anti-X antibody*" is the name of an autoantibody that targets the antigen "*X*". We relied on this regular syntax to extract names of autoantibodies and manually reviewed results. We also generated lexical variants by reducing each autoantibody name to a core string through the removal of selected prefixes and suffixes, then recombining this core with alternative affixes. For example, "*anti-smooth muscle antibody*" was reduced to "*smooth muscle*" by removing "*anti*" and "*antibody*", and

one resulting variant was "*smooth muscle autoantibody*". Overall, the dictionary includes a total of 10,916 variations for 285 unique autoantibodies that originated from HPO release 2024-07-01. On average, each autoantibody in the dictionary has 38.30 variations, including synonyms that were already listed in the HPO.

For each autoantibody type in the dictionary, PubMed is searched with a query of the form "*(""autoantibody_name_1""[Title/Abstract] OR ""autoantibody_name_2""[Title/Abstract] AND humans[MeSH Terms] AND (antibodies[MeSH Terms] OR autoantibodies[MeSH Terms])*". MeSH term filters were added to select only papers tagged with "human" and either "antibodies" or autoantibodies. This process returned a total of 56,750 titles and abstracts. The full query we used is available in the project's GitHub repository.

Despite MeSH terms filtering, some of the obtained texts were irrelevant due to not being written in English or not being related to autoimmunity. For this reason, we enforced additional filters: abstracts were automatically pre-annotated with exact matches of autoantibody names from our HPO-derived dictionary, and with disease and autoantibody names identified by GLiNER. This pre-annotation was performed with a version of the GLiNER specialized for the biomedical field ("*urchade/gliner_large_bio-v0.1*"). This model is an early version of GLiNER large that is finetuned on PubMed abstracts.

Only texts containing at least one HPO autoantibody term match, or one disease pre-annotation and one autoantibody pre-annotation were kept, reducing the set of texts to 44,890.

Among these, 120 texts were randomly selected for manual annotation, out of which 5 were manually eliminated due to either not being related to autoantibodies or not being English texts, resulting in a final set of 115 texts. The initial selection size was 120 because we aimed to have a corpus of around 100 texts and expected a few of them to be eliminated.

3.2 Annotation guidelines and process

Annotation was performed by 4 annotators with a background in either medical informatics, biology, or pharmacy. This group elaborated annotation guidelines through two preliminary annotation batches to identify main entity types, difficulties, and edge cases before the start of the annotation campaign (Hovy and Lavid, 2010). Annotation guidelines are available in the following GitHub repository: https://anonymous.4open.science/r/aabaac-FA3F. Five types of entities were annotated: "*Autoantibody*", "*Autoantibody target*", "*Disease*", "*Symptom or clinical sign*", and "*Autoantibody location*"; and ten different types of relationships: "*is associated with*", "*is not associated with*", "*may be associated with*", "*is caused by*", "*is target of*", "*is reference to*", "*discontiguous entity*" (this relation is used to link several separated parts of a single entity), "*is synonym of*", "*is located in*", and "*is subclass of*".

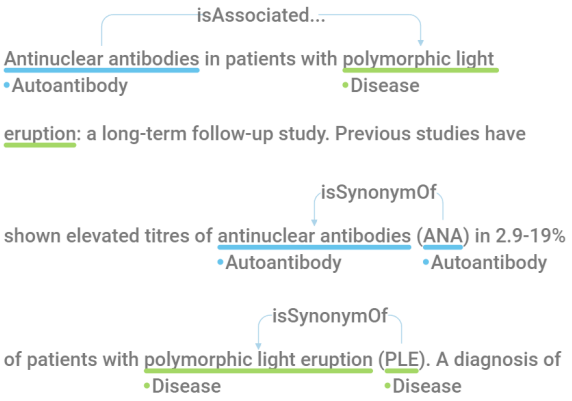

Figure 1: Example of text annotated in the Doccano web interface.

