ECEN 766 Final Project, 2025, pp. 1–9

doi:

Benchmarking Transformer-Based Metagenomic Functional Profiling: A Comparison with Alignment Methods

Anisha Gollapalli,[1] Jonathan Turck[2] and Jen Li Kao[3]

[1]Department of Electrical and Computer Engineering, Texas A&M University, [2]Department of Electrical and Computer Engineering, Texas A&M University and [3]Department of Electrical and Computer Engineering, Texas A&M University

Abstract

Metagenomic functional profiling is essential for understanding microbial community dynamics and metabolic potential. Traditional alignment-based methods such as HUMAnN3 rely on sequence homology to assign functional annotations but often fail to classify highly divergent genes, leaving a substantial fraction of reads unannotated. Transformer-based models offer a promising alternative by capturing long-range contextual features; DeepECTransformer, a BERT-based model pretrained on 22 million enzyme sequences spanning 2 802 EC numbers, has already demonstrated the ability to annotate previously uncharacterized proteins in E. coli K-12. Here, we benchmark DeepECTransformer against HUMAnN3 using a curated validation set of 7 884 prokaryotic proteins and a benchmarking dataset derived from the CAMI-II toy microbiome assemblies across five body sites, which comprises 251 559 high-confidence EC annotations. DeepECTransformer achieved a micro-F1 score of 0.89 in cross-validation and maintained high coverage (annotating over 95 % of ORFs), compared to HUMAnN3's average F1 of 0.12. These results establish transformer-based annotation as an potentially viable approach to metagenomic functional profiling, capable of extending discovery in enzyme diversity.

Availability: Code and data used in this work are available at:
`https://github.com/JonathanTurck02/ML_taxProfiling_ECEN766`
`https://doi.org/10.5281/zenodo.15192200`

Key words: Metagenomics, Functional Profiling, Deep Learning, Transformer Models, Sequence Annotation, Enzyme Classification, BERT, Alignment-Based Methods, CAMI Benchmarking, Microbial Communities.

Introduction

Problem Definition and Significance

Advances in next generation sequencing and high throughput techniques have enabled a rapid expansion of microbiota-associated genomic data across diverse environments and hosts. [4, 2]. Particularly shotgun metagenomics, or whole-metagenome sequencing, has enabled the sequencing of novel, uncultured microbial genes. Despite this, the understanding of the functional profile of these genes is still lacking.

Reference-based methods currently serve as the gold-standard in functional metagenome annotation. These approaches are homology based and therefore depend heavily on a curated reference database to annotate gene sequences. Due to this homology-based approach these methods benefit from high precision, however they often lack the ability to map genes/enzymes that diverge from the reference database [8, 5]. This shortfall of alignment methods often leaves a substantial amount of microbial genes unmapped and therefore unannotated. This issue is more pronounced in less characterized environments such as the non-human gut or environmental samples.

Recent developments in deep learning approaches allow for an alternative to reduce dependence on alignment-based methods. These methods can utilize context embedded within full gene sequences, that may allow for identification of divergent genes that have yet to be captured by updated reference databases and culture-based techniques.

Early studies implementing machine learning based methods for functional annotation of metagenomes suggest that these models have the capability to improve recall compared to alignment-based techniques [7]. Despite these improvements in recall, current one-vs-all (OvA) classifiers tend to lack in precision compared to traditional alignment-based techniques [1].

Despite this potential, evidence for the benchmarking of these methods remains unreliable. Furthermore, there is a lack of evidence for the real-world effectiveness of transformer-based models for full metagenome functional annotation. Rigorous benchmarking against well-established alignment-based tools is therefore essential to determine whether these newer methods can reliably complement, or even surpass, traditional approaches in functional profiling. Therefore, the aim of this study is to benchmark transformer-based models compared to alignment methods for functional profiling. The

pre-trained DeepECtransformer for enzyme prediction may be tested with gold-standard benchmarking datasets to directly compare its performance to assembly-based methods.

Related Work

Alignment-based methods, such as HUMAnN3, have been widely used for metagenomic functional profiling. These methods work by mapping sequencing reads to a curated pangenome [1]. If sequences fail to map to a pangenome, a translated search against a non-redundant protein database is performed. While these approaches excel at annotating genes with well-established homologs, sequences that fail to map remain unannotated. The issue of unmapped sequences is particularly prevalent in less-characterized metagenomic environments, where taxonomic and functional annotations are sparse in existing reference databases [2].

To address these limitations, hybrid and alignment-free approaches have been proposed. For example, Carnelian demonstrated the feasibility of machine-learning-based functional annotation by analyzing k-mer profiles. While this approach has improved recall, it lacks consistent precision in complex metagenomes compared to alignment-based methods [7, 1].

Transformer-based models, such as DeepECtransformer, represent a significant advancement in machine-learning-based functional annotation. By leveraging self-attention mechanisms, these models learn contextual relationships within entire amino acid sequences [9]. Early evaluations indicate that transformer-based models can identify previously unannotated genes [3]. However, comprehensive benchmarking of transformer-based models, specifically DeepECtransformer, remains insufficient. It is unclear whether the reported advancements will hold when applied to complex metagenomes.

This study positions DeepECtransformer within the broader ecosystem of metagenomic annotation tools by directly comparing its performance to that of HUMAnN3. It aims to evaluate both the overall precision and recall of these methods using gold-standard benchmarks. Ultimately, this study seeks to determine whether transformer-based models can effectively address the challenge of unmapped genes and enhance our understanding of microbiome functionality.

