INTERPRETABLE ENZYME FUNCTION PREDICTION VIA RESIDUE-LEVEL DETECTION

Zhao Yang, Bing Su, Jiahao Chen & Ji-Rong Wen
Gaoling School of Artificial Intelligence, Renmin University of China
{yangyz1230, bingsu, 2022000162, jrwen}@ruc.edu.cn

ABSTRACT

Predicting multiple functions labeled with Enzyme Commission (EC) numbers from the enzyme sequence is of great significance but remains a challenge due to its sparse multi-label classification nature, i.e., each enzyme is typically associated with only a few labels out of more than 6000 possible EC numbers. However, existing machine learning algorithms generally learn a fixed global representation for each enzyme to classify all functions, thereby they lack interpretability and the fine-grained information of some function-specific local residue fragments may be overwhelmed. Here we present an attention-based framework, namely ProtDETR (**Prot**ein **De**tection **Tr**ansformer), by casting enzyme function prediction as a detection problem. It uses a set of learnable functional queries to adaptatively extract different local representations from the sequence of residue-level features for predicting different EC numbers. ProtDETR not only significantly outperforms existing deep learning-based enzyme function prediction methods, but also provides a new interpretable perspective on automatically detecting different local regions for identifying different functions through cross-attentions between queries and residue-level features. Code is available at https://github.com/yangzhao1230/ProtDETR.

1 INTRODUCTION

The rapid advancement of genome sequencing has revealed numerous protein sequences, yet their functional annotations remain largely unknown (Acids research, 2021). Due to the time and cost constraints of experimental validation, computational methods for protein function prediction are essential, particularly for enzymes that play crucial roles in metabolic processes. The Enzyme Commission (EC) system classifies enzyme functions using four-level hierarchical numbers (e.g., 1.1.1.1). Enzyme function prediction presents a challenging multi-label classification problem, as each enzyme may catalyze multiple reactions, requiring accurate prediction of a few relevant EC numbers from over 6000 possibilities.

Recent years have seen significant advances in sequence-based enzyme function prediction through deep learning. HDMLF (Shi et al., 2023) leverages multi-sequence alignment with neural networks for EC number prediction, while CLEAN (Yu et al., 2023) addresses data imbalance through contrastive learning, particularly excelling with sparse training instances. Despite these performance improvements, interpretability remains a challenge. ProteInfer (Sanderson et al., 2023) approaches this through CNN-based activation mapping, while EnzBert (Buton et al., 2023) and DeepECtransformer (Kim et al., 2023) utilize Transformer attention mechanisms. However, EnzBert primarily targets mono-enzymes, and DeepECtransformer relies on a conventional multi-classification approach using global protein representations.

These deep learning-based methods actually follow the classification framework, i.e., they generally extract a fixed global representation for each enzyme and feed it into classifiers or compare it with templates of different EC numbers for classification. However, different enzyme functions largely depend on different local structures corresponding to specific active sites or residue fragments. If a local fragment or a weighted combination of residues in an enzyme determines a function, the fine-grained information of that fragment or combination is more effective in identifying that function. For the same enzyme, different functions may depend on different residue fragments and the discrim-

inative fine-grained information is also different. Such fine-grained residue-level or fragment-level information may be overwhelmed in the global protein-level representation. Moreover, gaining the interpretability at the detailed granularity for in-depth analysis becomes infeasible due to the loss of residue-level features.

In multi-label image classification, efforts have been made to learn different features for different classes. Query2label (Liu et al., 2021) introduces unique learnable queries for each class, termed *class queries*, which localize areas relevant to different objects within an image through cross-attention. This approach achieves fine-grained and interpretable modeling for multi-label classification. Nevertheless, this method cannot be directly transferred to multifunctional enzyme annotation due to the quadratic complexity of attention calculations for lengthy enzyme sequences and the vast number of queries needed to cover all EC numbers, resulting in prohibitively high computational demands.

We cast multi-label enzyme function prediction as a detection problem and introduce a novel framework, ProtDETR, by leveraging the advancements in object detection, as exemplified by DETR (Carion et al., 2020). Analogous to detecting all objects of interest in an image and determining their classes and locations, ProtDETR detects all functional residue fragments for determining relevant EC numbers. It preserves the sequence of residue-level features and employs *functional queries* to identify a specific enzymatic function or denote its absence from the sequence. Unlike *class queries* utilized in query2label, where more than 6000 queries are required, ProtDETR only learns 10 *functional queries* as the maximum number of annotated functions for all enzymes in Uniprot is less than 10. In this way, ProtDETR adaptively extracts fine-grained fragment-level representations from residue distributions located by different queries for classifying different functions, while significantly reducing the computational demands.

1. Different from existing methods that encode the enzyme into a fixed representation for multi-label classification, we cast enzyme function prediction as a fine-grained detection problem and propose a novel framework called ProtDETR, which adaptively detects different fragments from all residues of the enzyme by functional queries and extracts fine-grained fragment-level representations for classifying different functions.
2. ProtDETR achieves state-of-the-art (SOTA) results in both multifunctional and monofunctional enzyme prediction tasks. For instance, in the multifunctional enzyme prediction on the New-392 dataset, it not only matches the precision of existing SOTA methods but also improves the recall by 25% and enhances the F1-score significantly.
3. ProtDETR is the first model to offer EC number-specific interpretability. The cross-attention mechanisms between residue-level features and function queries not only illuminate the prediction mechanics of our model but also substantially bolster the reliability and utility of the predictions. This unique capability provides novel insights into the intricate catalytic mechanisms of multifunctional enzymes, propelling further research in enzyme function studies.

