# Implementation of Fine-Tuned BERT for Enzyme Classification Based on Gene Ontology

Matthew Martianus Henry, Christian Kenneth and
Bens Pardamean

September 5, 2024

# Implementation of Fine-Tuned BERT for Enzyme Classification Based on Gene Ontology

Matthew Martianus Henry
*Bioinformatics and Data Science Research Center*
*Bina Nusantara University*
Jakarta, Indonesia
matthewmartianush@gmail.com
(corresponding author)

Christian Kenneth
*Bioinformatics and Data Science Research Center*
*Bina Nusantara University*
Jakarta, Indonesia
christiankennethasikin@gmail.com

Bens Pardamean
*Computer Science Department, BINUS Graduate Program – Master of Computer Science*
*Bina Nusantara University*
Jakarta, Indonesia
bpardamean@binus.edu

*Abstract*— **Enzymes are biocatalysts with vital roles in biological functions and many industrial applications. Diverse enzymes are classified using Enzyme Commission (EC) nomenclature, making differentiation challenging. On the other hand, another biological information, gene ontology (GO), can describe the biological aspects of enzymes, covering related biological processes (BP), molecular functions (MF), and their locations within cells (CC). This study proposes a novel EC class and subclass classification of enzymes within the ontology subclass based on their GO semantics using a Bidirectional Encoder Representation of Transformer (BERT). The BERT model is first fine-tuned using the preprocessed GO term name and definition, with the enzymes in each ontology class (BP, MF, or CC) are also divided based on how the GO assigned, either through manual annotation (NONIEA) or electronically inferred (IEA). BERT successfully obtained 0.93, 0.60, 0.99, 0.90, 0.40, and 0.35 F1 scores during fine-tuning for BP IEA, BP NONIEA, MF IEA, MF NONIEA, CC IEA, and CC NONIEA, respectively. On the test set, the fine-tuned BERT significantly outperformed GOntoSim, a framework to calculate semantic similarity based on classical information theory, in EC class classification across all metrics with less inference time in all ontology subclass. Expanded further to the EC subclass, BERT can classify the enzyme on the EC subclass level in BP IEA and MF IEA ontology subclass. However, longer epochs are needed in fine-tuning. This result shows that the names and definitions of GO terms are distinguishable features in classifying enzymes as an alternative to the information content approach.**

***Keywords—BERT, fine-tuning, enzyme classification, gene ontology, GOntoSim***

## I. Introduction

Enzymes are biological molecules produced in all living cells and can speed up or catalyze various biological functions. It can be extracted and applied to many industrial sectors [1]. For example, trehalose synthase from bacteria can produce trehalose, an essential stabilizer in the food and cosmetic industries, and D-allulose epimerase for producing low-calorie sugar in many food industries [2], [3], [4]. Since different enzymes perform different functions, there is a need for a system that classifies these biocatalysts to facilitate extensive studies regarding specific enzymes. The International Union of Biochemistry and Molecular Biology (IUB) has established the Enzyme Commission (EC) classification and nomenclature system [5]. The system is based on its functionals, known as EC numbers. EC numbers are composed of four digits, each representing the main class, subclass, sub-subclass, and substrate class, respectively [6]. Recently, the EC is composed of 7 main classes, 77 subclasses, 308 sub-subclasses, and 7831 substrate classes [5]. The seven main classes are comprised of oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), ligases (EC 6), and translocases (EC 7), with the latter added recently in 2018. Due to the numerous functional classes, it becomes challenging to distinguish enzyme classes and classify novel enzymes. While various machine learning-based methods are proposed to classify enzymes using protein structures and sequences, none utilize gene ontology.

Gene Ontology (GO) is a biological knowledge representation of the functions of gene products. GO is divided into three main functional aspects: molecular function (MF), which describes the specific activity of the gene products; cellular component (CC), which indicates the cellular location where the products perform their function; and biological process (BP), which explains the broader tasks they are involved in [7]. Each central aspect comprises a directed acyclic graph (DAG), with each node representing a GO term containing a semantic description, while each edge contains the relationship between GO terms [8]. In functional bioinformatics studies, GO terms are used to annotate and characterize gene products. Gene products with similar functions will have the same GO terms with high semantic similarity and vice versa [7]. Proteins with similar functionalities can be classified by calculating the semantic similarities of gene products. Several classical methods have been established to calculate semantic similarities, such as Resnik [9], Wang, GOGO [10], and the latest and most advanced, GOntoSim [8]. However, with the advancement of artificial intelligence (AI), a new approach to classifying the gene product can be made using semantic similarity measurement from text semantics.