Proposed Methods

Research Plan

The DeepECtransformer will first be tested against the validation set used in Carnelian. The results will be compared to those that were recorded by the Carnelian model. This will allow us to ascertain the effectiveness of the DeepECtransformer with a dataset curated with the Critical Assessment of Metagenome Interpretation (CAMI)[6] and determine how the model will perform against the benchmark dataset. The inputs will be the amino acid sequence and the output will be the Enzyme Commission (EC) numbers.

While the DeepECtransformer is being trained, the benchmark dataset will be collected and assembled from CAMI and made ready for use with Prodigal (gene prediction) and DIAMOND(Mapping to UniRef90).

Once the benchmark data has been collected and the DeepECtrasformer's efficacy against the other approaches has been tested, we will be able to use the model to with the benchmark dataset and determine its accuracy in identifying the EC numbers. The model will then be refined to optimize its performance.

Models Overview

The current gold standard for functional annotation is alignment-based, as seen with HUMAnN3. HUMAnN3 is a homology-based profiler that uses an expanding dataset of archaea and bacteria that have been sequenced and annotated in past experiments. These experiments yielded high precision, but this profiler is also computationally intensive. Furthermore, while there are some methods to correctly profile some divergent sequences, this method struggles with anything too divergent or unknown, leaving many "unmapped" sequences.

The gold standard for functional profiling using Machine Learning is Carnelian. Carnelian, by comparison, uses a OvA approach to the unmapped sequences. Compared to HUMAnN3, this is less computationally intensive, and the binary model is relatively simple to implement. While this is useful, the binary model does have its limitations. For instance, there are many EC labels needing to be classified, and the OvA method needs a binary classifier for each class. As there are more EC numbers available, there will need to be more classifiers to available to accommodate this number. Thus, the classifier will be more complex and more computationally intensive. Another downside is the classifier does not take any context into account. Once again, there will be many sequences left unmapped.

Traditional homology-based functional annotation methods often fail to classify enzymes that lack known counterparts, leaving many genes uncharacterized. To address this limitation, DeepECtransformer leverages deep learning and transformer based models to improve enzyme function prediction.

DeepECtransformer is a Bidirectional Encoder Representations from Transformers(BERT)-based deep learning model designed to predict EC numbers directly from amino acid sequences. Unlike conventional methods that rely solely on sequence homology, DeepECtransformer incorporates a neural network and homology search to enhance prediction accuracy.

The model architecture consists of two Transformer encoders, which capture long-range dependencies in enzyme sequences. These encoders are followed by two convolutional layers, which extract local sequence features, and a final linear layer that predicts the EC number. The model has been trained on 22 million enzyme sequences, covering 2802 EC numbers. When combined with homology search, it extends its prediction capability to 5360 EC numbers, improving its ability to classify divergent and novel enzymes.

By integrating self-attention mechanisms, convolutional feature extraction, and homology-based validation, DeepEC-Transformer provides a powerful alternative to traditional alignment-based annotation methods, enabling the discovery of previously uncharacterized enzyme functions.

Figure 1 illustrates the architecture of DeepECtransformer.

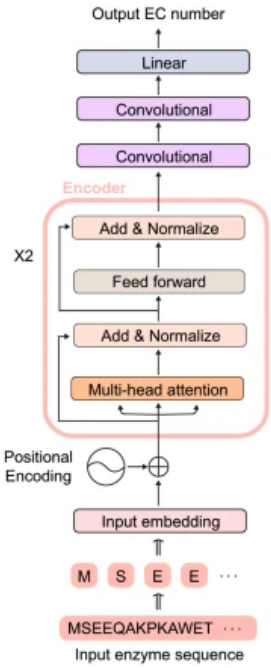

Fig. 1. Network architecture of DeepECTransformer

Experimental Procedures

We designed three experiments to systematically evaluate and improve the generalization performance of DeepECTransformer on enzyme classification tasks.

In the first experiment, we tested the default model with the ProtBERT encoder on a validation set derived from Carnelian and on five benchmark datasets representing diverse microbial environments. Each input protein sequence was passed through the model to predict one or more EC labels. Predictions were then compared to the ground truth using exact string matching, and performance was quantified using micro and macro evaluation metrics.

To investigate the performance gap between common and rare enzyme classes observed in the first experiment, we conducted a second experiment focused on sequence similarity analysis. Specifically, we compared the amino acid sequences of misclassified proteins with those that were predicted correctly. We used Jaccard similarity based on 3-mer token sets to measure how similar each misclassified sequence was to the correct set. The hypothesis was that low sequence similarity could correlate with prediction failure due to insufficient representation in the training set.

In the third experiment, we replaced the default encoder with ProtT5, a larger and more expressive protein language model pretrained on a broader dataset. We retrained DeepECTransformer from scratch using the same training data, keeping all other components fixed. This experiment aimed to assess whether a more powerful encoder could reduce performance bias and improve generalization, particularly on rare or dissimilar sequences. The ProtT5-based model was then evaluated using the same datasets and metrics as in the first experiment.

Data

Our project utilizes two primary datasets: a validation dataset for testing DeepECtransformer and a benchmarking dataset for comparative evaluation against existing methods.

The validation dataset is derived from Carnelian, a curated dataset introduced by Nazeen et al. (2020) [7]. Carnelian provides a high-quality benchmark for predicting enzyme function, containing 784 prokaryotic protein sequences labeled with unique 2010 EC numbers. This dataset is formatted for direct usage with DeepECTransformer, where amino acid sequences serve as input and EC labels provide ground truth for validation. By using this well-established dataset, we ensure a controlled and standardized evaluation of the model's performance.