2 METHOD

2.1 PROBLEM FORMULATION

Most enzymes are proteins and can be represented by amino acid sequences with length L. We denote such a protein sequence as $S = \{s_1, s_2, \ldots, s_L\}$, where s_i is one of 20 standard amino acids. Our objective is to predict the set of active enzymatic functions for each sequence. Traditionally, this is viewed as a multi-label classification problem with a label vector $\mathbf{y} = \{y_1, \ldots, y_M\}$, where $y_i = 1$ indicates activation of the i-th function and M can be very large (often over 6000). To align with a detection-based perspective, we instead treat each protein's functions as a *set of active function indices* Y. We also introduce a null symbol $\varnothing$ to represent the absence of a function. Let N be the maximum number of functions typically active in a single enzyme (in practice, $N = 10$ is sufficient for most enzymes). When an enzyme has fewer than N active functions, we pad the remaining slots with $\varnothing$. Formally,

$$Y = \{i \mid y_i = 1\} \cup \{\varnothing\}^{N-|\{i|y_i=1\}|}.$$

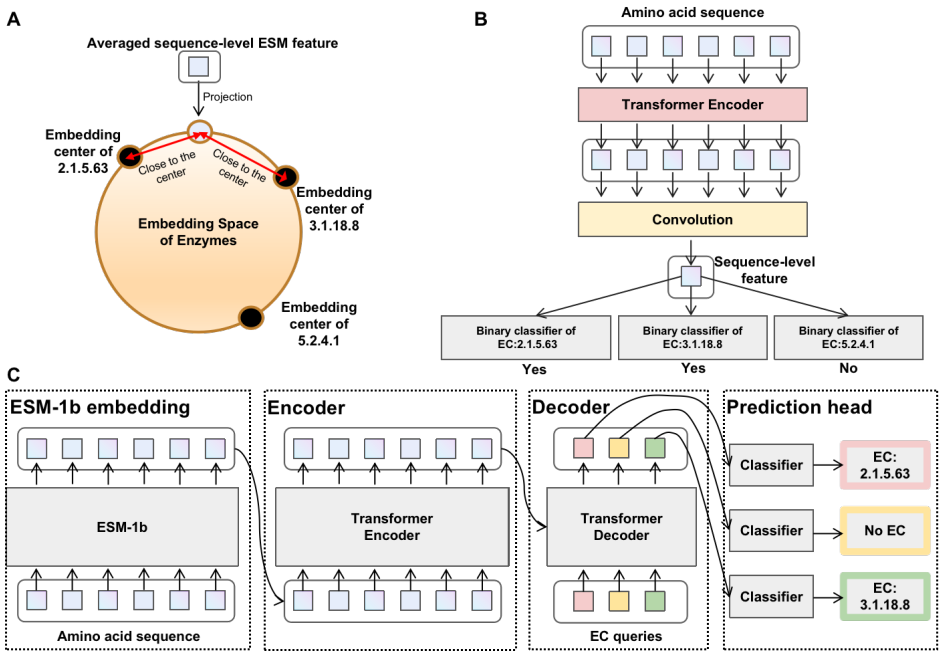

Figure 1: **Three distinct approaches to multifunctional enzyme annotation.** (A) CLEAN: contrastive learning-based approach using sequence-level representations and centroid distances. (B) DeepECtransformer: self-attention mechanism for modeling amino acid interactions. (C) ProtDETR: reformulates enzyme function prediction as a residue-level detection problem using function-specific queries.

Hence, $|Y| = N$ is fixed, which allows us to apply a set-based detection paradigm in an end-to-end manner.

2.2 Overview

Figure 1 contrasts ProtDETR with two prior SOTA methods. CLEAN (Yu et al., 2023) (Figure 1(A)) uses global protein-level representations learned by contrastive learning (Khosla et al., 2020), effectively clustering enzymes by EC numbers in a latent space but lacking residue-level interpretability. DeepECtransformer (Kim et al., 2023) (Figure 1(B)) extracts amino acid features via a Transformer Encoder, then aggregates these features into a single global representation for multi-label classification. Despite partial interpretability using self-attention, it relies on thousands of binary classifiers, one for each EC class, which is challenging to train under highly imbalanced, long-tailed data.

Our ProtDETR (Figure 1(C)) instead re-conceptualizes enzyme-function prediction as a *residue-level detection* task. Leveraging a Transformer encoder-decoder design, we use a fixed number ($N = 10$) of learnable query tokens—each query can capture the fine-grained local signatures of a possible function. Cross-attention between each query and the amino acid sequence selectively attends to crucial local fragments, such as active or binding sites, enabling more accurate and interpretable predictions.

2.3 ESM-1b Embedding

We begin by encoding each protein sequence using the pretrained ESM-1b model (Rives et al., 2021), a large language model trained on billions of protein sequences via masked language modeling (Kenton & Toutanova, 2019). Given a sequence S, we apply ESM-1b to obtain residue-level embeddings: $F_{\text{ESM}} = E(S) \in \mathbb{R}^{L \times d_{\text{ESM}}}$, where $E(\cdot)$ denotes the ESM-1b embedding function. These embeddings have shown strong performance in numerous protein-related tasks (Hsu et al., 2022; Meier et al., 2021; Hu et al., 2022). Consistent with prior work (Yu et al., 2023; Shi et al., 2023), we truncate sequences to $L = 1022$ to fit model constraints.

2.4 DETECTION TRANSFORMER ARCHITECTURE

We adopt a Transformer-based encoder-decoder architecture (Carion et al., 2020), adapted from DETR for enzyme function detection. First, we project F_{ESM} to d_{model} dimensions via a linear mapping, producing feature representations that a multi-layer Transformer *encoder* with M_e layers refines with self-attention:

$$F_{\text{trans}} = \text{Linear}(F_{\text{ESM}}) \in \mathbb{R}^{L \times d_{\text{model}}}, \quad F_{\text{encoded}} = \text{Encoder}\big(F_{\text{trans}}\big).$$

On the decoder side, we introduce N learnable query tokens $Q_{\text{EC}} \in \mathbb{R}^{N \times d_{\text{model}}}$, each representing a potential function. A Transformer *decoder* with M_d layers applies cross-attention between these queries and F_{encoded}, yielding

$$Q_{\text{decoded}} = \text{Decoder}\big(Q_{\text{EC}},\, F_{\text{encoded}}\big) \in \mathbb{R}^{N \times d_{\text{model}}},$$

where each query captures local context relevant to one possible enzyme function. Finally, a linear projection head maps these N decoded embeddings to probability distributions over $C + 1$ classes (i.e., C enzyme functions plus one null class):

$$\mathbf{p}_i = \text{Linear}\big(Q_{\text{decoded}_i}\big) \in \mathbb{R}^{C+1}, \quad \hat{y}_i = \text{argmax}(\mathbf{p}_i),$$

for $1 \leq i \leq N$, where $\hat{y}_i \in \{0, ..., C\}$. The predicted set of enzyme functions is then defined as $\hat{Y} = \{\hat{y}_i\}$, where $\hat{y}_i = C$ indicates that the i-th query detects no enzyme function.