In the realm of AI, text has been used to solve many problems, from text classification to topic modeling [11], [12], [13], [14]. The state-of-the-art model for such tasks is Bidirectional Encoder Representations from Transformers or BERT [15]. Through the novel self-attention mechanism, BERT has shown remarkable performance on various tasks. In molecular biology and bioinformatics, BERT has been used in DNA promoter prediction [16], protein property prediction [17], and mRNA design optimization [18]. Previous study has shown that fine-tuned BERT embeddings perform well in measuring GO terms similarity [7]. Yet, there is minimal application of this framework in areas of enzyme classification.

This study proposes a new method to classify enzymes based on their functionalities through their EC class, with GO terms names and definitions embedded to each enzyme serves as the basis for classification. BERT was leveraged to extract the semantic similarities from the GOs annotated in the enzyme. GOntoSim was also employed to compare the extracted semantic quality from the fine-tuned BERT. Section II describes the study methodology in detail, while the result and discussion are outlined in Section III and IV, respectively. The last section concludes the whole study and suggests improvement directions for future works based on the obtained result.

## II. Methodology

The methodology in this study is composed of data collection, data preprocessing, BERT finetuning, and classification performance analysis. The workflow summary is shown in Fig. 1, and the details are described below.

### A. Data Collection

The study used the preprocessed enzymes dataset from the Swiss-Prot database used in [8]. The enzymes were classified into three GO aspects and two subclasses based on the approach of GO term assignment: manual annotation (NONIEA) or with electronically inferred annotation (IEA). This classification resulted in six ontology subclasses: BP IEA, BP NONIEA, MF IEA, MF NONIEA, CC IEA, and CC NONIEA. There were 10517, 3092, 10242, 3377, 7406, and 2820 enzymes from BP IEA, BP NONIEA, MF IEA, MF NONIEA, CC IEA, and CC NONIEA, respectively. The EC classes information (1-6) for each enzyme were also gathered, which serves as the ground truth in the classification process. The enzyme must have at least a GO term associated with it, and the relations between them are not a one-to-one correspondence. In this study, 2020-01-01 GO graph was utilized to extract the unique GO term IDs, names, and definitions of each enzyme.

The resulting dataset for each GO aspect and assignment approach pairs consists of unique enzyme code based on UniProt ID mapped to the EC class label and GO features: GO term names and definitions. For instance, for tyrosinase in the CC IEA dataset, it will have a UniProt ID of B8NM74, labeled as EC ‘1’, and two GO features from GO:0016021 with the name “component of membrane” and definition “component membrane consisting gene products protein complexes least part peptide sequence embedded hydrophobic region membrane”; and from GO:0016020 with the name “membrane” and definition “lipid bilayer along proteins protein complexes embedded attached”. The GO features will be further preprocessed for BERT input.

### B. Data Preprocessing

The unique GO term definition has its punctuation and stop words removed, followed by a lemmatization using WordNet. The preprocessed GO term definition was then concatenated using the [SEP] token with the GO term name. Since each enzyme can have many GO terms attached, the preprocessed GO term texts were concatenated together for an enzyme with more than a single GO term. For example, in CC IEA, the preprocessed GO term text for enzyme tyrosinase with two GO terms attached (GO:0016021 and GO:0016020) is "component membrane consisting gene products protein complexes least part peptide sequence embedded hydrophobic region membrane [SEP] integral component of membrane lipid bilayer along proteins protein complexes embedded attached [SEP] membrane". The enzyme representation was made from the concatenation of two preprocessed GO term text, which is "component membrane consisting gene products protein complexes least part peptide sequence embedded hydrophobic region membrane [SEP] integral component of membrane" from GO:0016021 and "lipid bilayer along proteins protein complexes embedded attached [SEP] membrane" from GO:0016020. The enzymes in each ontology subclass, along with its preprocessed GO term text and corresponding clusters, were then split into training and testing sets, with 70% for the training set and 30% for the testing set. The split was stratified, ensuring the EC class proportions were identical in the training and testing set.