To make the annotation task easier for annotators, it was agreed to not annotate obvious relationships between nested entities, which were automatically added after manual annotation. For instance, when a "*Target*" is found inside the span of an "*Autoantibody*", which typically happens in structures like "*anti-X antibody*", the annotators were instructed not to add an "*is target of*" relation. Similarly, when an "*Autoantibody*" is found inside the span of a related "*Symptom or clinical sign*", the annotators were instructed not to add an "*is associated with*" relationship. Those were automatically added after manual annotation.

Each text was annotated independently by 2 annotators, but no single annotator annotated all texts. Each annotator was paired with each other annotator with equal frequency, and each pair annotated the same number of texts. Annotation was performed using Doccano (Nakayama et al., 2018) on texts where entities had been pre-annotated, but not relations. These pre-annotations were obtained by string matching and GLiNER. Doccano web interface is illustrated in Figure 1.

Texts were sent to the annotators in small numbered batches of around 4 texts, and annotators were free to annotate them in any order, in a single or multiple sessions, and at any time before the end of the campaign. Finally, an adjudication took place where annotations from both annotators for each text were confronted, and a consensual final annotation was decided by the head annotator.

4 Resulting corpus description

This section includes descriptive counts and statistics from the final adjudicated corpus of 115 texts. The variability of the length of the texts is high (from 50 to 3600 characters, as per Table 1), as is their annotation content (from 2 to 88 entities, and from 0 to 55 relations). This is partially due to the fact that some of the texts are limited to an article title with an empty abstract.

Metric	Min	Max	Average	Median
Entities per text	2	88	29.18	27
Relations per text	0	55	15.47	14
Text length (chars)	50	3600	1500.94	1500
Sentences per text	1	64	13.46	12

Table 1: General corpus statistics.

Imbalance is also observed in the number of entities of each type, with especially only 64 (see Table 2) occurrences of "*Autoantibody location*", whereas *Disease* is the most common entity with 1159 occurrences.

Entity type	Count	Avg/text	Median/text
Autoantibody	891	7.75	6
Autoantibody location	64	0.56	0
Autoantibody target	581	5.05	3
Disease	1159	10.08	8
Symptom or clinical sign	661	5.75	5
Total	3356	29.18	27

Table 2: Counts and distributions in the corpus by entity type.

Extremely variable numbers of occurrences can be observed for each relation type. Most relation annotations consist of "*is associated with*" (735, see Table 3), which can be explained by its rather broad definition, "*is target of*" (494), and "*is synonym of*" (226). Some relation types have almost never been used, such as "*is caused by*" (13), which could be explained by the cautious tone used in scientific literature, and the difficulty to establish causation with certainty, and "*is reference to*" (2), which could be explained by its constrained definition in our guidelines and redundancy with "*is synonym of*".

Relation type	Count	Avg/text	Median/text
discontiguousEntity	106	0.92	0
isAssociatedWith	735	6.39	5
isCausedBy	13	0.11	0
isLocatedIn	73	0.64	0
isNotAssociatedWith	24	0.21	0
isRefTo	2	0.02	0
isSubClassOf	61	0.53	0
isSynonymOf	226	1.97	1
isTargetOf	494	4.30	3
mayBeAssociatedWith	45	0.39	0
Total	1779	15.47	14

Table 3: Counts and distributions in the corpus by relationship type.

To evaluate the difficulty of the annotation task, inter-annotator agreement measures were computed between each pair of annotators, and between each annotator and the final annotations. Since the annotators simultaneously had to decide whether a text contained annotations, where in the text were these annotations located, if any, and of what type they were, Mathet's gamma coefficient (Mathet et al., 2015; Titeux and Riad, 2021) was used. Mathet's gamma is a measure that takes into account both disagreement on span limits (unitizing) and on label type (categorizing) for entity annotations. To measure agreement on relation annotations, three different comparison functions (strict match, flexible with tolerance for some disagreement, and super-flexible with tolerance for more disagreement) were used to compute F1-scores between the annotations of each annotator pair. We report a general average gamma coefficient between pairs of annotators of 0.67. The average gamma was 0.74 for "*Autoantibody*", 0.71 for "*Autoantibody location*", 0.44 for "*Autoantibody target*", 0.81 for "*Disease*", and 0.54 for "*Symptom or clinical sign*".

