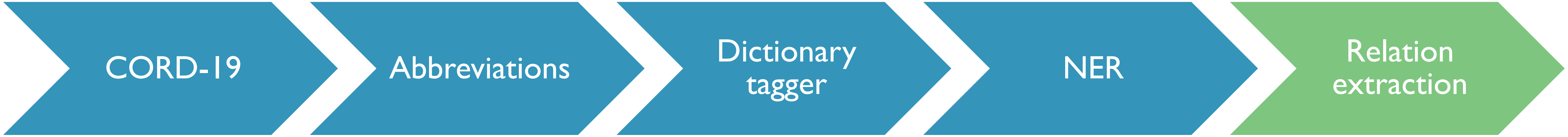
RELATION EXTRACTION USING BIOBERT

BY RASMUS LINDQVIST AND VIKTOR BARD

PROJECT OVERVIEW



CORD-19

Abbreviations

Dictionary tagger

NER

Relation extraction

OUR TASKS/GOAL

- Find relations between named entities in CORD-19 dataset from Kaggle using BioBERT
- Build BioBERT framework for relation extraction
- Evaluate BioBERT

BIOBERT INTRO

What is BioBERT ?

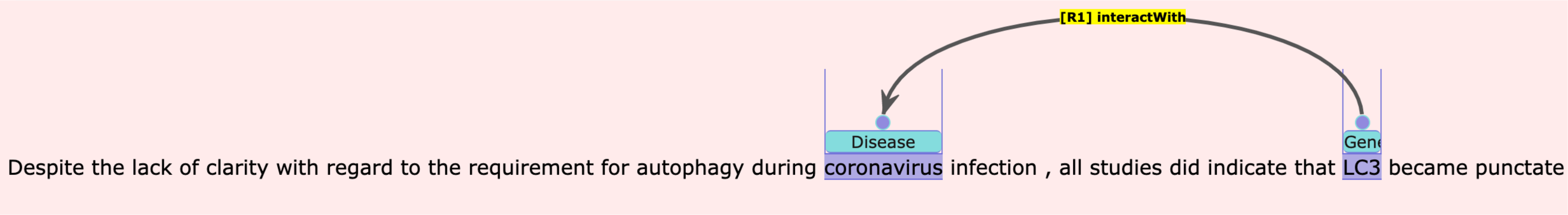
Pretrained model on biomedical data

Used for NER, RE and QA

Fine-tuning

Fine tune on task specific data-set

EXTRACTING RELATIONS BETWEEN ENTITIES



Despite the lack of clarity with regard to the requirement for autophagy during coronavirus infection , all studies did indicate that LC3 became punctate

”… authography during @disease infection , all studies did indicate that @gene became….”