To ensure an unbiased comparison between DeepECtransformer and HUMAnN3, we employ a benchmarking dataset derived from the Critical Assessment of Metagenome Interpretation (CAMI) standards[6]. CAMI provides a well-established framework for benchmarking metagenomic analysis tools, ensuring that results are evaluated against field-accepted standards. Unlike the validation dataset, which directly maps sequences to EC numbers, the benchmarking dataset requires the interpretation of EC numbers based on established functional profiling methods. This introduces an additional step of interpolating EC numbers from metagenomic data, as commonly done in HUMAnN3. Previous studies have demonstrated the feasibility of this approach, making it a reliable dataset for evaluating DeepECtransformer in a real-world metagenomic setting.

Using both data sets, we ensure a rigorous assessment of the DeepECtransformer, testing its performance on structured validation data, as well as its adaptability to realistic metagenomic scenarios.

In addition to the validation and benchmark datasets, we utilized the original training data from the DeepECTransformer paper for retraining purposes. This dataset was constructed from UniProt Knowledgebase (UniProtKB), specifically using Swiss-Prot and TrEMBL entries released in April 2018. To ensure high data quality and training efficiency, the original authors applied a series of filtering steps, which we followed consistently in our work. These included removing sequences without complete four-digit EC numbers, sequences containing non-standard amino acids, and those exceeding 1000 amino acids in length. Redundant sequences were eliminated, and only EC classes represented by at least 100 unique sequences were retained. This careful preprocessing ensured a well-balanced and reliable training set, enabling fair comparisons between the original and retrained models.

Evaluation Metrics and Baseline Comparisons

To comprehensively evaluate model performance, we adopted a combination of micro- and macro-averaged metrics, including precision, recall (sensitivity), and F_1 score. Micro-averaged metrics treat all predictions equally, aggregating over all classes, and are particularly sensitive to performance on dominant or frequent enzyme classes. In contrast, macro-averaged metrics compute per-class scores and average them equally, providing a better reflection of how well the model performs across rare or underrepresented EC classes. The use of both metrics offers a balanced perspective on overall accuracy and class-level fairness.

In benchmark experiments, we used HUMAnN3 as a baseline for metagenomic functional profiling. HUMAnN3 infers EC

numbers from gene family abundance using UniRef databases and is commonly used in microbiome analysis pipelines. While its prediction process is fundamentally different from DeepECTransformer, it provides a strong reference point for evaluating model performance in complex, real-world datasets.

For validation set comparisons, we also refer to performance reported by Carnelian, a prior enzyme annotation tool. Although Carnelian relies on k-mer-based embedding and classification, its usage on the same dataset enables a fair comparison of methodological improvements. Together, these baselines provide context for interpreting DeepECTransformer's strengths in both accuracy and generalizability.

Experiments and Assessments

The inputs for the model will be the amino acid sequences and the outputs will be the EC number.

The DeepECtransformer will first need to be tested by cleaning up the code and recreating the results from the Kim et al.[3] experiment using the available example data. When that is complete, the transformer model will be used with the validation dataset from the Carnelian experiment[7]. The model will use the same metrics used to measure the efficacy of Carnelian and HUMAnN3 (precision, F_1, AUC).

Table 1. Results of Previous Models

Model	Precision	F_1	AUC	Efficiency (CPU-hours)
HUMAnN3	0.97 ± 0.01	0.96 ± 0.05	0.85	52.5 ± 19.2
Carnelian	0.60 ± 0.08[1]	0.74 ± 0.04[1]	0.80	26.4 ± 2.7
DeepECtransformer	0.85 ± 0.10	0.82 ± 0.12	N/A	N/A

Note: The missing values for the DeepECtransformer is not available. We will be measuring this ourselves and use the Precision and F_1 as a reference.

The purpose of using DeepECtransformer is to bridge the gap between the precision of HUMAnN3 and the computational efficiency of Carnelian. Once that has been completed, we can use the transformer model with the benchmark dataset.

Once we have tested the model with the benchmark dataset, we will then start fine-tuning the model to get the most optimal based on precision, F_1, and AUC.

Benchmark Dataset Curation

Construction of the benchmarking dataset began with gold standard assemblies from simulated metagenomes sourced from the critical assessment of microbiome interpretation (CAMI). Specifically, the CAMI II Toy Human Microbiome Project (HMP) benchmarking sets were utilized. This dataset includes gold standard assemblies derived from ten simulated metagenomes from five human body sites: the airways, gastrointestinal tract, oral cavity, skin, and urogenital tract. CAMI provides the ground truth for metagenome taxonomic classification and a strong basis for evaluation of functional profiling given its complex environment. Despite CAMI's strength as a near "out-of-the-box" benchmarking set for metagenome interpretation it still lacks ground truth for functional profiles. Therefore, before implementing CAMI for benchmarking of DeepECtransformer versus current alignment-based methods (i.e. Humann3) gold-standard construction must be performed as first proposed in Franzosa et al., 2018.

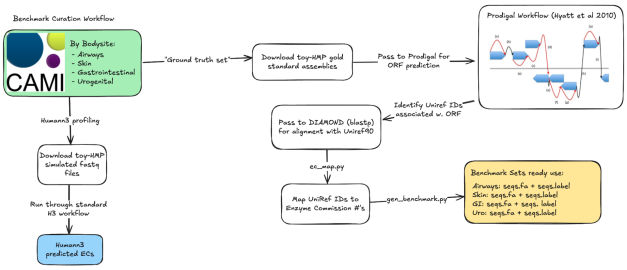

Fig. 2. Benchmark curation overview.