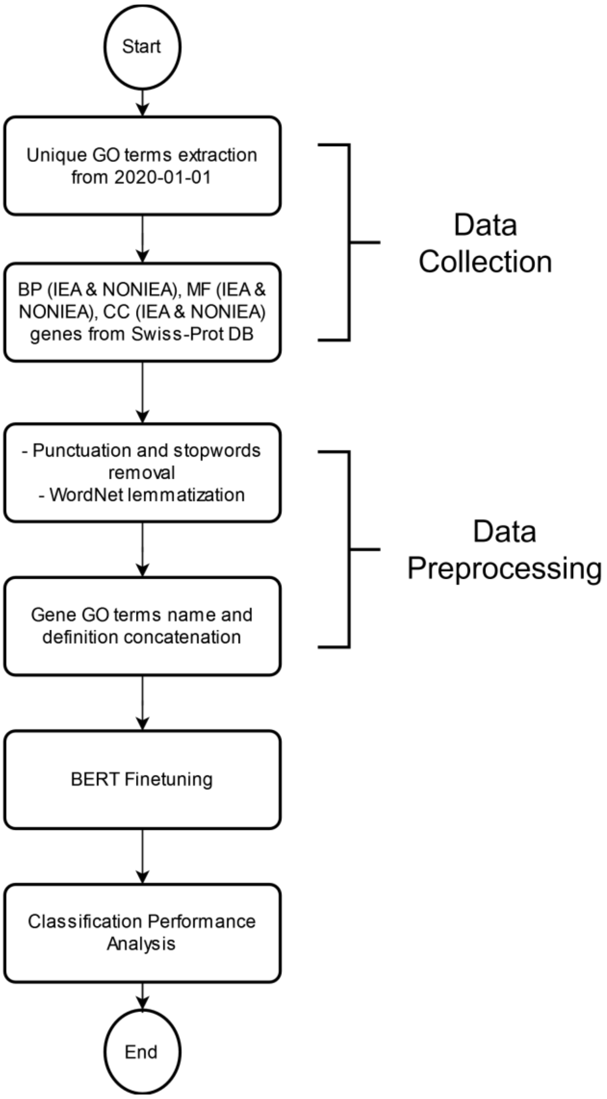
Start
Unique GO terms extraction from 2020-01-01
BP (IEA & NONIEA), MF (IEA & NONIEA), CC (IEA & NONIEA) genes from Swiss-Prot DB
Data Collection
- Punctuation and stopwords removal
- WordNet lemmatization
Gene GO terms name and definition concatenation
Data Preprocessing
BERT Finetuning
Classification Performance Analysis
End

Fig. 1. Research workflow

### C. BERT Fine-Tuning

The BERT model was fine-tuned for each ontology subclass, with the preprocessed GO term text in the training set serving as the BERT finetuning input. BERT finetuning aimed to classify the enzymes in each ontology subclass into their corresponding EC class. The BERT model was fine-tuned for five epochs for both the BP subclass and MF NONIEA, two for MF IEA, and ten for both CC subclass. Adam was employed as the optimizer in each finetuning, with a learning rate of 0.000003. The enzymes were processed per batch, with each batch composed of eight enzymes.

### D. Classification Performance Analysis

The performance of BERT finetuning was assessed by classifying the enzymes in the test set. The GOntoSim method was also applied to the test set for comparison. Since GOntoSim is a graph rule-based method, no prior training is required. The predicted EC classes were validated with the ground truth through the F1 score. However, since in [8], the cluster is assessed through Adjusted Rand Index (ARI) [19], Adjusted Mutual Information (AMI) [20], Fowlkes-Mallows score (FM score) [21], homogeneity score, completeness