TRAINING DATA

GAD

Genes and diseases

Semi-automatic annotation

EUADR

Drugs, disorders, genes

Manually annotated

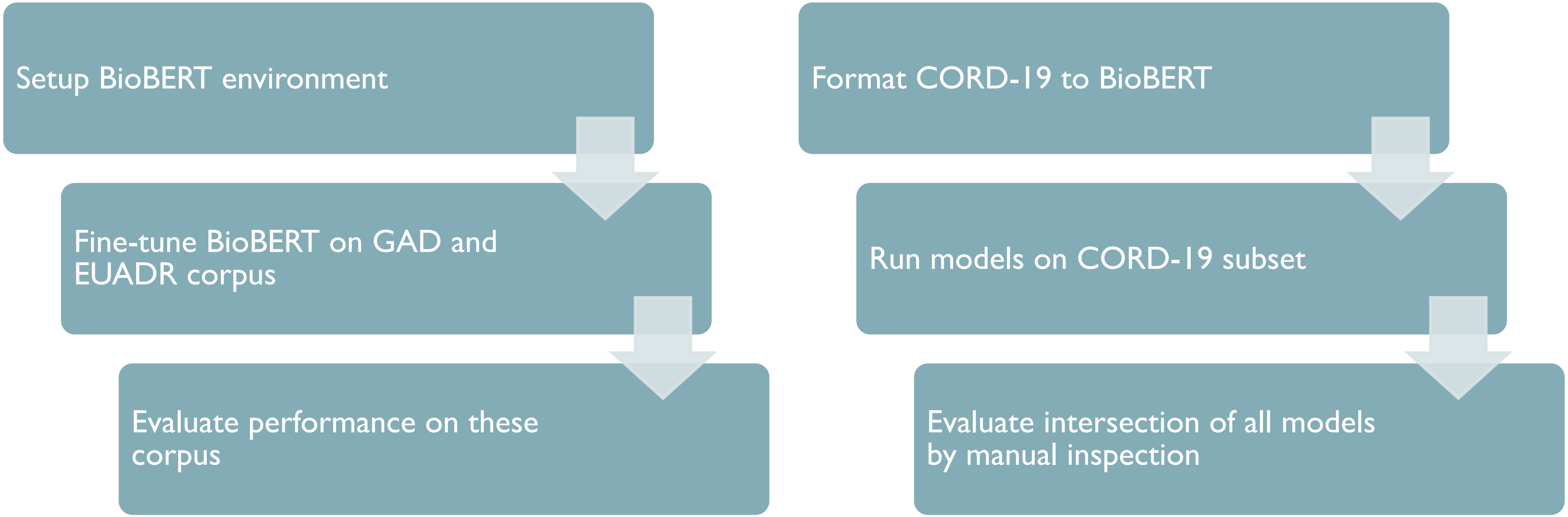
EVALUATION DATA

CORD-19

Current and past coronaviruses

Unlabeled data

METHOD



Setup BioBERT environment

Fine-tune BioBERT on GAD and EUADR corpus

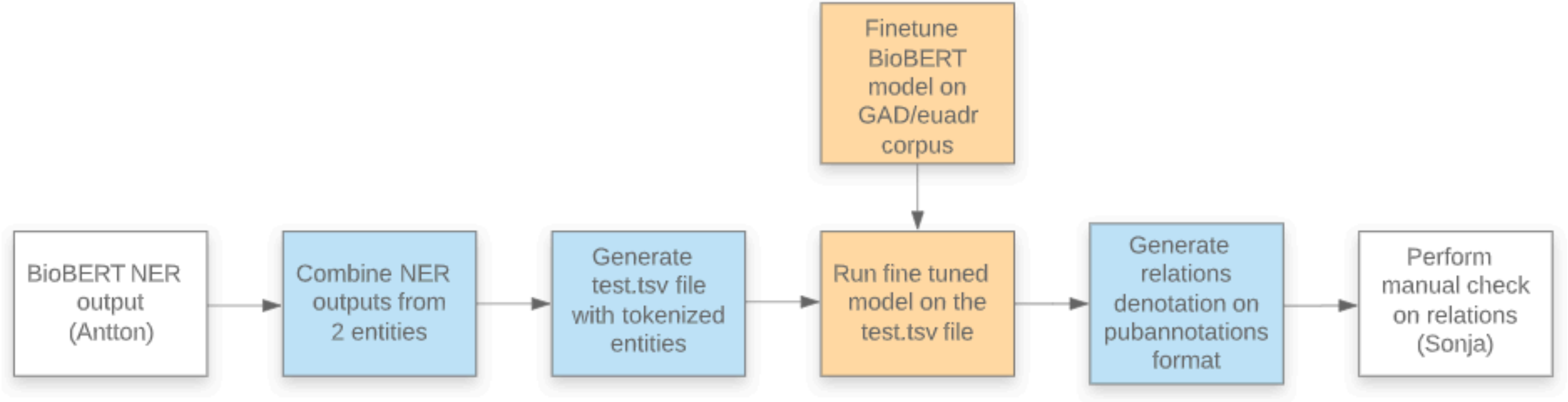
Evaluate performance on these corpus

Format CORD-19 to BioBERT

Run models on CORD-19 subset

Evaluate intersection of all models by manual inspection

FRAMEWORK



BioBERT NER output (Antton)