Following data retrieval from the CAMI II toy HMP repository protein coding genes were predicted for each assembly using Prodigal v2.6.3. Given that gold standard assemblies were used we were able to make use of Prodigal's single parameter allowing for single genome context which maintains higher specificity in open reading frame (ORF) prediction.

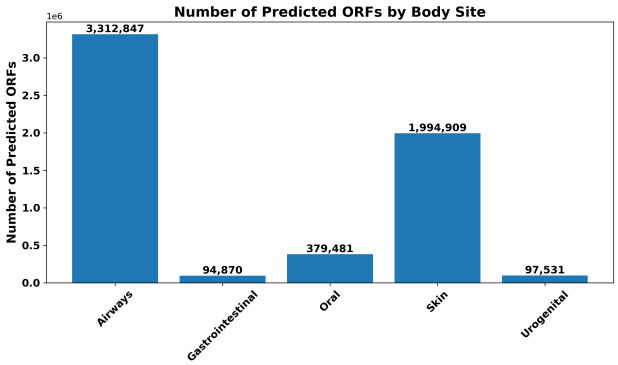

Fig. 3. Prodigal predicted ORFs by body site.

To functionally annotate the predicted open reading frames a homology search was performed with the UniRef90 database (2021.03) using DIAMOND v2.1.9. A DIAMOND database was generated using the UniRef90 FASTA file and the predicted translated genes were queried against it. If there were multiple hits for a single query only the top hit was retained. A mapping file based on UniRef90 accession numbers and enzyme commission (EC) numbers was retrieved from the Humann3 utility mapping database. The EC numbers were parsed with awk to match UniRef90 cluster representatives to their known EC annotations. To ensure high-level accuracy and emulate filtering behavior in alignment-based methods DIAMOND hits were post-processed based to only retain hits with a percent identity of $\geq$ 90%, a query coverage of $\geq$ 80%, and a subject coverage of $\geq$ 80% were retained. To perform this filtering and generate the final benchmarking protein sequences and label files, custom Python scripts were employed. These assigned labels were then used to create two files for use in benchmarking and supervised learning tasks. A seq.fasta file with the protein sequences generated from the predicted ORFs was generated. Another plain text file was generated containing the corresponding EC numbers for each protein sequences in seq.label. This input format corresponds to the validation set used in the initial testing of DeepECtransformer.

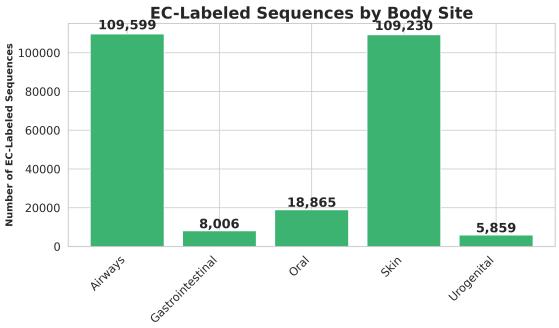

Fig. 4. Counts of final paired EC labeled sequences in benchmarking set by CAMI body site.

The combination of the benchmark curation across all CAMI 2 toy human microbiome project results in 251,559 labeled high confidence EC sequences. This data is available at `doi.org/10.5281/zenodo.15192200` for download. It is important to note that a high proportion of the ORFs predicted by Prodigal did not map to the UniRef90 database with high context and therefore these sequences are not included in the current benchmarking set. This leads into an additional proposed experiment to run DeepECtransformer against the unaligned sequences. This would give a dataset of low homology sequences that may be less biased towards alignment-based methods. A question we received on our initial proposal was if the interpolation process would introduce any bias towards alignment-based methods. While it is possible that this alignment based interpolation may have some bias towards established methods it provides high level of accuracy for ground truth. As such it allows for a gold standard benchmark that is an important initial check for considering ML-based approaches over currently accepted homology approaches. If DeepECtransformer performs well on this set, then focus can be shifted to prediction tasks of unaligned ORFs. This then allows for an initial testing set of DeepECtransformer for de novo prediction and has the potential for much higher recall compared to current alignment-based functional prediction procedures. Plans for running DeepECtransformer against these unaligned sequences has been applied to our project timeline. The simulated nature of the CAMI benchmarking set also provides a data set with far less potential in overlap to the training set of DeepECtransformer making this approximately 250,000 EC sequence testing set a solid benchmark against alignment-based methods.

HUMAnN3 Evaluation on Benchmark Set

To evaluate the baseline performance of HUMAnN3 (v3.9), associated FASTQ samples from the *CAMI-II Toy Human Microbiome Dataset* were retrieved by body site. A standard HUMAnN3 pipeline was executed in parallelized runs for each individual sample. No additional quality control was performed beyond the default HUMAnN3 pipeline. To expedite processing due to large input file sizes, HUMAnN3 was executed with the `--memory-use=minimum` flag.

During pipeline execution, four standard DIAMOND searches were performed for gene family assignment, aligning each sample against the following databases:

`uniref90_201901b_full.dmnd`, `uniref50_201901b_ec_filtered.dmnd`, `uniref50_201901b_full.dmnd`, `uniref90_201901b_ec_filtered.dmnd`.

In most cases, the HUMAnN3 pipeline mapped a median of 80% of reads across all body sites to either the curated ChocoPhlAn pangenome or UniRef clusters.