TABLE I. BERT FINE-TUNING RESULT ON EC MAIN CLASS

| Ontology Subclass | Epochs | Time Elapsed (s) | F1 Score | ARI | AMI | FM Score | Homogeneity | Completeness | V-Measure |
|---|---|---|---|---|---|---|---|---|---|
| BP IEA | 5 | 2035 | 0.93 | 0.91 | 0.85 | 0.94 | 0.85 | 0.85 | 0.85 |
| BP NONIEA | 5 | 600 | 0.60 | 0.54 | 0.44 | 0.71 | 0.40 | 0.49 | 0.44 |
| MF IEA | 2 | 742 | 0.99 | 0.98 | 0.97 | 0.99 | 0.97 | 0.97 | 0.97 |
| MF NONIEA | 5 | 380 | 0.90 | 0.88 | 0.82 | 0.92 | 0.80 | 0.83 | 0.82 |
| CC IEA | 10 | 1960 | 0.40 | 0.19 | 0.21 | 0.51 | 0.17 | 0.28 | 0.21 |
| CC NONIEA | 10 | 810 | 0.35 | 0.21 | 0.15 | 0.51 | 0.13 | 0.20 | 0.16 |

score, and the V-measure [22] clustering metrics, those metrics were also employed as well.

## III. RESULTS

### A. BERT Fine-Tuning

Table I shows the finetuning result from BERT across ontology subclasses. Unlike the other ontology classes, MF exhibited the lowest time elapsed in the IEA and NONIEA subclass. The finetuning time for the IEA subclass surpassed the NONIEA one, with BP IEA as the highest finetuning time among all the ontology subclasses.

Compared by the model performance, MF IEA was able to achieve remarkable result on all metrics with only two epochs, while MF NONIEA needs five epochs for convergence. Even though fine-tuned longer, the MF NONIEA finetuning results do not surpass the IEA one. The IEA superior performance also occurs in BP, where the IEA class needs less time for convergence with more satisfactory result than the NONIEA one. However, in the CC class, the metrics revealed inadequate results on both IEA and NONIEA even though it has been fine-tuned much longer than the other class. This suggests that using the text features alone, CC clusters are not well-separated.

### B. BERT and GOntoSim Performance Comparison

Table II shows that the fine-tuned BERT has lesser inference latency than GOntoSim in all EC classes. The margin is significant, especially in the BP ontology class, where the inference can take hours. The total time elapsed for BERT finetuning and inference in BP is far less than that of GOntoSim methods. However, compared with GOntoSim, the other ontology class's total finetuning and inference latency is higher. Nevertheless, the latency is tolerable, as the accuracy gained by BERT is higher, as shown in all the graphics in Fig. 2.

TABLE II. COMPARISON OF TIME ELAPSED FROM BERT AND GONTOSIM INFERENCE ON EC MAIN CLASS

| Ontology Subclass | Time Elapsed for GOntoSim Inference (s) | Time Elapsed for BERT Inference (s) |
|---|---|---|
| BP IEA | 21845 | 40 |
| BP NONIEA | 9833 | 11 |
| MF IEA | 283 | 69 |
| MF NONIEA | 55 | 22 |
| CC IEA | 98 | 9 |
| CC NONIEA | 34 | 10 |

In BP IEA (Fig. 2a), BERT achieved an outstanding score in F1 and other clustering metrics. This indicates that the underlying semantic distribution extracted in BP IEA is well-separated based on the EC main class labels. The predicted EC class is well-distributed and correctly labeled, as shown by the high clustering metrics and high F1 score. The same condition occurred in BP NONIEA (Fig. 2b). Concurrently, GOntoSim failed to achieve moderate scores in all the metrics except the FM score (Fig. 2a and Fig. 2b). The FM score indicates that GOntoSim achieved moderate precision and recall but is not sufficient to achieve moderate score on other clustering metrics. GOntoSim also displayed abundant mislabeling, hence the low F1 score. This shows that the semantics extracted from GOntoSim is complex and not well-structured.

BERT achieves even better semantics embeddings in MF across both ontology subclasses, as shown by the higher F1 score and clustering metrics achieved compared with BP. On the contrary, in MF IEA, GOntoSim achieved high clustering metrics, even approaching the performance of BERT, but still has a low F1 score (Fig. 2c). This extracted GOntoSim underlying structure is unique since it encompassed the well-structured property but with poor labeling. In MF NONIEA, the GOntoSim performance is also better than in BP NONIEA, even though the result is relatively moderate (Fig. 2d).

In CC, BERT still outperforms GOntoSim, but the quality of extracted semantics needs improvement. This is shown by the lower F1, and other clustering metrics scores compared with BP and MF (Fig. 2e and Fig. 2f). Both ontology subclasses show similar performance in BERT and GOntoSim, suggesting that the extracted semantics from both are similar in IEA and NONIEA. Similar to BP, GOntoSim achieves a higher FM score than any other clustering metrics, indicating the same semantic complexities as in BP.