Combine NER outputs from 2 entities

Generate test.tsv file with tokenized entities

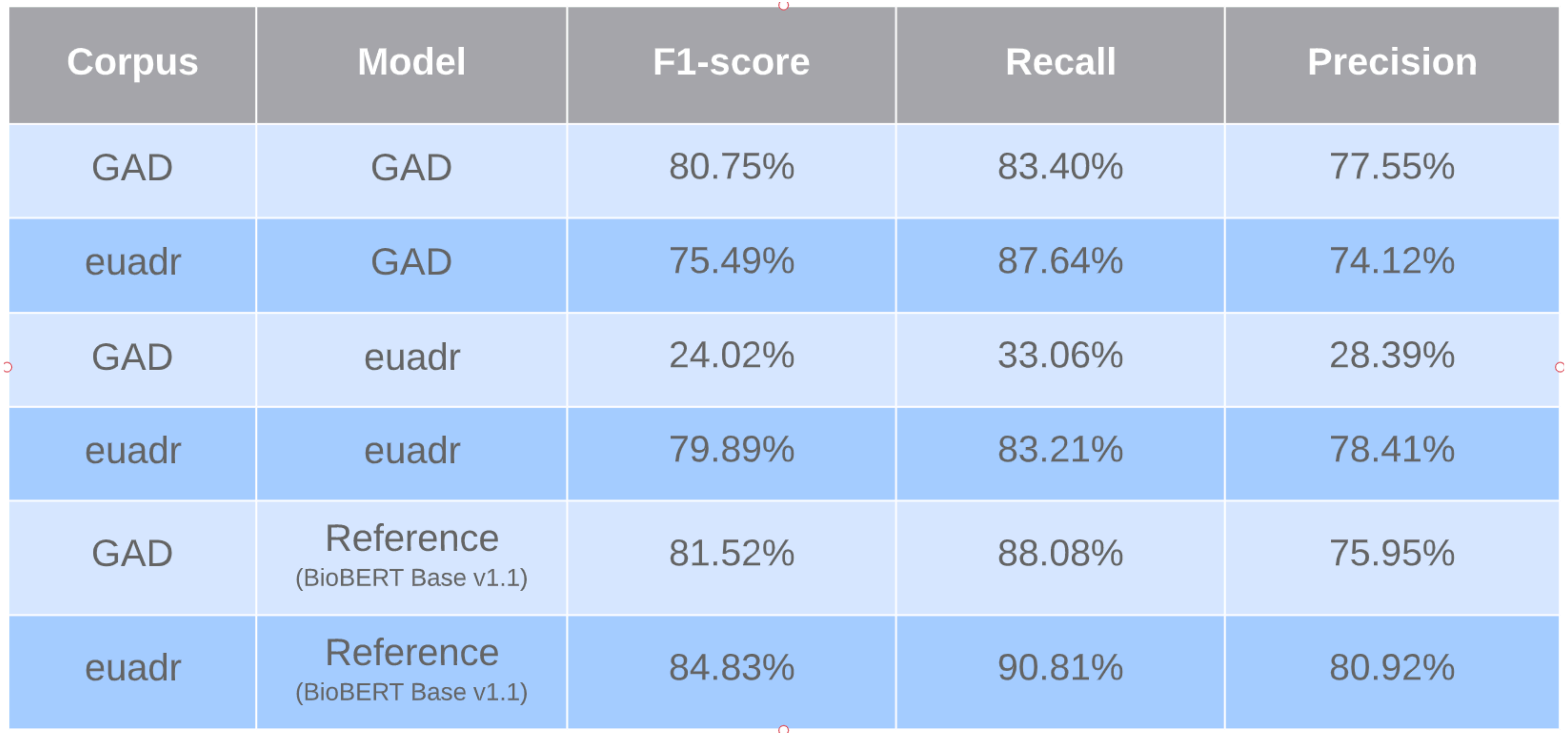
Finetune BioBERT model on GAD/euadr corpus

Run fine tuned model on the test.tsv file

Generate relations denotation on pubannotations format

Perform manual check on relations (Sonja)

