Zero-Shot Entailment Learning for Ontology-Based Biomedical Annotation Without Explicit Mentions

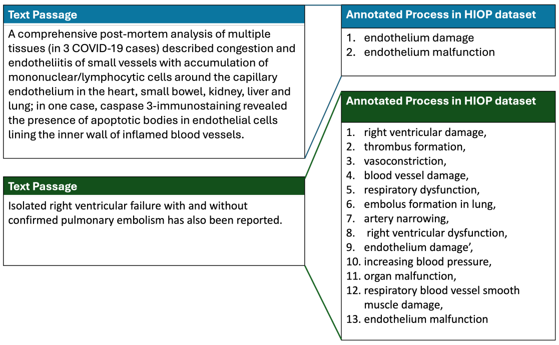
Rumana Ferdous Munne[1] **Noriki Nishida**[1] **Shanshan Liu**[1] **Narumi Tokunaga**[1]
Yuki Yamagata[2,3] **Kouji Kozaki**[1,4] **Yuji Matsumoto**[1]
[1]RIKEN AIP [2]RIKEN R-IH [3]RIKEN BRC [4]Osaka Electro-Communication University
{rumanaferdous.munne, noriki.nishida, shanshan.liu, narumi.tokunaga
yuki.yamagata, yuji.matsumoto}@riken.jp
kozaki@osakac.ac.jp

Abstract

Automatic biomedical annotation is essential for advancing medical research, diagnosis, and treatment. However, it presents significant challenges, especially when entities are not explicitly mentioned in the text, leading to difficulties in extraction of relevant information. These challenges are intensified by unclear terminology, implicit background knowledge, and the lack of labeled training data. Annotating with a specific ontology adds another layer of complexity, as it requires aligning text with a predefined set of concepts and relationships. Manual annotation is time-consuming and expensive, highlighting the need for automated systems to handle large volumes of biomedical data efficiently. In this paper, we propose an entailment-based zero-shot text classification approach to annotate biomedical text passages using the Homeostasis Imbalance Process (HoIP) ontology. Our method reformulates the annotation task as a multi-class, multi-label classification problem and uses natural language inference to classify text into related HoIP processes. Experimental results show promising performance, especially when processes are not explicitly mentioned, highlighting the effectiveness of our approach for ontological annotation of biomedical literature.

1 Introduction

Biomedical information extraction plays a critical role in advancing medical research, diagnosis, and treatment. Systematically analyzing extensive biomedical literature helps uncover insights into disease mechanisms, drug interactions, genetic associations, and treatment effectiveness. However, the task is challenging due to unclear terminology, implicit background knowledge, and diverse vocabularies. Manual annotation is time-consuming and labor-intensive, emphasizing the need for automatic extraction systems to efficiently manage large volumes of data.



Text Passage

A comprehensive post-mortem analysis of multiple tissues (in 3 COVID-19 cases) described congestion and endotheliitis of small vessels with accumulation of mononuclear/lymphocytic cells around the capillary endothelium in the heart, small bowel, kidney, liver and lung; in one case, caspase 3-immunostaining revealed the presence of apoptotic bodies in endothelial cells lining the inner wall of inflamed blood vessels.

Annotated Process in HIOP dataset

1. endothelium damage
2. endothelium malfunction

Text Passage

Isolated right ventricular failure with and without confirmed pulmonary embolism has also been reported.

Annotated Process in HIOP dataset

1. right ventricular damage,
2. thrombus formation,
3. vasoconstriction,
4. blood vessel damage,
5. respiratory dysfunction,
6. embolus formation in lung,
7. artery narrowing,
8. right ventricular dysfunction,
9. endothelium damage',
10. increasing blood pressure,
11. organ malfunction,
12. respiratory blood vessel smooth muscle damage,
13. endothelium malfunction

Figure 1: HoIP dataset example.

Biomedical annotation is the process of labeling or tagging specific terms or phrases in biomedical texts with predefined categories or classes. These entities can include diseases, genes, proteins, chemicals, drugs, biological processes, and other relevant biomedical concepts. In this paper, we aim to annotate biomedical articles from PubMed with HoIP processes. Homeostasis imbalance process ontology (HoIP) organizes a wide range of terms related to homeostasis imbalance courses and processes. It focuses on the course of COVID-19 infectious processes and cellular senescence (Yamagata et al., 2024). By annotating PubMed texts with HoIP process, researchers can better understand the mechanisms of COVID-19 diseases and how they progress. This helps in developing more effective treatments and diagnostic methods. Figure 1 shows examples from HoIP[1] dataset. Each text passage from the HoIP dataset is manually annotated by biomedical experts with a set of HoIP processes, though these processes do not always explicitly appear in the text. Shorter passages often have multiple annotations, while longer passages may have fewer. This demonstrates the selective annotation approach, where only the most relevant processes are captured.

Traditional methods for automatic biomedical entity use rule-based systems and dictionaries de-

[1] https://github.com/norikinishida/HOIP-dataset

Proceedings of the 31st International Conference on Computational Linguistics, pages 8148–8159
January 19–24, 2025. ©2025 Association for Computational Linguistics

rived from specific ontologies, relying on predefined term lists and patterns. While effective to some extent, they struggle with language variations, context, and require frequent manual updates. These methods assume that entities are explicitly mentioned, and such mentions provide clear features for extracting information. However, in real-world scenarios, entities often appear implicitly, which poses significant challenges. Machine learning methods usually depend on large amounts of annotated data, which can be challenging and expensive to acquire due to the manual effort and expertise needed for precise labeling. This data dependency can also limit their ability to handle new or unexpected variations, especially when annotated data is scarce. As a result, zero-shot learning is gaining popularity for its ability to perform tasks without needing large annotated datasets.