Relationship annotations are intrinsically more complex, and this is particulary the case in this work where annotators had 10 different relation types to choose from. To assess this difficulty quantitatively, we measured the inter-annotator agreement for relationships using an F1-score. We obtain a F1-score of 0.37 for all relation types together, using a custom flexible comparison metric tolerating some small overlap differences, and some entity type confusions.

Considering all relation types jointly, the primary source of disagreement was *isAssociatedWith*, because it was the most frequently used relation

type and because its broad definition allowed for subjective interpretation. When examining relation types individually, some exhibited lower agreement scores than *isAssociatedWith*; however, their contribution to the overall disagreement remained limited due to their lower frequency. In particular, *isRefTo* was used only twice and obtained an F1 score of 0. Given its limited use, removing this relation type from the annotation scheme could be considered. Overall, the relatively low inter-annotator agreement observed for relations highlights the difficulty of annotating texts from a specialized scientific domain. Furthermore, the complexity of the annotation scheme made the annotation task challenging for annotators.

5 Evaluating NER methods

5.1 Experimental setup

To demonstrate the utility of the AAbAAC corpus in autoimmunity NER, several methods were compared.

The corpus was randomly divided into train and test following an 80 - 20 split. This procedure was repeated independently five times, resulting in five distinct train–test partitions. Supervised approaches were fine-tuned separately on each training set, and all approaches were evaluated on the corresponding test sets. Reported results are the average performance across the five test sets.

Chunking of longer texts occurred after the split so that all chunks from the same text remain part of the same set. This prevents information leakage from the train set to the test set.

5.2 QuickUMLS

The first method tested is QuickUMLS (Soldaini and Goharian, 2016) on the task of recognition of mentions of "*Autoantibody*" and "*Disease*" entities only. One dictionary was created for each of these entity types. The dictionary for "*Disease*" was created by selecting all the UMLS terms from Semantic Type "*Disease or Syndrome*", and only keeping English strings associated with these terms. The "*Autoantibody*" dictionary was by combining the dictionary made from an intial list of HPO terms for text selection (see Subsection 3.1) and was then enriched from LOINC (Forrey et al., 1996) and manual additions. Moreover, name variations for autoantibodies were automatically added to the dictionary. QuickUMLS was used with default parameters except for the threshold, which was set to 0.9.

5.3 GLiNER

Although standard BERT-based models achieve state-of-the-art performance in NER, they rely on predefined annotation schemas and task-specific fine-tuning, in contrast to the GLiNER architecture, which supports more flexible, label-agnostic entity recognition (Zaratiana et al., 2024). GLiNER does not need a fixed list of entity types, but can work with any label passed as input at inference time. While studying a BERT model would be of interest in our use case, we decided to work with the GLiNER architecture (GLiNER 2.1 models: urchade/gliner_large-v2) on the task of NER for entity types related to autoimmunity, notably autoantibodies. Indeed, this allows for more flexibility with regards to the set of entity types to be identified, and our annotation scheme was initially not fixed. In this setup, texts longer than 384 tokens were chunked by cutting at the end of the last sentence before reaching the maximum length, and each chunk was tokenized using GLiNER's tokenizer.

GLiNER base models (small, medium, and large) were evaluated in a zero-, two-shot, and fine-tuned setting. In the GLiNER architecture, the labels are passed to the model along with the input text at inference time. The way labels are formulated may affect performances, and it may be useful to reformulate them or use synonyms to see what returns best results (e.g., "*Autoantibody target*" or "*Autoantigen*"). At this step, labels that seemed the most concise and clear were picked, but except from light changes, we did not conduct an extensive experiment to search for potential optimal alternatives. GLiNER models were evaluated on the task of NER for the following entities of interest: "*Autoantibody*", "*Autoantibody location*", "*Autoantibody target*", "*Disease*", "*Symptom or clinical sign*", on the 5 different test sets obtained from our split of the AAbAAC corpus. These entity types are the strings that were passed to GLiNER along with texts, as part of the input.