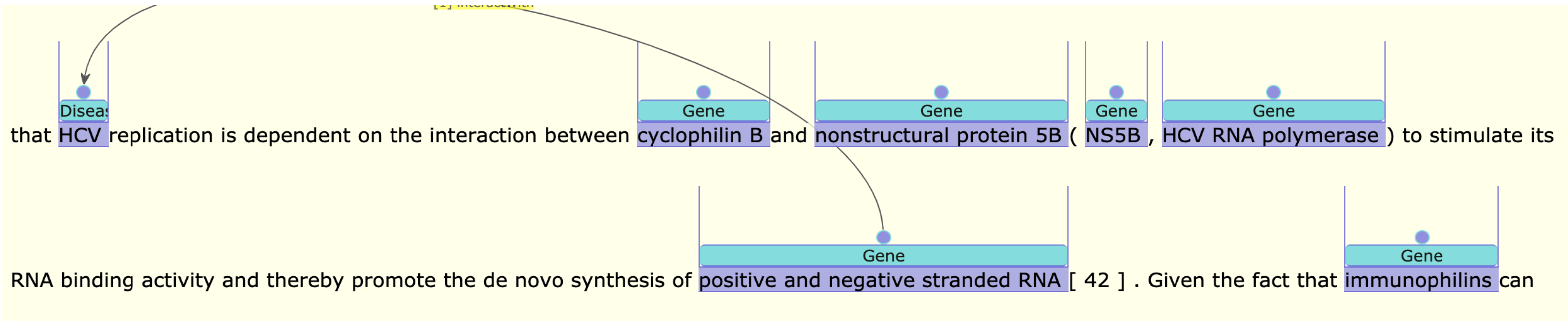
RESULTS



Corpus	Model	F1-score	Recall	Precision
GAD	GAD	80.75%	83.40%	77.55%
euadr	GAD	75.49%	87.64%	74.12%
GAD	euadr	24.02%	33.06%	28.39%
euadr	euadr	79.89%	83.21%	78.41%
GAD	Reference (BioBERT Base v1.1)	81.52%	88.08%	75.95%
euadr	Reference (BioBERT Base v1.1)	84.83%	90.81%	80.92%

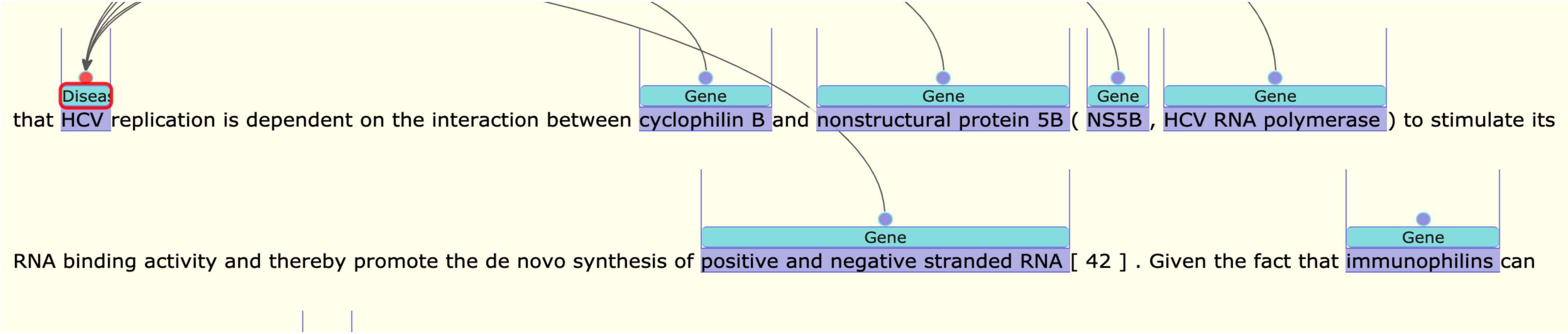
APPLIED TO COVID-19

Intersection



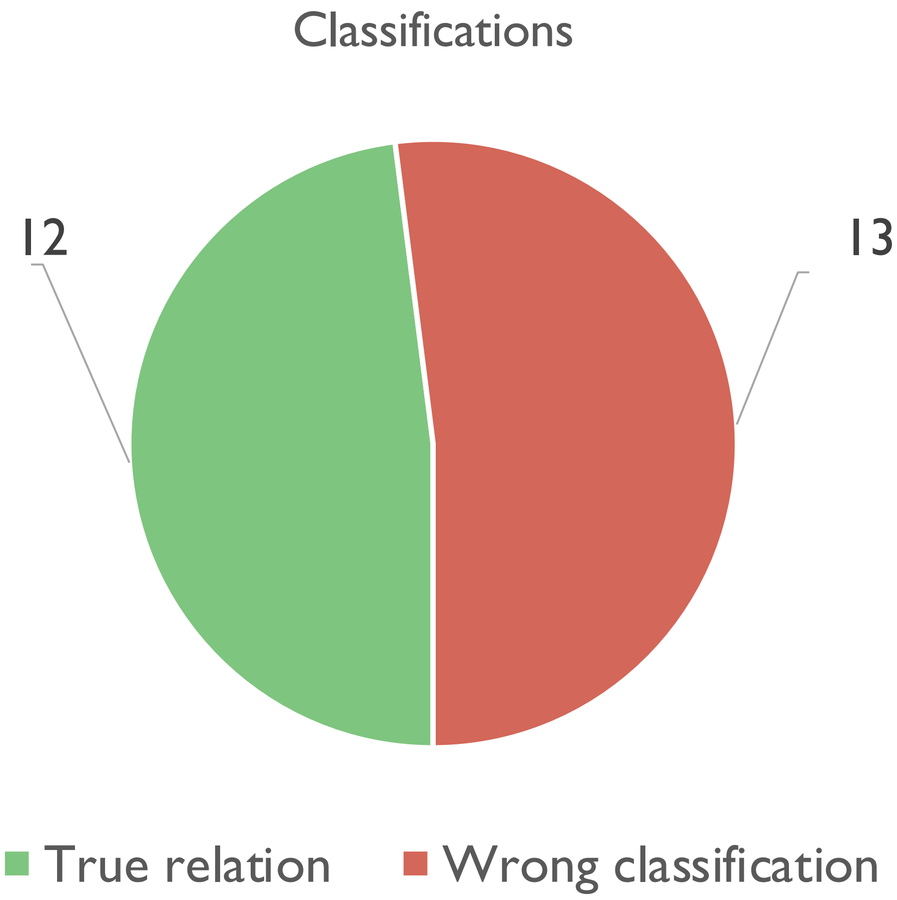
that HCV replication is dependent on the interaction between cyclophilin B and nonstructural protein 5B (NS5B , HCV RNA polymerase) to stimulate its RNA binding activity and thereby promote the de novo synthesis of positive and negative stranded RNA [42] . Given the fact that immunophilins can