### C. BERT Performance on Ontology Subclass EC Number Subclass Level

Since BERT drastically surpasses GOntoSim on the main class level, the study is extended to classify the EC on the subclass level using BERT. BERT was fine-tuned using the enzyme GO name and definition on the EC subclass level. The number of EC subclasses in each ontology subclass far exceeds the number of EC classes. BP has 44 and 43 EC subclasses for the IEA and NONIEA ontology subclass, respectively. Meanwhile, in both MF and CC ontology subclasses, 44 EC subclasses are present. Like in the EC class, the dataset was split into 70% training and 30% test sets, with a stratification strategy on the EC subclass. The training set is used to fine-tune the BERT model, while the

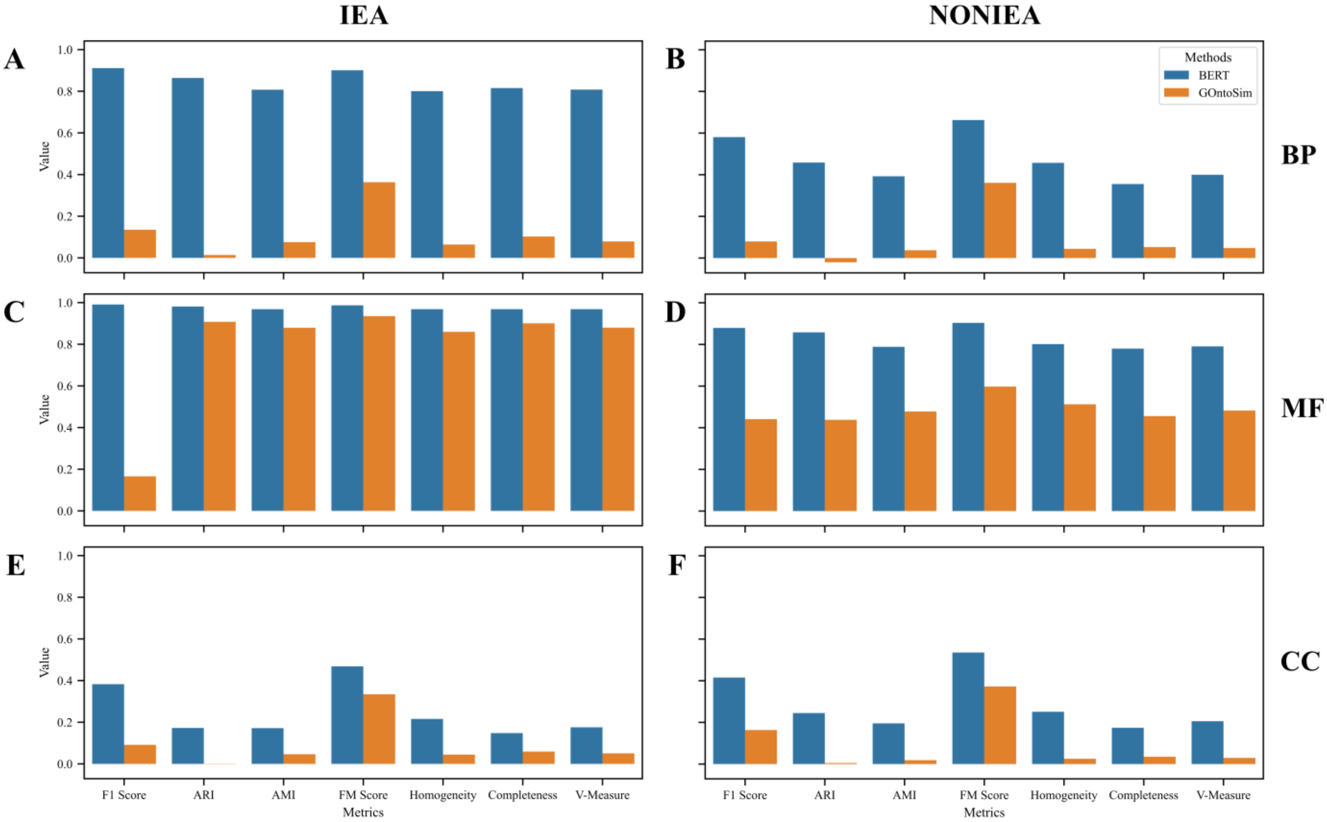

Fig. 2. Inference result comparison in (a) BP IEA, (b) BP NONIEA, (c) MF IEA, (d) MF NONIEA, (e) CC IEA, and (f) CC NONIEA ontology subclass.

rest is kept to assess the model performance. The inference result is shown in Table III.