HUMAnN3 gene families were mapped to level 4 enzyme commission (EC) numbers using the utility mapping script to match the output format of *DeepECtransformer* for downstream comparisons. HUMAnN3-assigned EC numbers were then compared to the precomputed gold-standard EC dataset derived from CAMI reference assemblies. Filtering was applied using HUMAnN3 standard thresholds: a minimum per-sample copies per million (CPM) of 1, and a minimum per-sample fraction of 0.20. These cutoffs normalize for differences in sequencing depth across samples and improve comparability of EC abundances.

Across all body sites, HUMAnN3 showed low precision (mean = 0.07) and low F1-score (mean = 0.12), while maintaining high sensitivity (mean = 0.91). Detailed performance metrics per body site are shown below.

Table 2. HUMAnN3 Per-site Benchmarking Results

Dataset	Precision	Sensitivity	F1
Benchmark(airways)	0.09	0.84	0.16
Benchmark(gastrointestinal)	0.05	0.95	0.10
Benchmark(oral)	0.05	0.96	0.10
Benchmark(skin)	0.10	0.83	0.17
Benchmark(urogenital)	0.05	0.96	0.09

A brief comparison of note is the large number of CPU hours needed to run HUMAnN3 on a large 500 GB dataset such as the CAMI-II toy microbiome dataset, which simulates deep shotgun metagenomic sequencing. Due to HUMAnN3's reliance on multiple alignment searches with Bowtie2 and DIAMOND, it required a per-sample CPU time of approximately 19.2 hours. While a direct comparison to DeepECtransformer is not compatible due to differences in the starting point of analysis (HUMAnN3 begins from FASTQ files, whereas DeepECtransformer begins from ORF sequences), this remains an important consideration and highlights a general limitation of functional profiling through exclusively alignment-based procedures.

Evaluation on Validation and Benchmark Datasets

To evaluate the robustness of DeepECTransformer, we tested the model on a validation dataset derived from the Carnelian resource as well as multiple benchmark datasets representing diverse microbial environments. These datasets include protein sequences with known Enzyme Commission (EC) annotations, allowing for quantitative comparison between predicted and reference labels. As in prior work, performance was evaluated using both micro- and macro-averaged precision, recall, and F1 score to capture class-level and global prediction quality.

Across all datasets, the model achieved consistently high micro-level metrics. As shown in Table4, micro F1 scores on the validation and benchmark datasets exceeded 0.89, indicating strong overall prediction accuracy. These results suggest that DeepECTransformer is highly effective at identifying enzyme functions when predictions are aggregated across all classes.

However, macro-level metrics, which average performance across each class independently, were notably lower. Table3 shows that this discrepancy was especially pronounced in the skin and airways datasets, where the gap between micro

and macro F1 scores exceeded 20 percentage points. This performance gap highlights a key limitation of the model: while it performs well on dominant enzyme classes with ample training data, its performance on rare or underrepresented classes is substantially weaker.

Such bias may be attributed to the long-tailed distribution of enzyme classes in biological data, where common functions are well represented but rare ones are sparse. The high micro scores reflect the model's proficiency in predicting frequent patterns, but lower macro scores suggest poor generalization across the full range of enzyme classes.

To investigate this issue further, we designed a second experiment to assess whether the sequences that were misclassified were intrinsically different from those predicted correctly. By analyzing their sequence-level similarity, we aimed to determine whether the model's prediction errors could be linked to unfamiliar or structurally distinct inputs.

Table 3. Performance (Macro) of DeepECTransformer on the validation and benchmark dataset

Dataset	Macro Precision	Macro F_1	Macro Sensitivity
Validation	0.81	0.81	0.84
Benchmark(uro)	0.79	0.78	0.81
Benchmark(gastro)	0.78	0.79	0.81
Benchmark(oral)	0.75	0.73	0.76
Benchmark(skin)	0.69	0.67	0.70
Benchmark(airways)	0.70	0.69	0.71

Table 4. Performance (Micro) of DeepECTransformer on the validation and benchmark dataset

Dataset	Micro Precision	Micro F_1	Micro Sensitivity
Validation	0.92	0.91	0.90
Benchmark(uro)	0.90	0.89	0.88
Benchmark(gastro)	0.90	0.89	0.89
Benchmark(oral)	0.91	0.91	0.90
Benchmark(skin)	0.91	0.91	0.91
Benchmark(airways)	0.90	0.90	0.90

Similarity Comparison of Correct vs Misclassified Sequences

To better understand the performance imbalance observed in the previous section, we conducted a sequence similarity analysis to compare misclassified proteins with those that were predicted correctly. The goal was to examine whether misclassification was associated with sequences that were substantially different from those the model had successfully classified. In this experiment, similarity was quantified using Jaccard similarity based on 3-mer representations of amino acid sequences.

For each misclassified sequence, we computed its average and maximum Jaccard similarity against all correctly predicted sequences in the same dataset. As shown in Table5, the similarity scores were consistently low across datasets. For instance, in the benchmark oral dataset, the average 3-mer similarity was only 0.0336, and the average maximum similarity was 0.1834. Over 84% of misclassified sequences had a maximum similarity score below 0.2, indicating minimal overlap with the set of correctly predicted sequences.

These findings suggest that misclassified inputs tend to be substantially different at the sequence level compared to what the model has learned to predict accurately. In other words, the model appears to struggle with unfamiliar or atypical sequences, likely due to the limited diversity of its training data. This reinforces the hypothesis that DeepECTransformer performs well on familiar patterns but fails to generalize to rare or novel protein sequences.

Motivated by this observation, we explored whether modifying the encoder could improve generalization to such difficult cases. This led to our third experiment, where we replaced the original ProtBERT encoder with a more expressive alternative, ProtT5.