For the 2-shot setting, we passed 2 short, manually made-up examples of each entity type before the actual input text. The examples followed the format: '"sentence": "The patient was admitted after complaining of thigh pain."\n"Symptom or clinical sign": "thigh pain"'. The same 2 example sentences for each type were passed along each input text. In rare cases, the input text was long and

close to the maximal allowed chunk length, and adding the extra examples has led to truncation, which may have affected the performance slightly. For the 2-shot experiment, the GLiNER threshold parameter, which acts as a confidence filter for the model's output was set to 0.3, but for 0-shot and fine-tuned configuration it was set to 0.5.

GLiNER base models were fine-tuned using the train set from each split, and the fine-tuned models were evaluated on the associated test set. For fine-tuning, the tokenized texts of the train set are shown to the model along with their entity annotations. For evaluation, the chunked but untokenized texts of the test set are passed to the model along with the types of entities to identify.

5.4 MedGemma

Google's MedGemma model for text (google/medgemma-27b-text-it) was evaluated in 0-shot, 2-shot setting and finally fine-tuned using AAbAAC. MedGemma being a generative model, the NER task was framed as a task where the model had to output a structured JSON answer containing identified entities, as a response to a prompt containing the input text and instructions. The LLM was instructed to write a JSON string containing the text spans identified as entities. The LLM's output was then parsed to extract valid JSON segments and exclude malformed output parts. The generated answer was limited to a length of 1024 tokens. The model was not asked to return the exact character offsets delimiting the spans, as preliminary tests with this method seemed unreliable. For this experiment, the chunks of texts were given at train time, untokenized, as input to the model along with corresponding entities. For evaluation, the chunks of texts were also passed untokenized to the model along with instructions, in a prompt. The fine-tuning was conducted using rank-16 LoRA adapters, with the 4-bit quantized version of the model.

6 Results

This section presents the averaged results for the various NER methods we evaluated.

Results of the QuickUMLS experiments can be found in table Table 4. It can be observed that performances are higher for the "*Disease*" label than for the "*Autoantibody* label. This could be explained by the fact that for the recognition of "*Disease*" we relied on the richness of UMLS vocabularies: our disease dictionary is 636,989 rows, whereas our manual dictionary of "*Autoantibody*" is smaller (7,464 rows) and handcrafted, and thus probably incomplete, which would explain the lower recall for autoantibodies.

Entity type	Precision	Recall	F1 (±SD)
ALL	0.47	0.40	0.43 (±0.04)
Autoantibody	0.50	0.24	0.33 (±0.06)
Disease	0.46	0.53	0.49 (±0.06)

Table 4: Precision (P), recall (R), and F1-score (F1) of QuickUMLS for autoantibodies and diseases, using custom made dictionaries.

Table 5 reports the performance of all the different model and experimental setup we evaluated. Overall best F1-score is obtained with the fine-tuned version of MedGemma. We particularly observe that in terms of F1-score, QuickUMLS performance is not significantly lower than GLiNER and MedGemma if those are not fine-tuned. F1-score for entity type "*ALL*" is the average over 5 test sets of the micro F1-score computed by counting true positives, false positives, and false negatives of all entity types together.

Regarding the various GLiNER configurations, the fine-tuned largest version of GLiNER is overall performing slightly better than other versions. 2-shot GLiNER models perform overall worse than 0-shot, possibly indicating our entity labels were somewhat more confusing than helpful to the model without a sufficient number of examples. Top precision is achieved by GLiNER large 2-shot, whereas top recall and F1-score are achieved by GLiNER large fine-tuned (Table 5). All methods return higher precision than recall, and this seems to be especially the case in the 2-shot configurations. It should be possible to modulate this to some extent by acting on the GLiNER threshold parameter, as well as on the number and type of examples.

	Model	P	R	F1 (±SD)
	QuickUMLS	0.47	0.40	0.43 (±0.04)
0-shot	GLiNER small	0.72	0.26	0.39 (±0.05)
	GLiNER medium	0.67	0.33	0.45 (±0.04)
	GLiNER large	0.74	0.30	0.42 (±0.03)
	MedGemma	0.61	0.31	0.41 (±0.05)
2-shot	GLiNER small	0.76	0.20	0.31 (±0.05)
	GLiNER medium	0.67	0.26	0.37 (±0.04)
	GLiNER large	**0.80**	0.21	0.34 (±0.04)
	MedGemma	0.67	0.29	0.41 (±0.04)
fine-tuned	GLiNER small	0.65	0.51	0.58 (±0.05)
	GLiNER medium	0.67	0.52	0.58 (±0.03)
	GLiNER large	0.67	0.53	0.59 (±0.04)
	MedGemma	0.71	**0.62**	**0.66** (±0.03)

Table 5: Precision (P), recall (R), and F1-score (F1) of different models for all entity types together.