In this paper, we propose text entailment-based zero-shot text classification methods. We view HoIP processes as classes and reframe the annotation task as a multi-class and multi-label classification problem. Since each passage may contain multiple HoIP processes and different passages might share similar associated processes, we use zero-shot text classification to predict potential classes/processes for each passage. We also explored a predictive method based on ontology mapping. In this approach, we align MeSH (Medical Subject Headings) with the HoIP ontology. Predicting HoIP processes directly from text is difficult due to implicit mentions, limited data, and complex language, so we use a mapping-based method that predicts MeSH headings and infers HoIP processes through established mappings. Both approaches show promising results, particularly given the challenges of the dataset. The entailment-based zero-shot classification proves to be the most effective for tackling this type of complex task. However, the mapping based predictive method serves as a valuable complementary model, further enhancing overall performance when combined with zero-shot predictions. The major contributions and findings of the paper are as follows:

- We created a biomedical annotator system that doesn't rely on explicit mentions. It automatically tags biomedical text with specific HoIP ontology terms, even when the knowledge is described indirectly in the text.
- We introduce two novel approaches: entailment-based zero-shot text classification (ZPA) and ontology mapping-based predictive model (MPA) for annotating biomedical text passages with HoIP processes, without relying on large labeled datasets.
- Textual entailment-based zero-shot text classification, treats the annotation task as a multi-class and multi-label classification problem and predict potential HoIP process candidate for a given text sample.
- Mapping-based predictive model (MPA) leverages ontology alignment by mapping MeSH headings to the HoIP ontology, enabling the inference of HoIP processes from text passages.
- Both methods showed promising results in correctly identifying HoIP processes despite the challenges posed by the dataset. We observed that zero-shot text classification performed better than the MeSH predictive model. However, the highest annotation coverage can be achieved through combining predictions from both models.

2 Homeostasis Imbalance Process Ontology (HoIP)

HoIP Ontology The Homeostasis Imbalance Process Ontology (HoIP) systematically classifies a wide range of processes triggered by homeostatic imbalances. It primarily focuses on cellular senescence and the infectious processes related to COVID-19. The HoIP ontology is manually annotated with COVID-19-related articles from PubMed, specifically focusing on COVID-19 infectious processes. These COVID-19-specific processes are used as our primary dataset (El Khettari et al., 2024) for this study.

HoIP Dataset The primary objective of this study is to (semi-)automatically annotate biomedical articles with entities from a specific ontology. We used the HoIP dataset, which is derived from the HoIP ontology. The dataset comprises passages from PubMed articles discussing biomedical processes related to COVID-19, each passage describing at least two specific processes. These processes are manually annotated into triples head entity, relation, tail entity, detailing relationships based on the HoIP ontology. The dataset was then split into training, development, and test sets, ensuring that passages

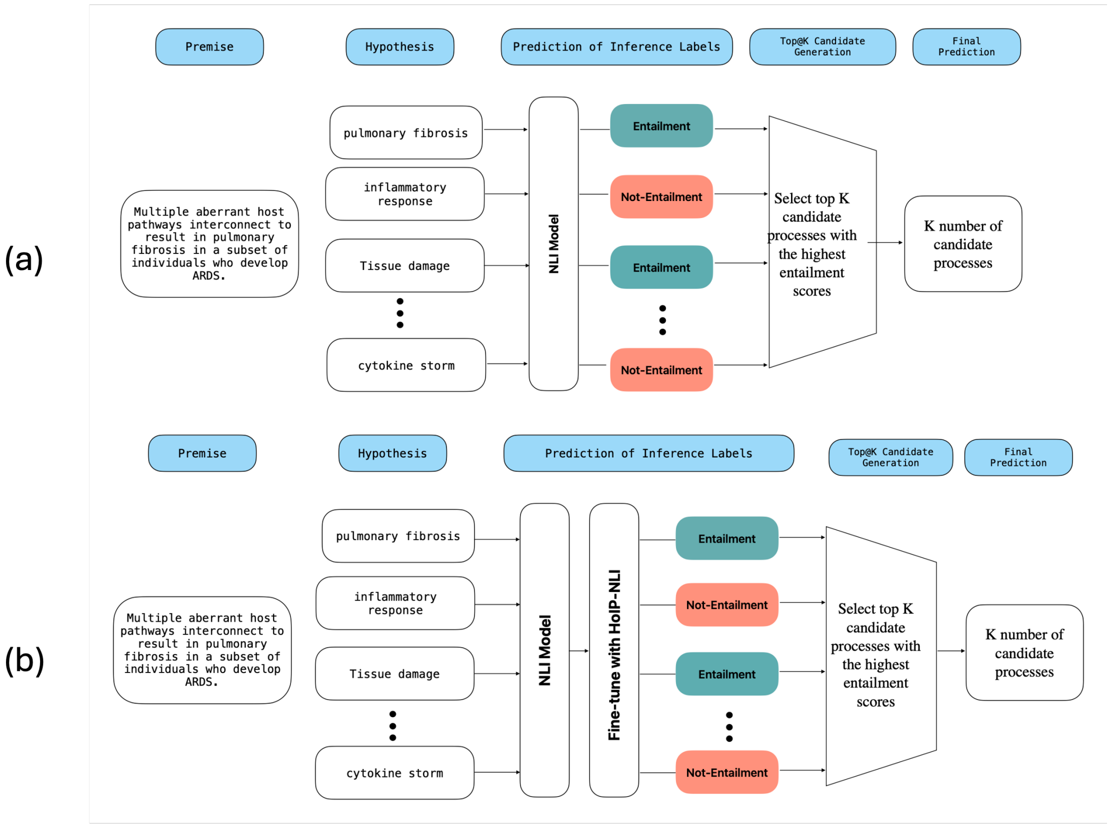

(a)
Premise
Hypothesis
Prediction of Inference Labels
Top@K Candidate Generation
Final Prediction
Multiple aberrant host pathways interconnect to result in pulmonary fibrosis in a subset of individuals who develop ARDS.
pulmonary fibrosis
inflammatory response
Tissue damage
cytokine storm
NLI Model
Entailment
Not-Entailment
Entailment
Not-Entailment
Select top K candidate processes with the highest entailment scores
K number of candidate processes
(b)
Premise
Hypothesis
Prediction of Inference Labels
Top@K Candidate Generation
Final Prediction
Multiple aberrant host pathways interconnect to result in pulmonary fibrosis in a subset of individuals who develop ARDS.
pulmonary fibrosis
inflammatory response
Tissue damage
cytokine storm
NLI Model
Fine-tune with HoIP-NLI
Entailment
Not-Entailment
Entailment
Not-Entailment
Select top K candidate processes with the highest entailment scores
K number of candidate processes

Figure 2: (a) Entailment based ZPA model for label fully unseen and (b) Ft-ZPA model fine-tuned with HoIP-NLI for label partially seen.

extracted from the same article were not scattered across different splits. The dataset statistics are shown in Table 1.