On the EC subclass level, BERT required longer finetuning time, marked by increased elapsed epochs. In the BP class, BERT needs 20 epochs to converge in IEA, with outstanding performance on the F1 score and other clustering metrics. This shows BERT ability to finally distinguish each EC subclass from the text semantics, even though longer learning time is needed. However, in the NONIEA ontology subclass, BERT only achieves moderate performance. As shown in Table III, BP IEA F1 score, and clustering metrics achieved higher score with a significant margin compared to the NONIEA.

The same condition also happened in MF. Specifically in MF NONIEA, BERT requires many more epochs than the MF IEA to achieve only a moderate score. In MF IEA, however, BERT successfully learned the semantics from the text, shown by the high F1 score that reveal mostly correct EC subclass prediction and a high clustering metrics that suggests well-separated EC subclass features, with lower epochs than MF NONIEA.

In CC, the performance is much worse, with low performance in both ontology subclasses, even though 50 epochs have elapsed, as exhibited in the two bottom rows of Table III. This shows BERT incapability to extract the EC subclass semantics from GO names and text.

## IV. Discussions

### A. *BERT Efficacy in EC Class*

In GO literature, CC refers to a location in the gene product where the molecular function occurs [23] (*e.g.* plasma membrane, cytoskeleton, or clathrin complex). While eucaryotes may have diverse molecular functions, their constituent components are homogeneous on some level. Their variations were only adapted to each unique biological process. Hence, differently functioning enzymes are able to have the same CC, leading to the same GO terms. For example, enzymes serine-tRNA ligase (UniProt ID: B8DW52), thioredoxin reductase (UniProt ID: P47348), and probable glutathione S-transferase (UniProt ID: D2YW48) are in EC class 6, 1, and 2, respectively. While those enzymes perform distinct functions, they reside in the same cellular component, the cytoplasm, which corresponds to the same GO annotation, GO:0005737. Another example shows enzyme palmitoyltransferase (UniProt ID: Q6BP80) in EC class 2 has six GO terms attached, which are GO:0005768, GO:0005794, GO:0016020, GO:0016021, GO:0031901, and GO:0000139. Meanwhile, in enzyme pheromone-processing carboxypeptidase (UniProt ID: E6R6G5) in EC class 3, the GO terms attached are GO:0005794, GO:0016020, and GO:0016021, which are the subset of GO terms in enzyme palmitoyltransferase in EC class 2. This subset proves the closely packed nature of the GO terms in CC.

Such overlaps rarely occur in BP or MF, which means the clusters in those classes can be distinguished by only examining their GO names and definitions. An overlap is observed occasionally, but with lower frequency. These enable BERT and even GOntoSim to have a better performance, compared with the one acquired in CC. This also happens because BP and MF have more information about the functionality of each enzyme class, resulting in a better score than CC, which only tells the location of the enzymes.

TABLE III. BERT INFERENCE RESULT ON EC SUBCLASS

| Ontology Subclass | Epochs Elapsed during Fine-tuning | F1 Score | ARI | AMI | FM Score | Homogeneity | Completeness | V-Measure |
|---|---|---|---|---|---|---|---|---|
| BP IEA | 20 | 0.81 | 0.91 | 0.90 | 0.92 | 0.91 | 0.89 | 0.90 |
| BP NONIEA | 20 | 0.34 | 0.60 | 0.62 | 0.65 | 0.71 | 0.62 | 0.66 |
| MF IEA | 20 | 0.95 | 0.99 | 0.99 | 0.99 | 0.92 | 0.98 | 0.99 |
| MF NONIEA | 50 | 0.68 | 0.57 | 0.69 | 0.62 | 0.75 | 0.71 | 0.73 |
| CC IEA | 50 | 0.15 | 0.04 | 0.17 | 0.22 | 0.30 | 0.17 | 0.22 |
| CC NONIEA | 50 | 0.06 | 0.05 | 0.07 | 0.20 | 0.18 | 0.11 | 0.13 |

However, BERT still shows promising results in the CC class. As depicted in Fig. 2e and Fig. 2f, BERT outperforms GOntoSim in predicting the clusters in CC by a significant margin. The testing performance in all metrics aligns with the finetuning result, which shows no overfitting. Text semantics acquired from the gene product name and definition still have worthier capabilities to cluster the enzymes.