2.5 TRAINING OBJECTIVE

Since enzyme sequences can have multiple functions and the order of our N query predictions is inherently arbitrary, we need to establish a matching between predictions and ground truth labels. Following Carion et al. (2020), we formulate this as a set prediction task and employ the Hungarian algorithm (Kuhn, 1955) to find an optimal one-to-one correspondence by minimizing a global matching cost:

$$\hat{\sigma} \;=\; \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}\big(Y_i,\; \hat{Y}_{\sigma(i)}\big),$$

where $\mathfrak{S}_N$ is the set of all permutations over N elements. Here, $Y_i \in \{1, ..., C\}$ denotes the i-th ground truth label with C representing the null class, and $\hat{Y}_{\sigma(i)}$ is the $\sigma(i)$-th predicted label. The cost $\mathcal{L}_{\text{match}}$ is based on the negative log-likelihood for each ground-truth class.

Once we obtain the optimal assignment $\hat{\sigma}$, we calculate the final cross-entropy loss over matched pairs:

$$\mathcal{L}(Y,\, \hat{Y}) \;=\; \sum_{i=1}^{N} -\log\, \hat{p}_{\,\hat{\sigma}(i)}\big(Y_i\big),$$

where $\hat{p}_{\,\hat{\sigma}(i)}\big(Y_i\big)$ is the predicted probability of the correct class for Y_i under the optimal matching. This ensures that each ground-truth function (including null) is paired with exactly one model prediction, enforcing a one-to-one correspondence in the multi-function setting. To address the extreme long-tail problem in enzyme function prediction, we further discuss the additional loss we used in Appendix E.

2.6 LOCALIZATION OF KEY ENZYMATIC SITES VIA ATTENTION

Residue-level attention scores in ProtDETR provide interpretability. First, in the final layer of the *encoder*, we aggregate the self-attention scores across all heads to find residues of high global importance:

$$\text{AttnAgg}(i) \;=\; \sum_{h=1}^{H} \sum_{j=1}^{L} \text{EncAttn}_h(i, j),$$

where $\text{EncAttn}_h(i, j)$ denotes the self-attention weight from residue i to residue j under head h in the last encoder layer. We visualize these scores to identify potentially significant regions (Appendix Fig. 7(A)).

Table 1: Benchmark Scores on New-392 and Price-149 Datasets

Method	New-392			Price-149		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
DEEPre	-	-	-	0.0415	0.0403	0.0386
ECPred	0.1778	0.0954	0.1000	0.0197	0.0197	0.0197
DeepEC	0.2978	0.2167	0.2297	0.1184	0.0724	0.0846
ProteInfer	0.4088	0.2843	0.3086	0.2434	0.1382	0.1662
DeepECtransformer	0.4268	0.3260	0.3350	0.5263	0.3026	0.3511
BLASTp	-	-	-	0.5083	0.3750	0.3852
CLEAN	**0.5965**	0.4811	0.4988	0.5844	0.4671	0.4947
ProtDETR	0.5943	**0.6083**	**0.5773**	**0.5773**	**0.5066**	**0.5078**

More uniquely, ProtDETR also provides *function-specific* insight via the *decoder* cross-attention. For each query q_{EC} and residue i, we similarly aggregate cross-attention weights from the final decoder layer:

$$\mathrm{QueryAttn}(q_{\mathrm{EC}}, i) \;=\; \sum_{h=1}^{H} \mathrm{DecAttn}_h(q_{\mathrm{EC}}, i),$$

where $\mathrm{DecAttn}_h(q_{\mathrm{EC}}, i)$ represents the cross-attention weight from query q_{EC} to residue i under head h. This reveals where each *individual function* query focuses in the sequence (Appendix Fig. 7(B)). As multifunctional enzymes typically have distinct sites for different activities, these query-specific attention maps provide deeper insight into the roles of particular residues in enabling diverse enzyme functions. These two interpretability methods are further explained in Appendix Figure 7.

3 EXPERITMENT

We evaluated ProtDETR's performance on both multifunctional and monofunctional enzyme benchmarks, demonstrating its superior effectiveness and interpretability. More model implementation details, including model architecture and hyperparameter selection, can be found in Appendix A.

3.1 PROTDETR ENABLES PRECISE MULTIFUNCTIONAL ENZYME CLASSIFICATION

We followed CLEAN's (Yu et al., 2023) experimental setup, training ProtDETR on the split100 dataset (220K instances from SwissProt) and evaluating on two benchmarks: New-392 (392 newly identified sequences) and Price-149 (experimentally verified annotations). Detailed data descriptions are provided in Appendix B.

Before evaluating our model on test sets, we aligned with the benchmarking approach of CLEAN, using MMSeqs2(Steinegger & Söding, 2017) to group data by sequence similarity, with thresholds from 10% to 70%. This led to the creation of datasets: split10, split30, split50, split70, and split100, where split100 represents the original, full dataset. We split each dataset with a ratio of 4:1 for training and testing, respectively, and conducted five-fold cross-validation. The average F1 score across these folds is reported. As shown in Fig 2, ProtDETR's F1 scores are initially lower than CLEAN's at the lower similarity levels (split10 and split30), equal at split50, and surpass CLEAN at the higher levels, with notable scores of 0.9332 versus 0.9163 for split70 and 0.9686 versus 0.9534 for split100. This improvement on larger datasets highlights the advantage of deep learning's data-driven nature, where ProtDETR's fine-grained modeling benefits from more data to enhance its prediction accuracy.

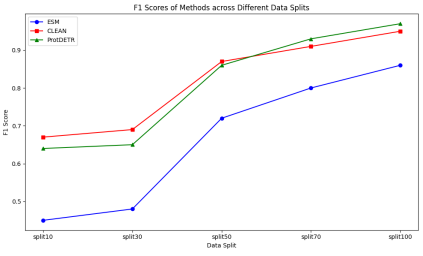

Figure 2: Performances in five-fold cross-validation.

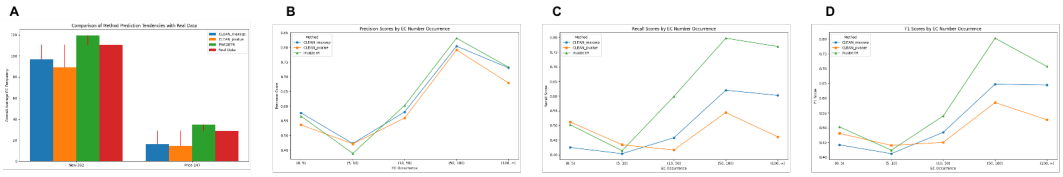

Figure 3: (A) Comparison of average EC number frequencies in the training set as predicted by ProtDETR, CLEAN, and observed in the actual test sets, demonstrating ProtDETR's predictions to more closely align with the true data distribution than those by CLEAN. (B-D) Across various EC occurrence frequencies, ProtDETR demonstrates comparable precision to CLEAN but significantly superior recall for more frequent EC numbers, enhancing its F1 scores.