This HoIP dataset presents several unique challenges:

- **Absence of Explicit Mentions:** Entities and processes are often not clearly named in the text, with mentions being partial, implicit, or missing, which makes it hard to match them directly with ontology entries.
- **Selective Annotation:** Only the most significant HoIP processes are annotated, focusing on the most relevant information and avoiding over-annotation of all related processes.
- **Variable Annotation Scope:** Annotations range from single lines to entire paragraphs, reflecting the varying lengths and complexities of the text.
- **Overlapping Text:** Text passages can overlap but may have different entities, adding complexity to the annotation process.

The complex annotation process of the dataset makes it challenging for traditional supervised models to achieve accurate alignment with the HoIP ontology. The selective annotation of only significant processes further complicates this task. Therefore, we need an advanced method that can handle implicit information, manage overlapping contexts, and ensure precise alignment with the HoIP ontology. (see Appendix A for ontology)

	Train	**Dev**	**Test**
# passages	255	35	37
# entities	1988	143	211
# triples	1848	137	177
Avg. words per passage	75.5	70.4	61.8
Avg. entities per passage	7.8	4.1	5.7
Avg. triples per passage	7.2	3.9	4.8

Table 1: Dataset statistics for the HoIP dataset.

3 HoIP Process Identification

Biological process entities in the HoIP dataset are annotated based on their presence in the text, either explicitly or implicitly, without marking the exact phrase in the passage. This approach reflects real-world conditions but makes automatic process identification difficult. Traditional Named Entity Recognition (NER) methods rely on explicit mention-text links for training, which are not available here. In the following sections, we will discuss our proposed methods for addressing this challenge.

3.1 Zero-shot Classification-Based Process Annotation (ZPA)

Zero-shot classification is becoming popular in biomedical information extraction due to its ability to handle limited annotated data. By leveraging knowledge from large general datasets, these models classify biomedical data without specific training. This is particularly useful for the HoIP ontology, where limited annotated data and implicit processes make traditional supervised methods inadequate. To convert HoIP process annotation into a classification task, we use the ZPA model, which applies zero-shot classification based on textual entailment. Figure 2. demonstrate the model architecture. In this model, the input text passage serves as the premise, HoIP processes act as hypotheses, and the zero-shot classifier identifies which processes are logically entailed by the text.

We simplify the traditional natural language inference (NLI) model by combining the "neutral" and "contradiction" categories into a single "not-entailment" class, effectively turning the task into a binary entailment problem. This approach helps the model better differentiate between entailed and non-entailed processes, improving precision and simplifying decision-making by using deep contextual understanding to map input text to relevant HoIP processes.

3.1.1 Converting labels into hypotheses.

The first step of the ZPA model is to convert the HoIP processes into hypotheses. To achieve this, we start by transforming the HoIP process label names into a format suitable for text entailment. We utilize two hypothesis templates: one that uses the direct format of the "HoIP process label" itself, and the other rephrases it into a more descriptive template, such as "This text is about <HoIP process label>." These templates help the model better understand the relationship between the input text passage (the premise) and each potential HoIP process (the hypothesis), allowing for a more accurate classification of the processes relevant to the passage. In our experiment, the first template performed the best.

3.1.2 Converting classification data into entailment data.

For a data split (train, dev and test), each input text, acting as the premise, has a positive hypothesis corresponding to the positive label, and all negative labels in the data split provide negative hypotheses. Note that unseen labels are not used as negative samples during training, so they are entirely zero-shot.

3.1.3 Entailment model learning.

We used three widely recognized state-of-the-art pretrained model for zero shot classification techniques: BART-Large-MNLI (Lewis et al., 2020), DeBERTa-Large-ZeroShot (He et al., 2021; Laurer et al., 2023), and RoBERTa-Large-ZeroShot (Laurer et al., 2023). Proposed model explores two setups namely label-fully-unseen (ZPA) and label-partially-seen (Ft-ZPA). For label-fully-unseen setup, we directly apply the pretrained entailment model on the test sets of zero-shot text classification task. For label-partially-seen setup fine-tuned the pretrained NLI model with our small-scale HoIP-NLI dataset.

Label-partially-seen: In zero-shot text classification, a common approach is to train a system on a subset of labels from a dataset and then evaluate it on the entire set of labels. This method is commonly applied to tasks such as topic categorization or emotion detection. The setup, referred to as 'Label-partially-seen', trains the model on a subset of labels, where some classes have very few examples, and others are entirely absent. Unlike traditional few-shot learning, which requires at least a few examples from each class, this approach challenges the model to generalize to both seen and unseen labels. During testing, each input passage (the "premise") is compared with all possible class labels (the "hypotheses"). When the model encounters new documents or previously unseen classes, it leverages pre-trained knowledge from models like BART-LARGE-MNLI, which are specifically designed for natural language inference, to generate accurate predictions.

Label-fully-unseen: This approach takes "zero-shot" to its most extreme form, where no annotated data is available for any labels. The idea is to develop a system using whatever methods are available and then test it on zero-shot text classifier datasets that cover completely new and open-ended aspects.

Fine tune with HoIP-NLI Since the HoIP dataset was not initially designed for an NLI task, we created the HoIP-NLI dataset to adapt it for this purpose. We converted the positive examples into a pairwise format, where each example includes the text input passage and the hypothesis, labeled with "entailment." For instances with multiple HoIP processes, we generated several positive pairwise examples, each corresponding to a different HoIP process. Additionally, for each positive example, we generated a random negative example by pairing the premise with an unrelated hypothesis, labeled as "not-entailment," to provide the model with a balanced set of entailed and non-entailed pairs for training. The process of constructing negative examples is discussed in section 5.2.1.