We also report in Table 6 per-label performances for some of the best-performing models: GLiNER large and MedGemma, both in 0-shot and fine-tuned configurations. Fine-tuning increases model F1-score for all entity types, though it sometimes slightly decreases precision: performance gains are mostly due to recall gains. Considerable performance imbalance can be observed between labels: for the fine-tuned models, the highest F1-score (0.79) is achieved by GLiNER large for the detection of "*Disease*", as per Table 6; and the lowest F1-score is achieved by MedGemma for the detection of "*Autoantibody location*" (0.33). In the case of "*Autoantibody location*", fine-tuning GLiNER large causes performance increase from an F1-score of 0.07 to 0.60; and fine-tuning MedGemma increases F1-score from 0.01 to 0.33. It is possible that some types of entities, such as "*Autoantibody location*" are not easily captured when the only information about them is their label, without examples. However, giving examples can lead to better performances, as illustrated by the performance differences between 0-shot and fine-tuned configurations.

7 Discussion

This work demonstrates the use of the AAbAAC corpus for autoimmunity information extraction through some basic NER experiments, including model fine-tuning. Indeed, models fine-tuned using the AAbAAc corpus perform better at detecting entities related to autoimmunity than the same models used out-of-the-box, or than a rule-based approach. However, so far, in the work presented here, no search for optimal fine-tuning parameters was conducted. The fine-tuned models' performances could probably be improved by performing a grid search or some other type of optimization prior to fine-tuning.

Model and Entity type	P	R	F1 (±SD)
GLiNER large 0-shot			
ALL	0.74	0.30	0.42 (±0.03)
Autoantibody	0.72	0.37	0.48 (±0.05)
Autoantibody location	0.10	0.06	0.07 (±0.05)
Autoantibody target	0.76	0.06	0.11 (±0.07)
Disease	0.82	0.48	0.60 (±0.06)
Symptom or clinical sign	0.56	0.13	0.21 (±0.06)
GLiNER large fine-tuned			
ALL	0.67	0.53	0.59 (±0.04)
Autoantibody	0.75	0.49	0.59 (±0.11)
Autoantibody location	0.47	0.88	**0.60** (±0.17)
Autoantibody target	0.43	0.28	0.34 (±0.07)
Disease	0.80	0.77	**0.79** (±0.02)
Symptom or clinical sign	0.52	0.39	**0.42** (±0.05)
MedGemma 0-shot			
ALL	0.61	0.31	0.41 (±0.05)
Autoantibody	0.65	0.45	0.53 (±0.06)
Autoantibody location	0.02	0.01	0.01 (±0.03)
Autoantibody target	0.42	0.07	0.11 (±0.02)
Disease	0.70	0.39	0.50 (±0.06)
Symptom or clinical sign	0.47	0.23	0.31 (±0.05)
MedGemma fine-tuned			
ALL	0.71	0.62	**0.66** (±0.03)
Autoantibody	0.79	0.70	**0.74** (±0.06)
Autoantibody location	0.35	0.35	0.33 (±0.18)
Autoantibody target	0.75	0.67	**0.70** (±0.04)
Disease	0.79	0.66	0.71 (±0.02)
Symptom or clinical sign	0.46	0.41	**0.42** (±0.04)

Table 6: Precision (P), recall (R), and F1-score (F1) per label for GLiNER large and MedGemma fine-tuned, both in 0-shot and fine-tuned configuration. Best F1-score per type of entity is reported in bold.

The presented AAbAAC corpus aims to be of use for complete information extraction pipelines for the study of autoimmunity. This includes both named entity recognition and extraction of relations between these entities, so the corpus also includes relation annotations. However, for now, no evaluation of the use of the corpus for information extraction was conducted. One future work may use the GLiNER2 (Zaratiana et al., 2025) architecture, which enables the joint extraction of both entities and relations simultaneously.