## B. *BERT Efficacy in EC Subclass*

Due to the increasing number of ground truth clusters in the EC subclass, BERT needs longer finetuning time. The abundant number of EC subclass makes the semantics harder to distinguish. This is because the EC subclass is a subset of the EC class. The genes, especially in BP and MF, are well-separated in the EC class by their contained GO names and definitions. Expanded further into the EC subclass, the difference between genes in the subclass level that originated from the same class is vague. This condition is even worse in CC, as the annotated GO terms in genes between each EC class in CC are not that different. Thus, more epochs elapsed for all ontology subclasses.

BERT only achieves remarkable classification and clustering performance in BP IEA and MF IEA. This shows that although the GOs in one EC class have similar functions, their functionality variations can be distinguished by the EC subclass. For example, in BP IEA, enzyme nitrogenase Mo-Fe protein beta chain (UniProt ID: P0CW52) in EC subclass 1.18 have two GO terms attached, which are GO:0009399 and GO:0055114. Enzyme thioredoxin reductase (UniProt ID: O66790), which is classified as EC subclass 1.8, has two GO terms attached as well, namely GO:0019430 and GO:0055114. These two enzymes that are both involved in reduction-oxidation reaction have similar GO terms (GO:0055114), which also indicates a reduction-oxidation process since both originated from the same EC class. Yet, the two still can be differentiated through the EC subclass by the other GO terms. However, this condition might not be true in the other ontology subclass, especially in CC, where the attached GO terms among the genes are alike since the EC class level.

## CONCLUSION AND FUTURE WORKS

This study leveraged BERT to cluster enzymes based on the names and descriptions of the GO terms attached. BERT has shown excellent performance in differentiating the genes on both EC class and subclass level in some ontology subclass through the GO terms name and definition. For the EC class, the GO terms semantics are easily differentiated in BP IEA, MF IEA, and MF NONIEA, shown by the high F1 score on the test set, which are around 0.93, 0.99, 0.90, respectively. The semantics for the EC subclass are distinguishable at BP IEA and MF IEA, shown by 0.81 and 0.95 F1 score, which are relatively higher than the other ontology subclass. The other clustering metrics positively correlates with the F1 score on both EC class and subclass. The EC class and subclass are not easily differentiable on ontology subclass like CC since the attached GO terms are alike among the genes. Apart from that, this study has successfully exhibited the utilization of text for enzyme classification.

In further studies, Low Rank Adapters (LoRA) emerges as a promising alternative to BERT finetuning. Alternatives from other paradigms can also be utilized, such as Node2Vec or Graph Convolutional Network (GCN), to extract the semantics from the GO term relationship from the GO graph.

## REFERENCES

[1] P. K. Robinson, “Enzymes: principles and biotechnological applications,” *Essays in Biochemistry*, vol. 59, pp. 1–41, Nov. 2015, doi: 10.1042/bse0590001.

[2] J. P. Trinugroho, F. Asadi, A. A. Hidayat, R. Nirwantono, and B. Pardamean, “DNA marker utilization for the sustainable production of trehalose,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 1297, no. 1, p. 012079, Feb. 2024, doi: 10.1088/1755-1315/1297/1/012079.

[3] R. Nirwantono, J. P. Trinugroho, D. Sudigyo, A. A. Hidayat, B. Mahesworo, and B. Pardamean, “Genome mining of potential enzyme from Chelatococcus composti for sustainable production of D-Allulose,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 1169, no. 1, p. 012083, Apr. 2023, doi: 10.1088/1755-1315/1169/1/012083.

[4] R. Nirwantono, A. A. Hidayat, J. P. Trinugroho, D. Sudigyo, and B. Pardamean, “An assessment of potential thermostable D-Allulose epimerases obtained from genome mining using a computational simulation approach,” *Commun. Math. Biol. Neurosci.*, 2023, doi: 10.28919/cmbn/8026.