Table 5. Similarity Comparison of Correct vs Misclassified Sequences

Dataset	Average Similarity	Average Max Similarity
Validation	0.042	0.20
Benchmark(uro)	0.035	0.13
Benchmark(gastro)	0.039	0.13
Benchmark(oral)	0.034	0.18
Benchmark(skin)	0.030	0.10
Benchmark(airways)	0.033	0.12

Encoder Replacement with ProtT5

Based on the similarity analysis in the previous experiment, we hypothesized that the model's limited generalization capability may be due in part to the representational capacity of the encoder. To address this, we substituted the original ProtBERT encoder in DeepECTransformer with ProtT5, a more expressive protein language model pretrained on a broader set of protein sequences.

The ProtT5-based model was retrained from scratch using the same training data and evaluated on the validation and benchmark datasets. Performance results are summarized in Table 6 and Table 7.

In terms of macro performance (Table 6), the ProtT5-based model yielded moderate F_1 scores, ranging from 0.61 to 0.76 across all datasets. The highest macro F_1 was observed on the validation set (0.76), while the lowest scores were seen in the oral (0.61), skin (0.61), and airways (0.64) datasets. Although these scores do not reflect a dramatic improvement over the ProtBERT-based results, the overall distribution is slightly more consistent across datasets, suggesting improved class-level balance.

At the micro level (Table 7), the model achieved high performance in several benchmark datasets, with a micro F1 score of 0.82 in the uro dataset and 0.80 in the gastro, oral, and skin datasets. In contrast, performance on the airways dataset was noticeably lower, with a micro F1 of 0.73. The validation set also exhibited a lower micro F1 of 0.71, indicating that overall prediction confidence may have decreased in this configuration. Nevertheless, the benchmark datasets maintained high sensitivity values ($\geq$ 0.76), highlighting the model's ability to retain good coverage.

One notable observation is that the gap between macro and micro scores was reduced compared to the original encoder. For example, in the skin and airways datasets, where ProtBERT exhibited a difference of over 20 percentage points between macro and micro F_1, the ProtT5 model narrowed the gap to approximately 10–13 points. This indicates a more balanced

prediction performance across enzyme classes, including rare or underrepresented ones.

Overall, while ProtT5 does not lead to universally higher scores, it improves prediction fairness across classes. This suggests that encoder selection plays a key role in balancing accuracy and generalization, especially in settings with long-tailed class distributions.

Table 6. Performance (Macro) of DeepECTransformer on the validation and benchmark dataset with ProtT5 Encoder

Dataset	Macro Precision	Macro F_1	Macro Sensitivity
Validation	0.79	0.76	0.79
Benchmark(uro)	0.74	0.68	0.69
Benchmark(gastro)	0.77	0.71	0.71
Benchmark(oral)	0.70	0.61	0.65
Benchmark(skin)	0.69	0.61	0.63
Benchmark(airways)	0.70	0.64	0.68

Table 7. Performance (Micro) of DeepECTransformer on the validation and benchmark dataset with ProtT5 Encoder

Dataset	Micro Precision	Micro F_1	Micro Sensitivity
Validation	0.66	0.71	0.78
Benchmark(uro)	0.85	0.82	0.79
Benchmark(gastro)	0.85	0.80	0.76
Benchmark(oral)	0.78	0.80	0.82
Benchmark(skin)	0.75	0.80	0.86
Benchmark(airways)	0.64	0.73	0.85

Model Comparison with ECPICK

To see if the DeepECTranformer is the most effective of the model, we wanted to test it using another pretrained model. To do this, we compared it to the ECPICK model. This model is a convolutional neural network consisting of a one-hot encoding layer, three convolutional layers and 1-max pooling, hierarchical layers to learn the hierarchy of the EC number, and an output layer. This model was used because it is a contemporary of DeepECTransformer.

The macro performance (Table 8), like the ProtT5-based model, has a moderate F_1 score, ranging from 0.52 to 0.75, and similar comparative results accross datasets.

Table 8. Performance (Macro) of DeepECTransformer on the validation and benchmark dataset with ProtT5 Encoder

Dataset	Macro Precision	Macro F_1	Macro Sensitivity
Validation	0.78	0.75	0.74
Benchmark(uro)	0.65	0.60	0.59
Benchmark(gastro)	0.65	0.58	0.59
Benchmark(oral)	0.59	0.53	0.52
Benchmark(skin)	0.60	0.54	0.53
Benchmark(airways)	0.55	0.52	0.51

The micro performance (Table 8), on the other hand, did not have as much similarity to the ProtT5-based model. The F_1 score for the validation set did better, but the scores for the benchmark dataset as a whole did worse. There is also a peculiarity that the micro precision, F_1, and sensitivity scores are all the same. This may be because this is a multi-class classification model.

Some of the pros of using ECPICK is that it is easier to set up. While DeepECTransformer took time to keep versions consistent and usable, ECPICK was quickly installed using pip. The time to validate both models was the same. Overall, the results gave more consistency between macro and micro scores, much like the ProtT5-based model.

Some cons is that ECPICK is not able to label some of the sequences, for example, this model did not annotate 915 of the 7884 sequences with in the validation dataset, whereas DeepECTransformer did not annotate 3 sequences. It is worth noting that both scores were poorer than the DeepECTransformer, in general.