HoIP Process Prediction We have transformed the process prediction task into a classification problem, where each input passage is assigned one or more classes from a set of 360 processes in our dataset. Since a single passage can be annotated with multiple processes, and multiple passages can share the same process, this task becomes a multi-class, multi-label classification problem. To address this, the model predicts the top@k candidate processes for each passage, based on the scores from the zero-shot classifier.

3.2 Mapping-Based Predictive Process Annotation (MPA)

Annotating HoIP processes is challenging due to the lack of explicit mentions in the text and limited training data. To overcome these challenges, we propose a mapping-based predictive approach that utilizes semantic similarities and predictive modeling techniques. Our MPA model includes: (1) Mapping similarities between MeSH and HoIP terms, (2) Predicting MeSH headings for text passages, and (3) Mapping these headings to HoIP processes.

Semantic Mapping To calculate the similarity between MeSH headings and HoIP processes, we use the descriptive data from each ontology. MeSH scope notes provide definitions and context for each MeSH term, while HoIP process definitions explain the roles of processes within HoIP. We convert these descriptions into embeddings with PubMedBERT and compare them using cosine similarity to assess their semantic similarity. The model refines mappings by adjusting HoIP mappings based on annotated training data, creating a comprehensive list of potential HoIP processes for each MeSH heading. Given the conceptual differences, MPA model supports many-to-many relationships, allowing for flexible alignment between MeSH and HoIP terms.

MeSH Headings Prediction We have extended WellcomeBertMesh[2] model, originally designed for tagging MeSH headings in PubMed abstracts, to generate headings for our specific input passages. WellcomeBertMesh utilizes PubMedBert for semantic indexing and incorporates a multi-label attention head to focus on relevant tokens for each label and it is trained using binary cross-entropy loss. (Details in Appendix C)

HoIP Process Inference During testing, MeSH headings are generated for each text passage and matched with established HoIP mappings. This process helps predict relevant HoIP processes, even when they are not explicitly mentioned in the text.

3.3 Optimizing Predictions: Combine MPA and ZPA (ZMPA)

ZPA generates k candidate HoIP processes, while the MPA method predicts HoIP processes based on aligned MeSH Headings. By integrating these approaches into the ZMPA model, we achieve better recall, covering more entities per text passage. Although this combination may increase the number of candidates and affect precision, the improved recall ensures higher annotation coverage and represents a significant advancement in capturing relevant processes. The complete illustration of the ZMPA model architecture and case study can be found in Appendix B and Section 6.

4 Experiments

The experiments presented in this section are designed to evaluate the effectiveness of our ZPA and ZMPA methods on both completely unseen data and partially seen data. Additionally, we assess the effectiveness of the proposed models in addressing these challenges:

Q1. Can the textual entailment-based zero-shot model (ZPA) accurately annotate challenging biomedical data, particularly when entities are implicitly mentioned or selectively annotated, without any direct training?

[2] https://huggingface.co/Wellcome/WellcomeBertMesh

Q2. How well do our models perform in handling a multi-class, multi-label classification task where each input text can be associated with multiple labels from a broad set of process classes?

Q3. How effective are these models when limited training data is available? Does the incorporation of partially seen data enhance the performance of the system?

Implementations setup We evaluate our zero-shot process annotation (ZPA) model using three distinct pretrained language models: BART-Large-MNLI , DeBERTa-v3-Large-ZeroShot-V2, and RoBERTa-Large-ZeroShot-V2C. These models are renowned for their expertise in textual entailment, natural language understanding, and zero-shot classification capabilities. For Ft-ZPA model we fine-tune these pretrained models using the HoIP-NLI dataset. All training sessions utilized the Adam optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. The fine-tuning process was executed using PyTorch and the Hugging Face Transformers library.

Evaluation Protocol In our performance evaluation, we used standard metrics: F1 score, accuracy, precision, and recall. Metrics were averaged across all samples to provide an overall measure of performance. To determine the final set of process candidates, we select top@k processes from the model's output, with k evaluated at 5, 10, and 20. This choice is based on the dataset, where the maximum number of labels per text is 19 and the average is 8 for training and 5 for test set. Selecting these k values ensures coverage of a broad range of potential processes, ensuring both comprehensive coverage and practical alignment with observed process frequencies.

5 Results and Discussion

5.1 Label-fully-unseen evaluation

In this setup, we are not conducting any training. Instead, we use a similarity-based model and a generative zero-shot model as baselines, alongside our entailment-based zero-shot model. We also report results from combining the MPA model with our proposed zero-shot classifier (ZMPA) for fully unseen settings.

Baselines

Similarity mapping Annotation (SMA) We use semantic similarity mapping to link HoIP processes with text passages by leveraging vector representations from advanced models like BioBERT, SciBERT, and PubMedBERT. Each process and passage is embedded into a high-dimensional vector space, capturing their meanings. We then compute cosine similarity between these vectors to assess how closely related each process is to the passage. No training is envolved in the process.

Generative Prompt based Annotaion (GPA) This approach uses advanced language models like LLaMA and ChatGPT to link text with candidate processes by assessing their relevance to the information in the text. We create prompts to query the models about each candidate's relatedness with the text. The models respond to questions about the relevance of each candidate, and positive responses help identify which processes are most relevant.

Discussion As shown in Table 2, Semantic similarity mapping struggles with detecting entities without explicit mentions because they rely on general embeddings and don't capture contextual clues or relationships effectively.

The Generative Prompt-based Annotation (GPA) model faces challenges with consistency, as responses from models like LLaMA (Touvron et al., 2023) and ChatGPT (OpenAI, 2024) can vary. Ambiguity also leads to vague or incomplete answers, especially for complex or resembling processes. Additionally, these models lack domain-specific knowledge, resulting in inaccurate or irrelevant annotations, which results in low accuracy in our experiments.