Following the hyperparameter optimization process as CLEAN, we adjusted the model's hyperparameters during the five-fold cross-validation on the split100 dataset. Once we determined a relatively suitable set of hyperparameters for training, we trained our model on the entire split100 dataset for a fixed number of epochs, in line with CLEAN's approach. We obtained ProtDETR$_{\text{split100}}$ for multifunctional enzyme annotation. We conducted comparisons of our model against the current SOTA methods on the test set, including CLEAN (Yu et al., 2023), DeepECtransformer (Kim et al., 2023), ProtInfer (Sanderson et al., 2023), DeepEC (Ryu et al., 2019), BLASTp, DEEPre (Li et al., 2018), and ECPred (Dalkiran et al., 2018). We evaluated our model using the same metrics as CLEAN, specifically weighted average precision, recall, and F1 score. These metrics helped assess our model's performance in multi-label classification challenges, especially in long-tail scenarios. As shown in Table 1 for the New-392 dataset, our model achieved a precision of 0.5943, comparable to the SOTA performance 0.5965 set by CLEAN, while our recall of 0.6083 was 25% higher than CLEAN's recall of 0.4811. This achievement emphasized our model's capacity for accurately annotating enzymes with potentially undiscovered functions (Poirson et al., 2024). The DeepECtransformer, which also models features at the residue level like ours, performed worse in comparison to ProtDETR and CLEAN. Results for the Price-149 dataset, depicted in Table 1, mirrored those for New-392, with our model matching CLEAN in precision but showing a notably higher recall of 0.5066 compared to CLEAN's 0.4671. These findings underlined the effectiveness of our approach as a tool for the annotation of multifunctional enzymes.

For a deeper analysis of the prediction outcomes, we calculated the average occurrences within the training set of EC numbers predicted by CLEAN, predicted by ProtDETR, and actually present in the test sets. As shown in Figure 3(A)), CLEAN's predictions typically fall below the true average occurrences observed in the test sets. In contrast, while ProtDETR's predictions slightly exceed the actual averages, they align more closely. Specifically, on the New-392 dataset, CLEAN's prediction for the average frequency was 89.40, whereas ProtDETR's was closer to reality at 119.41, with the actual frequency being 110.63. For the PRICE-149 dataset, the average frequencies predicted by CLEAN and ProtDETR were 14.69 and 34.75, respectively, against an actual frequency of 28.81.

To further explore the models' performance across various occurrence ranges, we combined the New-392 and Price-149 datasets to create a mixed dataset. We then divided the EC numbers from this dataset into several groups based on their occurrences within the split100 dataset: 0-5, 5-10, 10-50, 50-100, and over 100. Figure 3 ((B), (C), and (D)) illustrate the precision, recall, and F1 scores for both ProtDETR and CLEAN. Our results indicated that ProtDETR's precision is comparable to that of CLEAN across all examined occurrence intervals. Notably, ProtDETR's recall significantly outperformed CLEAN's in identifying EC numbers occurring more than 10 times. This improvement was also reflected in the F1 scores, underscoring ProtDETR's enhanced ability in handling higher-occurrence, head classes without compromising performance on lower-occurrence, tail classes.

3.2 ProtDETR is also Effective at Monofunctional Enzyme Prediction

Analysis of the data distribution reveals that enzyme function prediction, though framed as a multi-label task, predominantly involves single-labeled samples. As shown in Appendix Figure 6(A), within the split100 dataset, 215,439 out of nearly 220,000 enzymes possess a single EC annotation,

Table 2: Comparative performance on monofunctional enzyme prediction across different EC levels.

Model	Level	F1	Precision	Recall	Accuracy	Classes
ECPred	0	0.769	0.784	0.781	0.769	2
EnzBert	0	0.837	0.874	0.831	0.845	2
ProtDETR	0	**0.873**	**0.877**	**0.871**	**0.875**	2
ECPred	1	0.728	0.691	**0.841**	0.824	6
EnzBert	1	0.604	**0.784**	0.582	0.813	6
ProtDETR	1	**0.731**	0.780	0.716	**0.882**	6
ECPred	2	0.492	0.468	0.579	0.759	51
EnzBert	2	0.629	0.676	0.672	0.781	51
ProtDETR	2	**0.672**	**0.694**	**0.707**	**0.848**	51
ECPred	3	0.496	0.491	0.549	0.727	132
EnzBert	3	0.609	0.625	0.652	0.749	132
ProtDETR	3	**0.644**	**0.657**	**0.666**	**0.816**	132
ECPred	4	0.407	0.431	0.412	0.636	634
EnzBert	4	0.552	0.576	0.562	0.687	634
ProtDETR	4	**0.576**	**0.601**	**0.589**	**0.719**	634

with only 30 enzymes having more than six functions. Similarly, in the combined New-392 and Price-149 datasets (Appendix Figure 6 (B)), 480 out of 569 enzymes are monofunctional.

Given this prevalence of monofunctional enzymes, effective single-function prediction capability is crucial for addressing the enzyme function prediction challenge. While ProtDETR was originally designed for multifunctional annotation with 10 functional queries, we adapted it for monofunctional prediction by simply adjusting the number of queries to 1, while maintaining other hyperparameters consistent with $\text{ProtDETR}_{\text{split100}}$.

A detailed comparative analysis was conducted utilizing the ECPred40 dataset, which consists of a curated collection of monofunctional enzymes and some non-enzymes, assembled by EnzBert (Buton et al., 2023) based on the original dataset introduced by ECPred (Dalkiran et al., 2018). Two evaluations were carried out. The first evaluation was dedicated to distinguishing between enzyme and non-enzyme classifications. In the second evaluation, we focused exclusively on the enzymes present in the testing set, assessing the accuracy of predictions across the first to fourth EC levels. These two types of evaluations are referred to as level 0 and levels 1-4, respectively. The metrics used are macro F1, precision, recall, and accuracy, following EnzBert. We trained on the training set pre-divided by EnzBert (Buton et al., 2023) and selected the model with the best F1 score on the validation set for testing on the ECPred40 test set. The results are shown in Table 2, where level 0 refers to distinguishing between enzymes and non-enzymes.