[5] A. Chang *et al.*, “BRENDA, the ELIXIR core data resource in 2021: new developments and updates,” *Nucleic Acids Research*, vol. 49, no. D1, pp. D498–D508, Jan. 2021, doi: 10.1093/nar/gkaa1025.

[6] Z. Tao, B. Dong, Z. Teng, and Y. Zhao, “The Classification of Enzymes by Deep Learning,” *IEEE Access*, vol. 8, pp. 89802–89811, 2020, doi: 10.1109/ACCESS.2020.2992468.

[7] D. Duong *et al.*, “Evaluating Representations for Gene Ontology Terms,” Sep. 18, 2019. doi: 10.1101/765644.

[8] A. B. Kamran and H. Naveed, “GOntoSim: a semantic similarity measure based on LCA and common descendants,” *Sci Rep*, vol. 12, no. 1, 2022, doi: 10.1038/s41598-022-07624-3.

[9] P. Resnik, “Using Information Content to Evaluate Semantic Similarity in a Taxonomy,” Nov. 29, 1995, *arXiv*: arXiv:cmp-lg/9511007. Accessed: Jun. 27, 2024. [Online]. Available: http://arxiv.org/abs/cmp-lg/9511007

[10] C. Zhao and Z. Wang, “GOGO: An improved algorithm to measure the semantic similarity between gene ontology terms,” *Sci Rep*, vol. 8, no. 1, p. 15107, Oct. 2018, doi: 10.1038/s41598-018-33219-y.

[11] A. A. Hidayat, R. Nirwantono, A. Budiarto, and B. Pardamean, “BERT-based Topic Modeling Approach for Malaria Research Publication,” in *2022 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, Jakarta,

Indonesia: IEEE, Nov. 2022, pp. 326–331. doi: 10.1109/ICIMCIS56303.2022.10017743.

[12] A. Budiarto, R. Rahutomo, H. N. Putra, T. W. Cenggoro, M. F. Kacamarga, and B. Pardamean, "Unsupervised News Topic Modelling with Doc2Vec and Spherical Clustering," *Procedia Computer Science*, vol. 179, pp. 40–46, 2021, doi: 10.1016/j.procs.2020.12.007.

[13] R. Rahutomo, F. Lubis, H. H. Muljo, and B. Pardamean, "Preprocessing Methods and Tools in Modelling Japanese for Text Classification," in *2019 International Conference on Information Management and Technology (ICIMTech)*, Jakarta/Bali, Indonesia: IEEE, Aug. 2019, pp. 472–476. doi: 10.1109/ICIMTech.2019.8843796.

[14] M. Isnan, G. N. Elwirehardja, and B. Pardamean, "Sentiment Analysis for TikTok Review Using VADER Sentiment and SVM Model," *Procedia Computer Science*, vol. 227, pp. 168–175, 2023, doi: 10.1016/j.procs.2023.10.514.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 24, 2019, *arXiv*: arXiv:1810.04805. doi: 10.48550/arXiv.1810.04805.

[16] S. Li *et al.*, "CodonBERT: Large Language Models for mRNA design and optimization," Sep. 12, 2023. doi: 10.1101/2023.09.09.556981.

[17] N. Q. K. Le, Q.-T. Ho, V.-N. Nguyen, and J.-S. Chang, "BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection," *Comput. Biol. Chem.*, vol. 99, p. 107732, 2022, doi: 10.1016/j.compbiolchem.2022.107732.

[18] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, Apr. 2022, doi: 10.1093/bioinformatics/btac020.

[19] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985, doi: 10.1007/BF01908075.

[20] N. X. Vinh, J. Epps, and J. Bailey, "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[21] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, Sep. 1983, doi: 10.1080/01621459.1983.10478008.

[22] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420, Jun. 2007.

[23] P. D. Thomas, "The Gene Ontology and the Meaning of Biological Function," in *The Gene Ontology Handbook*, vol. 1446, C. Dessimoz and N. Škunca, Eds., in Methods in Molecular Biology, vol. 1446. , New York, NY: Springer New York, 2017, pp. 15–24. doi: 10.1007/978-1-4939-3743-1_2.