ZPA employs an entailment-based zero-shot classifier and gives the best results compared to all the baselines due to its ability to leverage a pretrained model's understanding of language and logic. This method excels with limited annotated data by using natural language inference (NLI) to determine logical relationships between text pairs, thus generalizing well from fewer examples. Among the three pretrained models we used, **bart-large-mnli** achieved the best performance. This is because the model is well-suited for textual entailment tasks due to its pretraining on a diverse range of textual entailment data. In the extreme scenario "label-fully-unseen", it achieves recall rates between 26.25% and 54.10% by selecting the top 5 to 20 candidate processes. However, precision is affected by the dataset's class imbalance and multi-label nature. Incorporating the MPA model improves recall by nearly 5% but increases the number of candidates. Despite these challenges, the model performs comparably well in zero-shot scenarios, leveraging text entailment effectively. Among the three pretrained models we use **bart-large-mnli** achieves the best performance as the generative pretrained model is well suited for textual entailment task.

5.2 Label partially seen evaluation

In this setup, we have annotated data for partial labels (HoIP-NLI), which serves as our training set. For evaluation, we compare our proposed entailment-based zero-shot model, which is fine-tuned with HoIP-NLI, against a binary classifier fine-tuned on HoIP-NLI. Additionally, we report the results of combining the MPA model with our proposed zero-shot classifier in this setting.

Binary-BERT (supervised) We fine-tune BERT-based models like BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), and PubMed-BERT(Gu et al., 2020) for a binary classification task to determine whether a passage entails any of the 360 processes. During testing, the model selects the label with the highest probability in single-label scenarios, while in multi-label cases, it chooses all labels higher than an 'entailment' threshold.

Discussion Supervised binary classification tends to perform well with seen labels because it effectively learns patterns specific to these classes from the training data. However, its performance drops for unseen classes due to limited generalization beyond the training examples. Our main challenges are the lack of enough annotated data and missing explicit entity mentions making it hard for the model to identify new entities. As shown in Table 3, the Ft-binaryBERT model performs poorly. In contrast, both the ZPA and ZMPA models show significant improvements after fine-tuning with HoIP-NLI. Despite the small amount of training data, both models achieve over a 7% improvement in recall, with notable gains in precision and F1-score.

5.2.1 Negative sample Influence

In Figure 3, we show the influence of negative samples over the fine-tuned ZPA model. We applied three different negative sampling methods.

- NS-R (Random Sampling): For each "Entailment" instance, a "Not-entailment" example is generated by randomly selecting a process name from the dataset that does not appear in the positive set.
- NS-S (Similarity-Based Sampling): We use a pretrained BERT model to create embeddings for each process name and calculate their cosine similarity. For each "Entailment" instance, we find the top most similar process names. If the closest match isn't in the "Entailment" set, it is labeled as "Not-entailment."
- NS-D (Dissimilarity-Based Sampling): We generate negative examples by identifying the process name that is least similar to each "Entailment" instance, based on cosine similarity of their BERT embeddings.

NS-R works better for our model as shows in Figure 3. because it introduces diversity and reduces bias, helping the model generalize more effectively. NS-S can lead to over-fitting, while NS-D samples might be too far from the true class, making it harder for the model to learn useful distinctions. For our model, a 1:1 positive-to-negative sample ratio works best.

5.3 Ablation study

In this section, we conduct an ablation study to assess the impact of different model components. We evaluate the Mapping-based model (MPA), the zero-shot classification model (ZPA), the hybrid model ZMPA, and their fine-tuned versions (Ft-ZPA and Ft-ZMPA). Table 4 shows that MPA achieves a recall of 23.07% and precision of 28.65%, indicating limited coverage. ZPA improves recall to 54.01% but reduces precision to 26.00%, with an F1-score of 31.28%. Fine-tuning ZPA (Ft-ZPA) increases recall to 61.08% and slightly improves precision to 26.24%, resulting in a better balance with an F1-score of 33.53%. ZMPA, combining MPA and ZPA, raises recall to 64.14% but lowers precision to 23.30%, achieving an F1-score of 31.45%. The fine-tuned hybrid model, Ft-ZMPA, has the highest recall at 71.39%, though precision is 23.87%, with an F1-score of 32.50%. The study reveals that while Ft-ZPA balances recall and precision, Ft-ZMPA excels in maximizing recall. For tasks requiring extensive annotation coverage, increased recall across models ensures nearly all relevant processes are captured, which is crucial for real-world biomedical applications. Figure 4 illustrates how each model com-

Model	Pre-train Model	Top@5			Top@10			Top@20		
		Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score
SMA	BioBERT	3.89	4.89	3.81	5.78	4.45	4.55	8.97	3.45	4.56
	PubMedBERT	6.53	7.56	5.89	8.54	6.22	6.17	12.82	4.56	6.13
	SciBERT	6.45	7.56	6.29	12.71	8.22	9.05	16.77	5.89	8.09
GPA	GPT	10.02	26.05	14.47	-	-	-	-	-	-
	Llama	12.51	15.41	13.81	-	-	-	-	-	-
ZPA	roberta-large-zeroshot-v2c	23.01	31.60	26.76	28.76	28.13	28.01	37.34	19.78	25.89
	deberta-v3-large-zeroshot-v2	22.36	30.39	25.38	35.35	29.54	29.54	52.18	24.36	30.70
	bart-large-mnli	26.25	**35.68**	26.83	40.55	**32.48**	31.65	54.01	**26.00**	31.28
ZMPA	roberta-large-zeroshot-v2	42.04	25.77	31.34	43.61	22.35	29.10	51.83	16.94	25.50
	deberta-v3-large-zeroshot-v2	40.72	26.98	31.77	50.21	24.84	31.89	60.87	21.63	30.21
	bart-large-mnli	**44.61**	30.51	**32.45**	**54.89**	28.10	**33.84**	**64.14**	23.30	**31.45**