For GLiNER-based models, so far, only fine-tuning of the original generalist GLiNER family models was performed, however it might be of interest to attempt similar experiments with different GLiNER variants. For instance, versions dedicated to the biomedical field, such as *Ihor/gliner-biomed-large-v1.0* (Yazdani et al., 2025) are available. Attempting to fine-tune a domain-specific BERT model, such as PubMedBERT (Gu et al., 2022), for example, would also be a possible future work direction and an interesting comparison.

8 Conclusion

This work introduces AAbAAC, a manually annotated corpus of texts that was created for information extraction in the field of autoimmunity. AAbAAC includes annotations of entities and relationships that are relevant to the study of autoantibodies and autoimmune diseases. Preliminary NER experiments on this new corpus demonstrate that fine-tuning models with AAbAAC yields a performance increase over dictionary-based approaches as well as zero-shot or few-shot settings. Future work will focus on further tuning different models for this domain, for both tasks of NER and relation extraction.

9 Limitations

AAbAAC corpus is relatively small: 115 annotated texts of variable length, all of them being titles and abstracts of scientific papers published on PubMed. This is enough to deliver substantial performance gains on NER tasks for the entity types considered in this work; however, ultimately the knowledge that can be derived from this corpus is finite and may not reflect the entire existing spectrum of autoantibodies, autoimmune diseases, and the various appellations used in the literature to refer to these concepts, especially since many autoimmune diseases are rare (Miller, 2023). For some very rare, or unusually named autoantibodies not annotated in the AAbAAC corpus, detection improvements may be limited. In addition, since this corpus consists of texts drawn from the scientific literature, its impact on NER in clinical texts could be lower due to differences between written styles and possibly different naming conventions for autoantibodies.

Regarding relation annotations, some of the relation types in our scheme have almost never been instantiated in the corpus. For example, this is the case with *isCausedBy* (13 occurrences, see Table 3), a stronger relation than the more common *isAssociatedWith* (735 occurrences), or *isRefTo* (2 occurrences).

The annotation process took place using the Doccano tool, which does not allow to specify annotation attributes. Therefore, to annotate entities that are composed of several discontiguous segments of text, the *discontiguousEntity* relation is used. But this complexity is lost to an NER model that only considers entity annotations, and this may have caused some performance drop in the presented experiments.

Other models from the state of the art could have been interesting to experiment with as well, in particular BERT-type models fine-tuned for the biomedical literature, or non-proprietary LLM. Their inclusion could open perspectives to extend our study to consider the computational cost of various approaches and attempt to specialize or distill efficient but costly models.

Acknowledgments

The authors acknowledge the Filières de Santé Maladies Rares BRAIN-TEAM and Filnemus for funding. The authors thank B. Belloir and S. Bernichtein from BRAIN-TEAM and R. Soussi from Filnemus for their help. The authors highlight the contributions of the BRAIN-TEAM, Filnemus, FAI²R, Fimarad, Fimatho, FILFOIE, MHEMO, and DéfiScience networks. The authors also thank C. Lucano and C. Fabrizzi (ORPHANET) for their help and our research teams, HeKA and Pathophysiological basis of skeletal dysplasia, for their constant support.

References

Haseeb Ahsan. 2023. Origins and history of autoimmunity—A brief review. *Rheumatology & Autoimmunity*, 3(1):9–14.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue):D267.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Thuy Trang Dinh, Trang Phuong Vo-Chanh, Chau Nguyen, Viet Quoc Huynh, Nam Vo, and Hoang Duc Nguyen. 2022. Extract antibody and antigen names from biomedical literature. *BMC Bioinformatics*, 23:524.

A. W. Forrey, C. J. McDonald, G. DeMoor, S. M. Huff, D. Leavelle, D. Leland, T. Fiers, L. Charles, B. Griffin, F. Stalling, A. Tullis, K. Hutchins, and J. Baenziger. 1996. Logical observation identifier names and codes (LOINC) database: a public use set of codes and names for electronic reporting of clinical laboratory test results. *Clinical Chemistry*, 42(1):81–90.