Remarkably, ProtDETR's performance surpassed that of EnzBert and ECPred across almost all EC levels and metrics, as illustrated in Table 2, except for a comparable precision to EnzBert at Level 1 (0.780 vs 0.784) and a lower recall compared to ECPred at the same level (0.716 vs 0.841). Despite these specific instances, ProtDETR consistently outperformed the benchmarks, averaging an improvement of approximately 0.03 points over the nearest competitor. This signified ProtDETR's precision in not just identifying proteins as enzymes but also in its granular prediction capabilities across both broad and detailed EC categorizations.

3.3 Exploring Prediction Interpretability with ProtDETR

Efforts by recent deep learning models (Sanderson et al., 2023; Buton et al., 2023; Kim et al., 2023) have shown potential in identifying catalytic sites during enzyme function prediction. However, these methods often provided broad, averaged insights rather than precise interpretability. Subsection 2.6 demonstrates how ProtDETR achieves fine-grained interpretability through both its encoder and decoder architecture. The decoder's cross-attention mechanism enables EC number-specific interpretations, allowing identification of functional sites associated with particular enzyme functions.

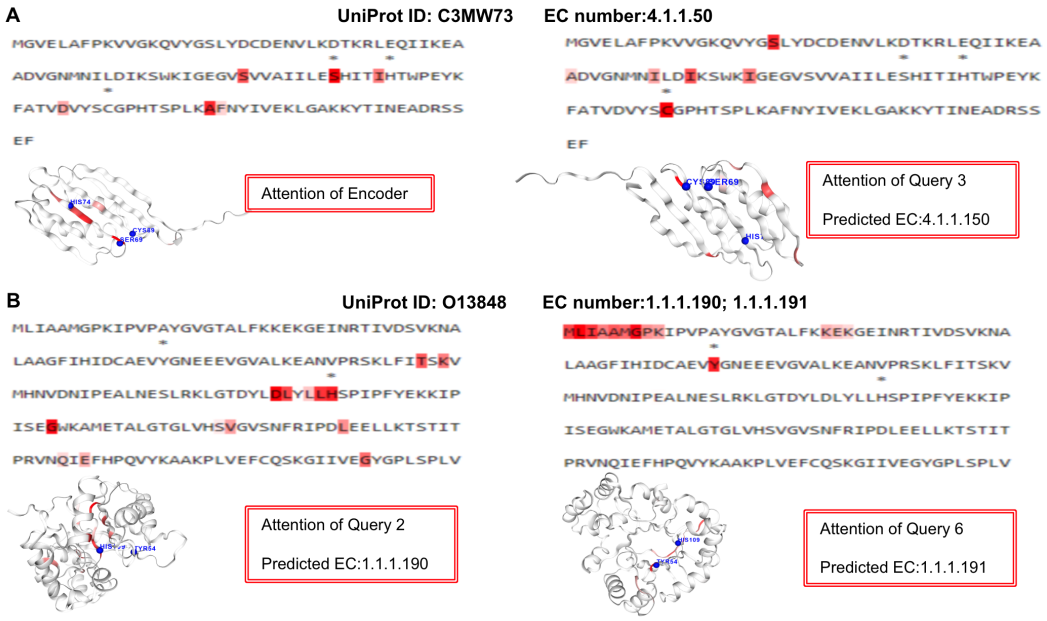

Figure 4: **Visualization of ProtDETR's interpretability**. (A) Monofunctional enzyme C3MW73 showing encoder-query complementarity. (B) Multifunctional enzyme O13848 demonstrating EC-specific attention patterns.

Benchmarking with the M-CSA Database We first quantitatively evaluated our interpretability capabilities on the M-CSA benchmark, with details provided in Appendix D.

Case Study and Visualization. To complement our M-CSA benchmark quantitative analysis, we conducted case studies on multifunctional enzymes from UniProt. Figure 4 (A) and (B) visualize our findings, where sequence importance is shown with the top 5% critical residues in red, and active sites are marked by asterisks and blue spheres in 3D structures. Figure 4 (A) illustrates ProtDETR's attention scores for Uniprot ID C3MW73, associated with EC number 4.1.1.50 and active sites at residues 69, 74, and 89. The left panel showcases the encoder's focus, particularly on active sites 64 and 79, while the right panel highlights query 3's attention, accurately predicting the enzyme's function with an emphasis on active site 89. This case underscores the collaborative identification of all pertinent active sites by both the encoder and the query that makes the accurate prediction. Figure 4 (B) examines Uniprot ID O13848, a multifunctional enzyme linked to EC numbers 1.1.1.190 and 1.1.1.191, with active sites at 54 and 109. The left panel reveals query 2's attention, correctly inferring EC 1.1.1.190 with a focus on active site 109, whereas the right panel shows the attention of query 6, predicting EC 1.1.1.191 and concentrating on the 54. Different queries have predicted their respective correct enzymatic functions, and these two queries focus on different active sites, suggesting that ProtDETR may possess EC-number specific interpretability.

4 Conclusion

This study presents ProtDETR, a novel detection-based framework. By treating enzyme function prediction as a detection problem, ProtDETR not only surpasses existing methods in performance but also provides EC number-specific interpretability through its unique approach. This methodological innovation demonstrates the potential of detection frameworks in computational biology and opens new avenues for understanding enzymatic processes, contributing significantly to the broader field of enzyme research.

MEANINGFULNESS STATEMENT

Enzymes are fundamental molecular machines that drive virtually all biological processes in living systems. Understanding their functions is crucial for comprehending life itself. Our work contributes to meaningful representation of life by developing an interpretable way to predict enzyme functions from their sequences. Unlike black-box approaches, ProtDETR provides insights into how specific protein regions contribute to different biological functions, mirroring nature's modular design principles. This interpretability helps bridge the gap between sequence and function, advancing our understanding of life's molecular machinery in a mechanistically meaningful way.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, 2020.

N Acids research. Uniprot: the universal protein knowledgebase in 2021. *Nucleic acids research*, 49:D480–D489, 2021.

SF Altschul, W Gish, W Miller, and EW Myers. Lipman. dj (1990) j. *Mol. Biol*, 215:403–410.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Nicolas Buton, François Coste, and Yann Le Cunff. Predicting enzymatic function of protein sequences with attention. *Bioinformatics*, 39(10):btad620, 2023.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 782–791, 2021.

Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3339–3348, 2018.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2988–2997, 2021a.

Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1601–1610, 2021b.

Alperen Dalkiran, Ahmet Sureyya Rifaioglu, Maria Jesus Martin, Rengul Cetin-Atalay, Volkan Atalay, and Tunca Doğan. Ecpred: a tool for the prediction of the enzymatic functions of protein sequences based on the ec nomenclature. *BMC bioinformatics*, 19:1–13, 2018.