Table 2: Evaluation in Label Fully Unseen setting

Model	Fine-tune Model	Top@5			Top@10			Top@20		
		Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score
Ft-binary BERT	bert-base-uncased + HoIP-NLI	9.34	19.11	12.22	14.24	20.15	16.85	20.34	16.35	17.78
Ft-ZPA	roberta-large-zeroshot-v2-HoIP-NLI	26.01	43.10	32.44	31.16	28.82	29.94	48.56	21.48	29.72
	deberta-v3-large-zeroshot-v2-HoIP-NLI	24.01	42.21	30.60	38.72	**39.83**	36.49	55.35	25.01	32.78
	bart-large-mnli-HoIP-NLI	31.73	**48.89**	34.46	47.41	38.41	38.63	61.08	**26.24**	**33.53**
Ft-ZMPA	roberta-large-zeroshot-v2-HoIP-NLI	47.43	34.55	36.50	50.93	27.88	36.03	58.27	18.94	28.58
	deberta-v3-large-zeroshot-v2-HoIP-NLI	44.32	33.03	36.15	56.65	30.54	39.69	65.21	21.37	32.19
	bart-large-mnli-HoIP-NLI	**55.68**	39.42	**37.81**	**68.15**	34.56	**40.02**	**71.39**	23.87	32.50

Table 3: Evaluation in Label Partially Seen setting

ponent and the number of candidates contribute to continuous improvements in recall.

5.4 Robustness of Entailment-Based Zero-Shot Performance

To evaluate the robustness of our entailment-based zero-shot learning (ZSL) model, we assessed its performance on a different dataset, demonstrating the model's generalizability beyond the initial problem. Specifically, we used the Medical-Abstracts dataset from Kaggle, as proposed by Schopf et al. (2022) (Schopf et al., 2022), which consists of 28,880 abstracts across five patient condition classes. The dataset is divided into 14,438 training abstracts and 2,888 test abstracts.

The goal of the task is to classify these abstracts into the five condition classes. Table 5 presents a comparative analysis of F1-scores between Lbl2TransformerVec and ZPA model pretrained in DeBarta. Our model shows comparable results to the existing methods. However, in the "partially seen label" setting, we significantly enhance performance by improving nearly 10% through fine tuning on a small NLI dataset that we created from the training split. This fine tuning step allowed the model to better handle unseen labels, further demonstrating the effectiveness of our approach in real-world scenarios.

Settings	**Model**	**F1-score**
Similarity	Lbl2TransformerVec*	55.84
Label Fully Unseen	Zero-shot Entailment*	57.88
	ZPA	57.18
Label Partially Seen	Ft-ZPA	**67.34**

Table 5: Comparison of F1-scores across different models and settings.*Results from(Schopf et al., 2022)

6 Case Study

In biomedical research, understanding the underlying mechanisms of diseases is crucial for effective diagnosis and treatment. The Homeostasis Imbalance Process (HoIP) ontology serves as a pivotal tool, categorizing diverse processes implicated in disease progression. Annotating PubMed text with HoIP processes offers valuable insights into the underlying molecular pathways and disease dynamics. Our dataset comprises segments from PubMed texts alongside their corresponding annotated HoIP processes. For instance, the text "Isolated right ventricular failure with and without confirmed pulmonary embolism has also been reported" is associated with annotated HoIP processes like "embolus formation in lung," "vasoconstriction," and "thrombus formation." This example clearly illustrates that the HoIP processes are not explicitly referenced within the text. We utilize two annotation methods for predicting HoIP processes from the text. Our first method Predictive Model from MeSH relies on semantic understanding, yielding

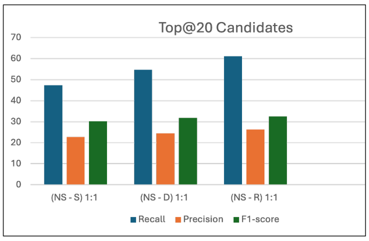


Figure 3: Influence of Negative Samples on Performance

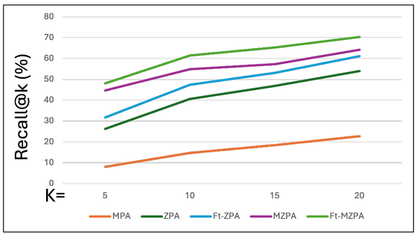


Figure 4: Recall@K results for each model component ;K=[5,10,15,20]

Model	R@20	P@20	F1-score
MPA	23.07	**28.65**	21.52
ZPA	54.01	26.00	31.28
Ft-ZPA	61.08	26.24	**33.53**
ZMPA	64.14	23.30	31.45
Ft-ZMPA	**71.39**	23.87	32.50

Table 4: Ablation Study. The results for the top 20 candidates.

predictions such as "vasoconstriction," "increasing blood pressure," and "detaching of blood clot." Meanwhile, our second method zero-shot text classification leverages advanced natural language inference techniques, resulting in predictions such as "right ventricular damage," "decreasing cardiac output," and "embolus formation in lung." By combining the results from both methods, we generate a comprehensive prediction of HoIP processes for the text, providing a robust understanding of the molecular mechanisms at play. This integrated approach enhances our ability to discern disease pathways and inform biomedical research and clinical practice effectively.

7 Related Work

Recent advances in biomedical text annotation leverage supervised models, pretrained language models, and Large Language Models (LLMs). Supervised methods, such as the BiLSTM-CRF model (Gong et al., 2021) and the hybrid approach of Li et al. (2020), excel in process classification. Pretrained models like BERT (Devlin et al., 2019) and BioBERT (Lee et al., 2020) enhance performance by utilizing extensive biomedical corpora and fine-tuning on specific datasets. Traditionally, these studies assume biomedical entities are explicitly mentioned multiple times and rely on these mentions for features and struggle when mentions are implicit. Perera et al. (2015) introduces one of the few methods for implicit entity recognition in clinical documents, proposing an unsupervised approach that leverages knowledge base definitions. Only a few studies focus on multi-label classification task due to its complexity. Rios and Kavuluru (2018) use label embedding to attend the text representation in the developing of a multi-label classifier. Zero shot classification based methods are gaining popularity in annotating biomedical documents. Yin et al. (2019) proposed to treat zero-shot text classification as a textual entailment problem, while Gera et al. (2022) tackled the task with a self-training approach. Koutsomitropoulos (2021) uses zero-shot learning to validate ontology-based annotations in biomedical texts. Pàmies et al. (2023) enhances zero-shot text classification through weak supervision and textual entailment techniques, while Košprdić et al. (2024) generalizes across unseen entities in zero-shot and few-shot settings. GPT models have gained popularity for NLP tasks. ChatGPT, with specific prompt design, has been applied in zero-shot clinical NER (Hu et al., 2023). Although these models have made significant progress in biomedical annotation, they fall short in addressing challenges such as limited labeled data, implicit entity recognition, out-of-vocabulary concepts, and the complexities of multi-label classification. Annotating biomedical text using the HOIP process poses all of these challenges. Our proposed methods show promising performance in overcoming these limitations.