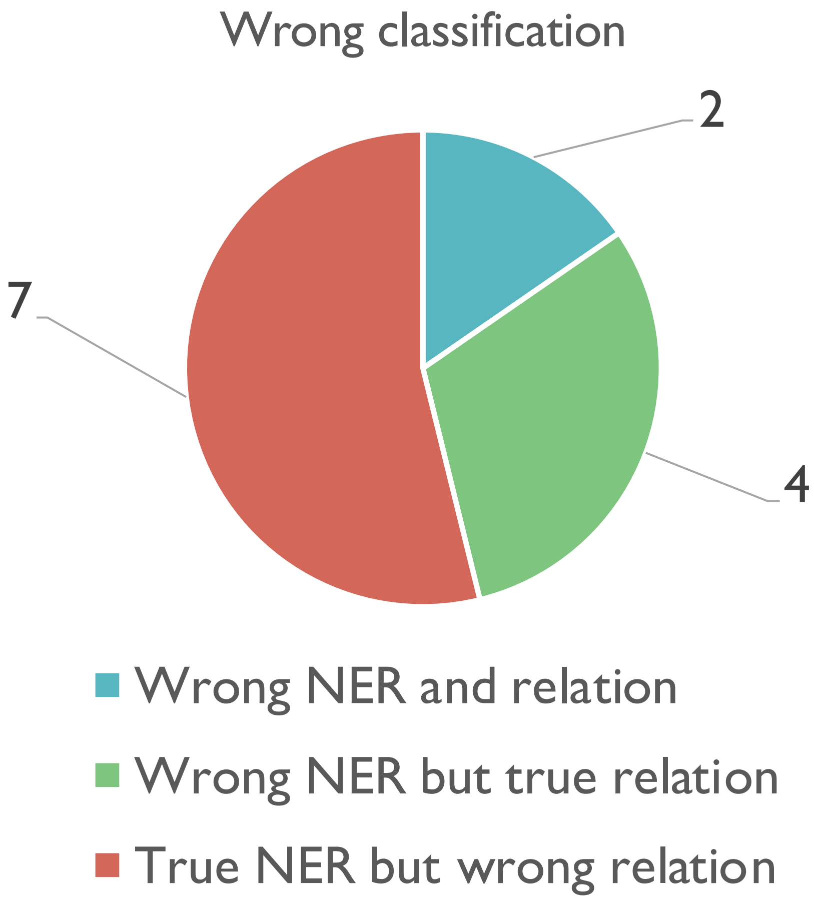
Union



that HCV replication is dependent on the interaction between cyclophilin B and nonstructural protein 5B (NS5B , HCV RNA polymerase) to stimulate its RNA binding activity and thereby promote the de novo synthesis of positive and negative stranded RNA [42] . Given the fact that immunophilins can

MANUAL INSPECTION RESULTS





CONCLUSION

- Little metric evaluation on CORD-19
- Manual inspection
- Comparable results to BioBERT article
- Computationally heavy

Questions? + Questions? = Questions?

QUESTIONS?