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021.

Peter Flach and Meelis Kull. Precision-recall-gain curves: Pr analysis done right. *Advances in neural information processing systems*, 28, 2015.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *ICML*, 2022. doi: 10.1101/2022.04.10.487779.

Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang Ding. Exploring evolution-aware &-free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 2022.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19702–19712, 2023.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2, 2019.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

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(1):7370, 2023.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Yu Li, Sheng Wang, Ramzan Umarov, Bingqing Xie, Ming Fan, Lihua Li, and Xin Gao. Deepre: sequence-based enzyme ec number prediction by deep learning. *Bioinformatics*, 34(5):760–769, 2018.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2018.

Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303, 2021.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Juline Poirson, Hanna Cho, Akashdeep Dhillon, Shahan Haider, Ahmad Zoheyr Imrit, Mandy Hiu Yi Lam, Nader Alerasool, Jessica Lacoste, Lamisa Mizan, Cassandra Wong, et al. Proteome-scale discovery of protein degradation and stabilization effectors. *Nature*, pp. 1–9, 2024.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and catalytic site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic acids research*, 46(D1):D618–D623, 2018.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Jae Yong Ryu, Hyun Uk Kim, and Sang Yup Lee. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proceedings of the National Academy of Sciences*, 116(28):13996–14001, 2019.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. Proteinfer, deep neural networks for protein functional inference. *Elife*, 12:e80942, 2023.

Zhenkun Shi, Rui Deng, Qianqian Yuan, Zhitao Mao, Ruoyu Wang, Haoran Li, Xiaoping Liao, and Hongwu Ma. Enzyme commission number prediction and benchmarking with hierarchical dual-core multitask learning framework. *Research*, 6:0153, 2023.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J Haunsberger, and Johannes Söding. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics*, 20:1–15, 2019.

F Altschul Stephen. Basic local alignment search tool. *J. mol. Biol.*, 215:403–410, 1990.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023.

Chengxin Zhang, Peter L Freddolino, and Yang Zhang. Cofactor: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic acids research*, 45(W1):W291–W299, 2017.

Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P Xing. Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

A HYPERPARAMETER SELECTION STRATEGY

The hyperparameter configurations for our model were primarily influenced by the guidelines proposed in DETR Carion et al. (2020), tailored to meet the specific demands of our enzyme function prediction task. DETR originally suggested using 6 encoder layers, 6 decoder layers, and 8 attention heads, with a inference threshold set at 0.7. Given the flexibility of adjusting thresholds during the testing phase, our initial focus was on fine-tuning the encoder and decoder layers as well as the number of attention heads. We evaluated three different configurations: (3, 3, 4) and (6, 6, 8) – the standard DETR setup.

Our assessments demonstrated that models with fewer parameters generally outperformed their more complex counterparts, likely due to the prolonged training epochs required for larger models to achieve convergence. Comparative results on split100's 5-fold cross-validation for the configurations (3, 3, 4) and (6, 6, 8) are presented in Table 3 and Table 4, respectively. We conducted a total of one hundred training epochs on our model. It is evident that the configuration (3, 3, 6) consistently outperformed (6, 6, 8) across all epochs and inference thresholds, also requiring fewer epochs to converge while using fewer parameters. Consequently, we have adopted the (3, 3, 4) configuration for our final hyperparameter settings in the task of multifunctional enzyme annotation.

Table 3: F1 scores for configuration (3,3,6) across different inference thresholds and epochs

Infer Threshold	20 Epochs	50 Epochs	100 Epochs
0.5	0.9486	0.9543	0.9571
0.7	0.9551	0.9577	0.9606
0.9	0.9617	0.9623	0.9645
0.99	0.9600	0.9678	0.9686

Table 4: F1 scores for configuration (6,6,8) across different inference thresholds and epochs

Infer Threshold	20 Epochs	50 Epochs	100 Epochs
0.5	0.9331	0.9225	0.9629
0.7	0.9296	0.9198	0.9576
0.9	0.9296	0.9197	0.9505
0.99	0.9296	0.9197	0.9387

Table 5: Configurations of Model HyperParameters for Different Training Sets

	$\text{ProtDETR}_{\text{split100}}$	$\text{ProtDETR}_{\text{ECPred40}}$
Encoder Layer	3	3
Decoder Layer	3	3
Num heads	4	4
Number of Queries	10	1
Hidden dim	256	256
FFN DIM	2048	2048
Dropout	0.1	0.1
Learning Rate	1e-4	1e-4

This training culminated in the deployment of our $\text{ProtDETR}_{\text{split100}}$ on the complete split100 dataset. We adopted a fixed epoch strategy similar to that used in CLEAN Yu et al. (2023), which trained on the full split100 dataset for 7000 epochs. However, we limited our training to 50 epochs dedicated to multifunctional prediction. This decision was based on the observation that, by the 50th epoch, our model's performance in five-fold cross-validation on the split100 dataset had already significantly surpassed that of CLEAN, with no substantial gains from additional training.

A pivotal adjustment was made to the Number of Queries parameter, predicated on our analytical insight. Given DETR's framework, where an image is unlikely to host more than 70 objects, thus

setting 100 queries, we reasoned that an enzyme with up to 7 or 8 functions warrants 10 queries by analogy. This foundational setting was preserved in ProtDETR$_{\text{split100}}$. For ProtDETR$_{\text{ECPred40}}$, tailored to monofunctional enzyme prediction, we adjusted the query count to one, reflecting the task's monofunctional focus. Due to the smaller dataset size of ECPred40, we trained the model for a total of 20 epochs. During the validation phase of ProtDETR$_{\text{ECPred40}}$, peak performance was observed at the 17th epoch, which guided our decision to finalize the model at this stage.

Table 5 delineates the specific hyperparameters employed for each model version, detailing our tailored approach to enzyme function prediction across different training sets.

B Details of Multifunctional Enzyme Classification Dataset

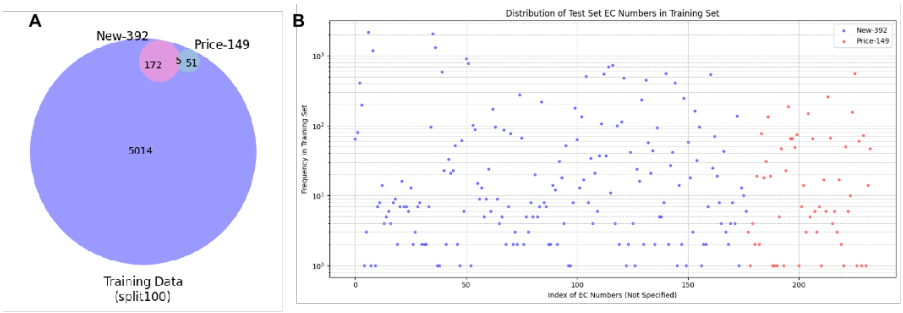

Figure 5: (A) A Venn diagram illustrates the overlap of EC numbers between the training and test sets, highlighting shared and unique annotations. (B) The frequency of EC numbers in the New-392 and Price-149 test sets reveals a long-tailed distribution, with New-392 encompassing a wider array of head classes.