8 Conclusion

In conclusion, our proposed system provides an efficient solution for automating biomedical text annotation by addressing challenges like implicit entities and limited labeled data. Using an entailment-based zero-shot approach, we reformulate the annotation task into a multi-class, multi-label classification problem, enabling accurate predictions of relevant processes from unstructured biomedical texts. The integration of ontology mapping boosts flexibility and coverage, making the system scalable and efficient for large biomedical datasets. This approach has the potential to significantly advance medical research, diagnosis, and treatment by streamlining and improving annotation processes.

Limitations

Although our proposed annotation method for biomedical documents without relying on mentions has shown promising results, several challenges remain. The task is inherently complex due to its multi-class, multi-label nature, which requires precise identification of all relevant labels while minimizing false positives. Additionally, while our model leverages textual entailment, this can sometimes lead to over-generalization or overlook subtle details. Additionally, due to the imbalanced class distribution and variability in label counts, we use a top@k approach for predictions. This method may not fully address the challenges of managing candidate selection. Future work should explore developing adaptive thresholds to enhance prediction accuracy and better handle the variability in label distribution. Moreover, the potential to enhance prediction accuracy by incorporating additional metadata—such as document structure, context, or other relevant information has not yet been fully explored in our current model. Leveraging such metadata could provide more detailed insights and improve the overall performance of the annotation system. Therefore, future work should consider integrating these additional sources of information to address the existing limitations and refine the annotation process further.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, Minneapolis, MN, USA.

Oumaima El Khettari, Noriki Nishida, Shanshan Liu, Rumana Ferdous Munne, Yuki Yamagata, Solen Quiniou, Samuel Chaffron, and Yuji Matsumoto. 2024. Mention-agnostic information extraction for ontological annotation of biomedical articles. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 457–473.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Conference on Empirical Methods in Natural Language Processing*.

Lejun Gong, Xingxing Zhang, Tianyin Chen, and Li Zhang. 2021. Recognition of disease genetic information from unstructured text data based on bilstm-crf for molecular mechanisms. *Security and Communication Networks*, 2021:1–8.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Yan Hu, Iqra Ameer, Xu Zuo, Xueqing Peng, Yujia Zhou, Zehan Li, Yiming Li, Jianfu Li, Xiaoqian Jiang, and Hua Xu. 2023. Zero-shot clinical entity recognition using chatgpt. *arXiv preprint arXiv:2303.16416*.

Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milošević. 2024. From zero to hero: Harnessing transformers for biomedical named entity recognition in zero-and few-shot contexts. *Artificial Intelligence in Medicine*, page 102970.

Dimitrios Koutsomitropoulos. 2021. Validating ontology-based annotations of biomedical resources using zero-shot learning. In *The 12th International Conference on Computational Systems-Biology and Bioinformatics*, pages 37–43.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2023. Building Efficient Universal Classifiers with Natural Language Inference. *arXiv preprint*. ArXiv:2312.17543 [cs].

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiangyang Li, Huan Zhang, and Xiao-Hua Zhou. 2020. Chinese clinical named entity recognition with variant neural structures based on bert methods. *Journal of biomedical informatics*, 107:103422.

Rumana Ferdous Munne and Ryutaro Ichise. 2023. Entity alignment via summary and attribute embeddings. *Logic Journal of the IGPL*, 31(2):314–324.

OpenAI. 2024. Chatgpt. `https://www.openai.com/chatgpt`. Accessed: September 17, 2024.

Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-Silva, César Parra-Rojas, Aitor González-Agirre, Francesco Alessandro Massucci, and Marta Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–296.

Sujan Perera, Pablo Mendes, Amit Sheth, Krishnaprasad Thirunarayan, Adarsh Alex, Christopher Heid, and Greg Mott. 2015. Implicit entity recognition in clinical documents. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 228–238.

Md Mostafizur Rahman, Daisuke Kikuta, Yu Hirate, and Toyotaro Suzumura. 2024. Graph-based audience expansion model for marketing campaigns. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2970–2975.

Md Mostafizur Rahman, Atsuhiro Takasu, and Gianluca Demartini. 2020. Representation learning for entity type ranking. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 2049–2056.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 3132. NIH Public Access.

Tim Schopf, Daniel Braun, and Florian Matthes. 2022. Evaluating unsupervised text classification: zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, pages 6–15.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuki Yamagata, Tatsuya Kushida, Shuichi Onami, and Hiroshi Masuya. 2024. Homeostasis imbalance process ontology: a study on covid-19 infectious processes. *BMC Medical Informatics and Decision Making*, 23(4):1–13.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

A HoIP Ontology

The Homeostasis Imbalance Process Ontology (HoIP) systematically classifies a wide range of processes triggered by homeostatic imbalances. It primarily focuses on cellular senescence and the infectious processes related to COVID-19. The HoIP ontology is annotated based on COVID-19 related articles in PubMed using Protégé 5.5.0[3] and the Web Ontology Language (OWL). The COVID-19 infectious processes are manually annotated from PubMed articles. Passages corresponding to the annotated terms are included, and article identifiers (e.g., PubMed ID (PMID: 25301932), DOI) are provided using the database cross-reference annotation property. Process relationships are annotated with object properties. Causal relationships between processes are primarily annotated using the ‘has result’ relationship while sub-processes are identified using the ’has part’ relation. HoIP defines a "COVID-19 infectious course" as a sequence of processes that describe the infectious mechanisms.These courses are organized into an is-a (subclass of) hierarchy by severity, ranging from mild to severe. Notably, the "COVID-19 severe course" includes a subclass associated with acute respiratory distress syndrome (ARDS). These COVID-19-specific processes are used as our primary dataset for this study.