Michael A. Gargano, Nicolas Matentzoglu, Ben Coleman, Eunice B. Addo-Lartey, Anna V. Anagnostopoulos, Joel Anderton, Paul Avillach, Anita M. Bagley, Eduard Bakštein, James P. Balhoff, Gareth Baynam, Susan M. Bello, Michael Berk, Holli Bertram, Somer Bishop, Hannah Blau, David F. Bodenstein, Pablo Botas, Kaan Boztug, and 157 others. 2024. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Research*, 52(D1):D1333–D1346.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Scott M. Hayter and Matthew C. Cook. 2012. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. *Autoimmunity Reviews*, 11(10):754–765.

Eduard Hovy and Julia Lavid. 2010. Towards a 'science' of corpus annotation: A new methodological challenge for corpus linguistics. *Open Journal of Modern Linguistics*, 9(3):206–214.

Yann Mathet, Antoine Widlöcher, and Jean-Philippe Métivier. 2015. The unified and holistic method gamma (γ) for inter-annotator agreement measure and alignment. *Computational Linguistics*, 41(3):437–479.

Frederick W. Miller. 2023. The increasing prevalence of autoimmunity and autoimmune diseases: An urgent call to action for improved understanding, diagnosis, treatment and prevention. *Current opinion in immunology*, 80(102266).

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with UMLS concepts. In *1st Conference on Automated Knowledge Base Construction, AKBC 2019, Amherst, MA, USA, May 20-22, 2019*.

Marco Naguib, Xavier Tannier, and Aurélie Névéol. 2024. Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Ana Rath, Annie Olry, Ferdinand Dhombres, Maja Miličić Brandt, Bruno Urbero, and Segolene Ayme. 2012. Representation of rare diseases in health information systems: The orphanet approach to serve a wide range of end users. *Human Mutation*, 33(5):803–808.

Adam Remaki, Jacques Ung, Pierre Pages, Perceval Wajsburt, Elise Liu, Guillaume Faure, Thomas Petit-Jean, Xavier Tannier, and Christel Gérardin. 2025. Improving Phenotyping of Patients With Immune-Mediated Inflammatory Diseases Through Automated Processing of Discharge Summaries: Multicenter Cohort Study. *JMIR Medical Informatics*, 13:e68704.

Eric Sayers. 2018. E-utilities quick start. In *Entrez® Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

Yinon Shapira, Nancy Agmon-Levin, and Yehuda Shoenfeld. 2010. Defining and analyzing geoepidemiology and human autoimmunity. *Journal of Autoimmunity*, 34(3):J168–J177.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Sandeep Subramanian and Madhavi K. Ganapathiraju. 2017. Antibody exchange: Information extraction of biological antibody donation and a web-portal to find donors and seekers. *Data*, 2(4):38.

Hadrien Titeux and Rachid Riad. 2021. pygamma-agreement: Gamma γ measure for inter/intra-annotator agreement in python. *Journal of Open Source Software*, 6(62):2989.

Randi Vita, Nina Blazeska, Daniel Marrama, IEDB Curation Team Members, Sebastian Duesing, Jason Bennett, Jason Greenbaum, Marcus De Almeida Mendes, Jarjapu Mahita, Daniel K. Wheeler, Jason R. Cantrell, James A. Overton, Darren A. Natale, Alessandro Sette, and Bjoern Peters. 2025. The immune epitope database (IEDB): 2024 update. *Nucleic Acids Research*, 53:D436–D443.

Amy Y. Wang, Jeremiah H. Sable, and Kent A. Spackman. 2002. The SNOMED clinical terms development process: refinement and analysis of content. *Proceedings of the AMIA Symposium*, pages 845–849.

Dan Wang, Liuhui Yang, Ping Zhang, Joshua LaBaer, Henning Hermjakob, Dong Li, and Xiaobo Yu. 2017. AAgAtlas 1.0: a human autoantigen database. *Nucleic Acids Research*, 45(D1):D769–D776.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: Schema-driven multi-task learning for structured information extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 130–140. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.