To evaluate the prediction capabilities of ProtDETR for multifunctional enzymes, we adopted the experimental setting used by CLEAN Yu et al. (2023). Specifically, ProtDETR was trained on the split100 dataset, which is composed of about 220K instances from the expert-reviewed SwissProt section of the UniProt database Acids research (2021). Each instance is labeled with one or more EC numbers. The model's performance was tested using two benchmarks: New-392 and Price-149. The New-392 dataset includes 392 enzyme sequences corresponding to 177 unique EC numbers, extracted from the SwissProt version released after CLEAN's training in April 2022. This dataset simulates a realistic scenario by training the model on historical SwissProt data and then predicting the functions of newly identified sequences. The Price-149 dataset, curated by ProteInfer Sanderson et al. (2023), contains experimentally verified annotations and poses a challenge due to inconsistencies or errors in annotations by other automatic annotation methods, such as those found in KEGG Kanehisa & Goto (2000).

The relationship between EC numbers in both the training and test sets is depicted in Figure 5(A), showing that each test set has a low overlap with the training set. Additionally, New-392 and Price-149 share only five common EC numbers, suggesting these datasets effectively test the model's ability to predict across a diverse range of EC numbers. Figure 5(B)illustrates the EC numbers appearing in the test sets, with the vertical axis indicating their frequency of occurrence in the training set. Each test set contains a mix of both highly frequent (head) categories, defined as EC numbers observed more than a hundred times in the training set, and infrequent (tail) categories, for example, those seen only once. Notably, only New-392 includes instances of extremely frequent categories, with occurrences exceeding a thousand times.

C More Intuition for Monofunctional Enzyme Prediction

Analysis of the data distribution reveals a clear pattern in enzyme functionality. As shown in Figure 6 (A), in the split100 training dataset, the vast majority (215,439 out of 220,000) of enzymes are annotated with exactly one EC number. Only a small fraction has multiple functions, with 10,583 having two EC numbers and merely 30 enzymes having more than six functions.

This strong bias towards monofunctional enzymes is also evident in our test sets. Figure 6 (B) shows that in the combined New-392 and Price-149 datasets, 480 out of 541 enzymes are monofunctional, with only 61 enzymes having multiple functions. This consistent pattern underscores the importance of accurate single-function prediction in enzyme annotation tasks.

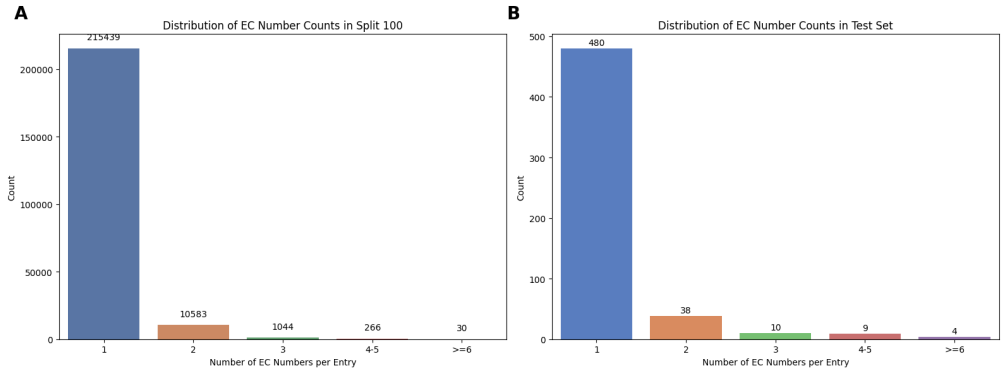

Figure 6: (A) The distribution of EC numbers per enzyme in the split100 dataset indicates that a large majority of enzymes are annotated with a single EC number. (B) Statistics from the combined test sets of New-392 and Price-149 highlight the prevalence of monofunctional enzymes.

D More Interpretability Experiments

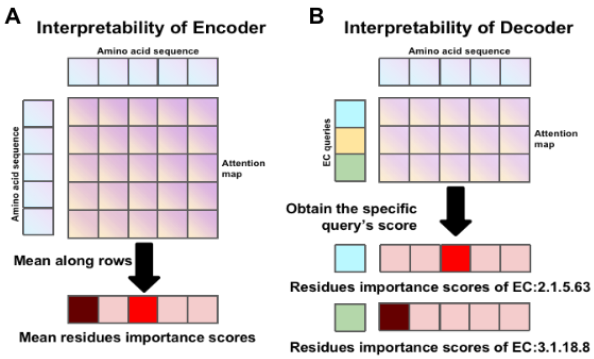

Figure 7: **Interpretability Capabilities of ProtDETR:** (A) The encoder employs "Attn Agg" to assess the average attention each residue receives within the self-attention mechanism, highlighting general areas of interest. (B) "Query Attn" is utilized, where the decoder employs cross-attention from each query to a residue, calculating unique focus levels of each query on different sites, providing EC number-specific interpretations.

In a quantitative comparison, our study benchmarked ProtDETR against EnzBert, selected due to the lack of quantitative interpretability experiments from DeepECtransformer. We utilized the M-CSA (Mechanism and Catalytic Site Atlas) database (Ribeiro et al., 2018). This database documented 992 enzymes and their active sites. We applied PRG-AUC and maximum F-Gain score (Max F-Gain) metrics (Flach & Kull, 2015), aiming to showcase our model's precision in highlighting enzymatic functionalities and its ability to pinpoint catalytically significant residues. In simple terms, our model assigned an importance score to each residue, representing the probability of each residue being predicted as an active site. These scores were then compared with the actual labels of active sites to calculate performance metrics.

Table 6 showcases the interpretability comparison between ProtDETR and EnzBert, with a baseline established by randomizing token importance scores to evaluate effectiveness beyond random chance. EnzBert utilized a variety of interpretability methods, including neural network gradient-

Table 6: Quantitative assessment of enzyme active site recognition capabilities on the M-CSA dataset benchmark.