B ZMPA: Optimizing Predictions

Figure 5. illustrates the ZMPA architecture with a real-time example from the HoIP dataset. It shows how the ZMPA model integrates predictions from both the ZPA and MPA models to maximize entity coverage. The The ZPA model bridges the gap in biomedical information extraction by leveraging zero-shot classification to map implicit processes with high precision and adaptability. Additionally, the MPA model maps HoIP processes to MeSH headings (Munne and Ichise, 2023), and the integration of ontology mapping (Rahman et al., 2024, 2020) enhances the model’s flexibility and coverage, making it a robust solution for handling complex biomedical datasets. For the Ft-ZMPA variant, we use the fine-tuned ZPA (Ft-ZPA) in place of the standard ZPA to enhance performance.

[3] https://protege.stanford.edu

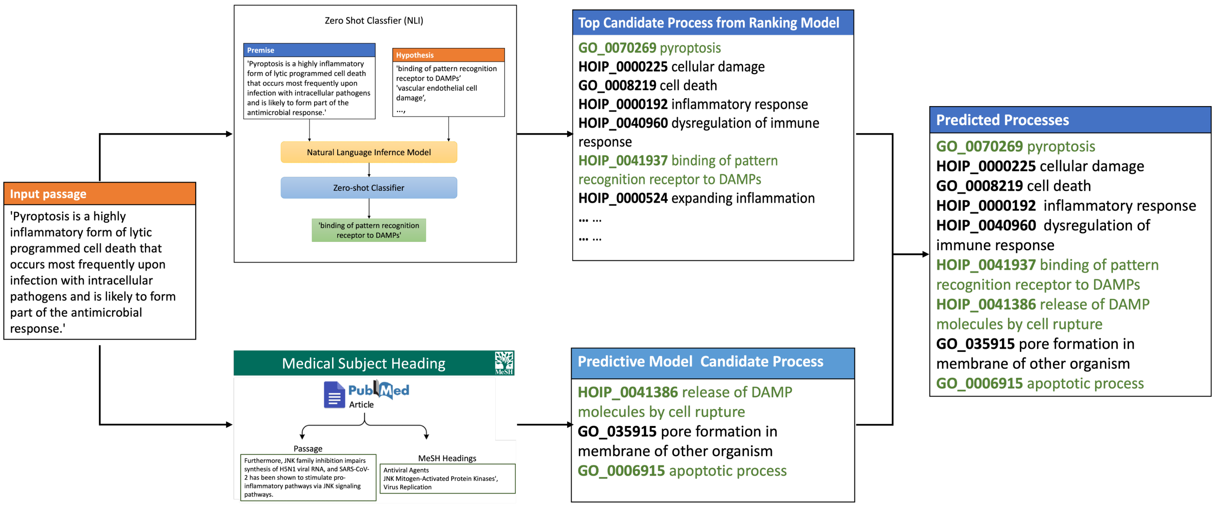

Input passage
'Pyroptosis is a highly inflammatory form of lytic programmed cell death that occurs most frequently upon infection with intracellular pathogens and is likely to form part of the antimicrobial response.'
Zero Shot Classfier (NLI)
Premise
'Pyroptosis is a highly inflammatory form of lytic programmed cell death that occurs most frequently upon infection with intracellular pathogens and is likely to form part of the antimicrobial response.'
Hypothesis
'binding of pattern recognition receptor to DAMPs'
'vascular endothelial cell damage',
…,
Natural Language Infernce Model
Zero-shot Classifier
'binding of pattern recognition receptor to DAMPs'
Medical Subject Heading
MeSH
PubMed
Article
Passage
Furthermore, JNK family inhibition impairs synthesis of H5N1 viral RNA, and SARS-CoV-2 has been shown to stimulate pro-inflammatory pathways via JNK signaling pathways.
MeSH Headings
Antiviral Agents
JNK Mitogen-Activated Protein Kinases',
Virus Replication
Top Candidate Process from Ranking Model
GO_0070269 pyroptosis
HOIP_0000225 cellular damage
GO_0008219 cell death
HOIP_0000192 inflammatory response
HOIP_0040960 dysregulation of immune response
HOIP_0041937 binding of pattern recognition receptor to DAMPs
HOIP_0000524 expanding inflammation
… …
… …
Predictive Model Candidate Process
HOIP_0041386 release of DAMP molecules by cell rupture
GO_035915 pore formation in membrane of other organism
GO_0006915 apoptotic process
Predicted Processes
GO_0070269 pyroptosis
HOIP_0000225 cellular damage
GO_0008219 cell death
HOIP_0000192 inflammatory response
HOIP_0040960 dysregulation of immune response
HOIP_0041937 binding of pattern recognition receptor to DAMPs
HOIP_0041386 release of DAMP molecules by cell rupture
GO_035915 pore formation in membrane of other organism
GO_0006915 apoptotic process

Figure 5: ZMPA model architecture

C WellcomeBertMesh

WellcomeBertMesh is utilising the latest state of the art model in the biomedical domain which is PubMedBert from Microsoft and In addition, it integrates a Multilabel attention head, a crucial component that enables the model to dynamically focus on different tokens for each label. This functionality enhances the model's ability to assess the relevance of various labels, thereby improving its performance in complex biomedical text analysis tasks

WellcomeBertMesh is trained the model using data from the BioASQ competition which consists of abstracts from PubMed publications. They use 2016-2019 data for training and 2020-2021 for testing which gives us 2.5M publications to train and 220K to test. This is out of a total of 14M publications. It takes 4 days to train WellcomeBertMesh on 8 Nvidia P100 GPUs. The model achieves 63% micro f1 with a 0.5 threshold for all labels.

$$J(\theta) = -\frac{1}{NK} \sum_{i=1}^{N} \sum_{j=1}^{K} [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})] \tag{1}$$

where: N is the number of samples, K is the number of labels. $\hat{y}_{ij} \in [0, 1]$ is the predicted probability for the i-th sample and the j-th label. $y_{ij} \in \{0, 1\}$ is the true value for the i-th sample and the j-th label.