Table 9. Performance (Macro) of DeepECTransformer on the validation and benchmark dataset with ProtT5 Encoder

Dataset	Macro Precision	Macro F_1	Macro Sensitivity
Validation	0.81	0.81	0.81
Benchmark(uro)	0.73	0.73	0.73
Benchmark(gastro)	0.71	0.71	0.71
Benchmark(oral)	0.76	0.76	0.76
Benchmark(skin)	0.68	0.68	0.68
Benchmark(airways)	0.64	0.64	0.64

HPRC Environment Setup and Code Adaptation

DeepECTransformer was originally published in 2023, and its codebase reflects the software environment that was standard at that time. As a result, many of its dependencies, such as Python 3.6, CUDA 10.2, and Transformers version 3.5.1, are now outdated and incompatible with current platforms. In our case, the computing environment available to us included newer versions such as Python 3.8 and CUDA 11.3.1, which led to significant compatibility issues. This was especially evident when attempting to run the model on Google Colab, where the system configurations could not be downgraded easily.

To resolve these challenges, we migrated our work to the Grace cluster at High Performance Research Computing (HPRC), which allowed us greater control over the software environment. Using Conda, we created a custom environment that closely replicated the original dependencies used in DeepECTransformer. Because the Grace cluster does not have internet access by default, we manually downloaded and transferred necessary files, such as pre-trained tokenizers required by the Transformers library. With this setup, we were able to successfully run the model using CPU resources.

Enabling GPU execution required additional steps. Since the Grace cluster supports CUDA versions starting from 11.3.1, we had to upgrade both the PyTorch and Transformers libraries to versions compatible with this CUDA version. These upgrades introduced a number of incompatibilities with the existing model code, particularly due to changes in the Transformers API. A summary of the version differences between the original and updated environments is shown in Table 10.

To maintain consistency with the original model behavior, we manually added or adjusted several attributes, including the attention implementation type, the positional embedding type, and the gradient checkpointing setting. In addition, positional embeddings are now included by default in newer versions of the Transformers library. Since the original DeepECTransformer

did not use positional embeddings, we explicitly disabled their effect by setting their weights to zero and preventing them from being updated during training. We also verified that the embeddings contained no active values to ensure that they would not influence the model output. Further modifications were made to the attention layers to define any missing attributes and ensure that all components executed correctly.

Through these efforts, we were able to successfully run DeepECTransformer on both CPU and GPU within a modernized software environment, while preserving the intended functionality of the original model.

Table 10. Comparison of Original and Updated Environment Configurations

Dataset	Original	Ours
Python	3.6	3.8
CUDA	10.2	11.3.1
Transformers	3.5.1	4.2.2
Pytorch	1.7.0	1.12.1

Conclusion

In this work, we sought to benchmark whether transformer-based functional profilers can reliably compare to alignment methods in metagenomic functional annotation. Most significantly, we found that DeepECtransformer, with the ProtBERT encoder, achieved a micro-F_1 ¿ 0.89 across all the datasets (validation and five simulated body site metagenomes), while HUMAnN3 averaged an F_1 of about 0.12 despite high sensitivity. Experiments altering DeepECtransformer's architecture showed promise in altering micro and macro performance. Similarity analysis between correct versus misclassified sequences by DeepECtransformer showed low average 3-mer similarity, indicating the model may struggle with more atypical sequences. Using a transformer model also appeared to be more effective than a CNN (ECPICK), judging by the precision, sensitivity, and F_1 scores. Utilizing DeepECtransformer also showed a precision markedly higher than Carnelian and HUMAnN3, while also having much lower computational utilization compared to alignment-based workflows. From these findings, it appears that transformer models are viable to be used in addition to alignment-based pipelines on well-characterized environments and their subsequent enzymes. Still, DeepECtransformer has significant potential for improvements in the long tail of rare classes in the functional space.

Limitation

Despite these findings, this work does come with some limitations of note. For one, there is potential for the benchmark set to harbor some construction bias towards alignment-based methods, given EC labels are partially interpolated from a DIAMOND alignment step. It is a valid concern that this could undervalue performance for recall on truly novel proteins that may not be adequately represented in reference databases. While CAMI-II simulated datasets allow for ground truth annotation, it is key to note that simulated datasets still may miss some environmental complexity, strain variation, or sequencing artifacts that may be present in real samples. One of the main limitations is that this is an older model that does not have as much support; therefore, getting around problems with Python versions and GPU was a challenge at first. When evaluating the micro scores, the results show that common classes were able to be identified more consistently than rarer classes. However, the rarer classes showed poorer generalizations and are harder to identify. As such, sequence similarity impacts prediction quality. While ECPICK was good for quick analyses, it does not work as well on larger datasets.

Unaligned open reading frames from Prodigal are both a possible limitation and a potential area for future work. In this study, unaligned ORFs were excluded from quantitative benchmarking; however, future work should perform comparative analysis on these genes to survey their potential for novel function. Underrepresentation of these clusters in the UniRef database may also account for the macro-level deficits observed in the DeepECtransformer performance, as it appears to be biased towards more abundant EC classes. This may be due to a lack of appropriate training data associated with rare EC numbers that remain poorly captured in current reference databases.

Future Work

There is more research being done on this subject. One thing we can do is to start using newer versions of the model to keep up to date on this research and be more able to ask for help on current software and techniques. Recent alpha release of HUMAnN4 adds potential to change the landscape of current state of the art alignment based methods. While this work is still in its early stages and has yet to be published changes to database version may consitute substantial changes. Future work should look to assess new methods under similar conditions tested in this work.

One of our experiments, where we replaced the encoder, showed promise in improving classification. There is also promise in doing more experiments using different encoders. We can also try different loss functions and see if that improves the performance.

We can also try augmenting the data and keep validating and training new models. Likewise, we can add more new sequences to annotate. This would make the widen the amount of sequences to be annotated and allow for more consistency in correctly identifying rarer classes.

Overall, we believe that machine learning based methods have the potential to expand upon alignment approaches in functional profiling tasks of the metagenome and future work should seek to synthesize these approaches for best performance.

Appendix

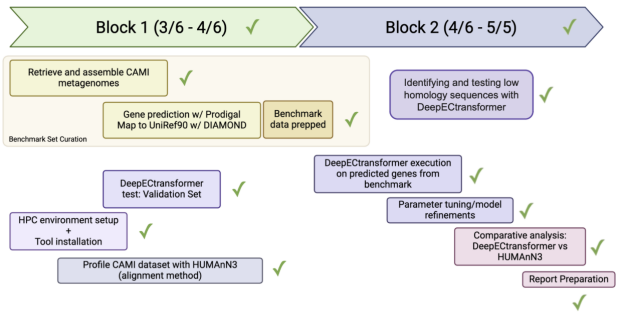

Fig. 5. Final GANTT Execution

Contributions

JT, AG, and JLK jointly conceived the study and designed all experiments. JT carried out the HUMAnN3 benchmarking analyses, curated the datasets, and implemented the associated workflows. JLK performed Experiments 1–3 to evaluate DeepEC and enabled GPU deployment of the DeepECTransformer through model version matching and compatibility fixes, while AG conducted Experiment 4 assessing ECPICK and, together with JLK, supported benchmarking of the DeepEC Transformer model. AG also deployed the DeepEC Transformer on the HPRC infrastructure. JT, AG, and JLK collaboratively drafted the manuscript, analyzed and interpreted the results, and critically revised its content.

References

1. Francesco Beghini, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltez Thomas, Mireia Valles-Colomer, George Weingart, Yancong Zhang, Moreno Zolfo, Curtis Huttenhower, Eric A Franzosa, and Nicola Segata. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *eLife*, (10), 2021.
2. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature*, 486:207–214, 2012.
3. Gi Bae Kim, Ji Yeon Kim, Jong An Lee, Charles J. Norsigian, Bernhard O. Palsson, and Sang Yup Lee. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nature Communications*, 14(7370), 2023.
4. J. Li, J. Wang, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, and P. Bork. An integrated catalog of reference genes in the human gut microbiome. *Nature Biotechnology*, 32:834–841, 2014.
5. Pernille Menzel, Kee Yuan Ng, and Anders Krogh. Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, 7:11257, 2016.
6. Fernando Meyer, Adrian Fritz, Zhi-Luo Deng, David Koslicki, Till Robin Lesker, Alexey Gurevich, Gary Robertson, Mohammed Alser, Dmitry Antipov, Francesco Beghini, Denis Bertrand, Jaqueline J. Brito, C. Titus Brown, Jan Buchmann, Aydin Buluç, Bo Chen, Rayan Chikhi, Philip T. L. C. Clausen, Alexandru Cristian, Piotr Wojciech Dabrowski, Aaron E. Darling, Rob Egan, Eleazar Eskin, Evangelos Georganas, Eugene Goltsman, Melissa A. Gray, Lars Hestbjerg Hansen, Steven Hofmeyr, Pingqin Huang, Luiz Irber, Huijue Jia, Tue Sparholt Jørgensen, Silas D. Kieser, Terje Klemetsen, Axel Kola, Mikhail Kolmogorov, Anton Korobeynikov, Jason Kwan, Nathan LaPierre, Claire Lemaitre, Chenhao Li, Antoine Limasset, Fabio Malcher-Miranda, Serghei Mangul, Vanessa R. Marcelino, Camille Marchet, Pierre Marijon, Dmitry Meleshko, Daniel R. Mende, Alessio Milanese, Niranjan Nagarajan, Jakob Nissen, Sergey Nurk, Leonid Oliker, Lucas Paoli, Pierre Peterlongo, Vitor C. Piro, Jacob S. Porter, Simon Rasmussen, Evan R. Rees, Knut Reinert, Bernhard Renard, Espen Mikal Robertsen, Gail L. Rosen, Hans-Joachim Ruscheweyh, Varuni Sarwal, Nicola Segata, Enrico Seiler, Lizhen Shi, Fengzhu Sun, Shinichi Sunagawa, Søren Johannes Sørensen, Ashleigh Thomas, Chengxuan Tong, Mirko Trajkovski, Julien Tremblay, Gherman Uritskiy, Riccardo Vicedomini, Zhengyang Wang, Ziye Wang, Zhong Wang, Andrew Warren, Nils Peder Willassen, Katherine Yelick, Ronghui You, Georg Zeller, Zhengqiao Zhao, Shanfeng Zhu, Jie Zhu, Ruben Garrido-Oter, Petra Gastmeier, Stephane Hacquard, Susanne Häußler, Ariane Khaledi, Friederike Maechler, Fantin Mesny, Simona Radutoiu, Paul Schulze-Lefert, Nathiana Smit, Till Strowig, Andreas Bremges, Alexander Sczyrba, and Alice Carolyn McHardy. Critical assessment of metagenome interpretation: the second round of challenges. *Nature Methods*, 19(4):429–440, Apr 2022.
7. Sumaiya Nazeen, Yun William Yu, and Bonnie Berger. Carnelian uncovers hidden functional patterns across diverse study populations from whole metagenome sequencing reads. *Genome Biology*, 21(47), 2020.
8. Torsten Thomas, Jack Gilbert, and Folker Meyer. Metagenomics - a guide from sampling to data analysis. *Microb Inform Exp*, 2(1):3, feb 2012.
9. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.