Backbone	Method Type	PRG-AUC ($\times$100)	Max F-Gain (%)
Random	-	42.54 ± 4.37	69.85 ± 1.04
EnzBert_SwissProt21	Grad	75.01	81.27
	Grad $\times$ Input	63.62	78.66
	Integrated Grad	76.41	81.70
	Attn Last Layer	87.80	85.62
	Rollout	66.08	76.77
	TGLRP	90.92	88.56
	TGradCam	81.00	82.77
	Attn Agg	**98.02**	**96.05**
ProtDETR_SwissProt21	Attn Agg	<u>96.50</u>	<u>94.54</u>
ProtDETR_split100	Attn Agg	89.17	88.28
	Query Attn	83.08	82.30

based techniques (such as Grad, Grad X Input, Integrated Grad (Sundararajan et al., 2017), TGradCam (Chefer et al., 2021)) and attention mapping approaches (like Rollout (Abnar & Zuidema, 2020), TGLRP (Chefer et al., 2021)). Specifically, Attn Last Layer refers to analyzing attention from the last layer of the Transformer Encoder, whereas Attn Agg involves averaging attention analysis across all layers of the Transformer Encoder. EnzBert's evaluation found that Attn Agg was the most effective method.

To ensure a fair comparison, $\text{ProtDETR}_{\text{SwissProt21}}$ was developed using the SwissProt21 dataset, the same dataset that EnzBert used for this task. This dataset contained approximately 500K protein sequences, including both enzymes and non-enzymes. Similarly, since this dataset comprised only monofunctional enzymes, we restricted the model to a single functional query. We adopted the Attn Agg method (shown in Figure 7(A)) for active sites scoring. As shown in Table 6, $\text{ProtDETR}_{\text{SwissProt21}}$ yielded a PRG-AUC(X 100) of 96.50 and a Max F-Gain of 94.54%, indicating that the interpretability of our model's encoder closely matched that of EnzBert. Furthermore, this significantly surpassed the performance of random guessing for active site prediction.

To investigate the interpretability capabilities of ProtDETR's decoder, we evaluated $\text{ProtDETR}_{\text{split100}}$ focusing on its process for classifying multifunctional enzymes. We employed a Query Attn mechanism (shown in Figure 7(A)), which allocated importance scores based on each query's cross-attention scores to discern residue significance. However, given that the M-CSA dataset comprised monofunctional enzymes, we opted to consider only the query with the highest confidence in EC number prediction among all queries, presuming this query most likely reflected the correct prediction. Averaging attention across all queries significantly diminished performance, as most queries did not predict an EC number. Despite being trained on a dataset notably smaller than EnzBert's SwissProt21, $\text{ProtDETR}_{\text{split100}}$ exhibited commendable interpretability. It outperformed EnzBert's interpretative techniques in 6 out of 8 cases with the Attn Agg method and in 3 out of 8 with Query Attn. This indicated the decoder of ProtDETR also had the capability of localizing active sites.

E Addressing the Long-Tailed Challenge

The long-tailed distribution of enzyme classes in training data presents a significant challenge in multifunctional enzyme annotation, as depicted in Figure 8. Each EC number level exhibits a pronounced long-tailed distribution. Additionally, non-enzymatic instances significantly outnumber specific enzyme classes, exacerbating the class imbalance. In machine learning, such long-tail distributions typically result in significant performance degradation, as models tend to favor the more frequent classes at the expense of rare classes. To address this issue, we explore three distinct strategies to mitigate the effects of class imbalance and enhance the model's performance across all classes.

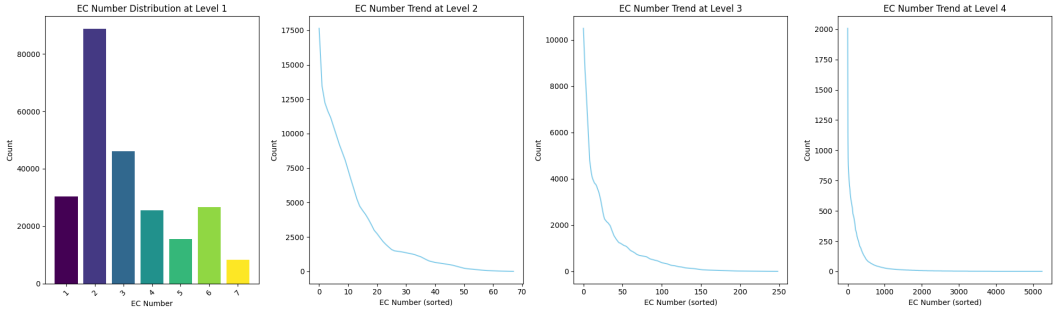

Figure 8: **Long-tailed distribution in the split100 dataset, observable across all four hierarchical levels of EC numbers.** This widespread pattern underscores the significant challenge of class imbalance in enzyme function prediction.

Baseline Adjustment (BA): Following DETR, we apply a constant weight to non-enzyme classes within our loss function to mitigate their predominance due to high occurrence rates. However, this method does not directly address the imbalance within enzyme classes. The weight w is a hyperparameter, set to 0.1 based on DETR guidelines.

$$\mathcal{L}_{\text{BA}} = -\sum_{i=1}^{N} \left(w \cdot \mathbb{1}_{\{Y_i = \varnothing\}} + \mathbb{1}_{\{Y_i \neq \varnothing\}} \right) \cdot \log(\hat{p}_i) \tag{1}$$

Inverse Frequency Reweighting (IFR): This method corrects class imbalance by adjusting weights inversely proportional to class frequencies, thus enhancing the influence of rarer classes. The weight for each class w_c is set inversely proportional to its frequency in the training dataset, enhancing the model's attention to rarer classes.

$$\mathcal{L}_{\text{IFR}} = -\sum_{i=1}^{N} \frac{1}{\text{freq}(c_i)} \cdot \log(\hat{p}_i) \tag{2}$$

Logit Adjustment (LA): Logit Adjustment is a straightforward yet effective technique commonly employed in long-tail image classification tasks to enhance model performance across diversely distributed classes Menon et al. (2020). By adding the logarithm of class frequencies to the logits, this method increases the difficulty of learning for high-frequency classes, aiming to balance the training across different class distributions. This logit modification forces the model to pay more attention to less frequent classes, potentially reducing the dominance of frequent classes in the loss function:

$$\text{adjusted logits} = \text{logits} + \log(\text{freq}(c_i)) \tag{3}$$

After adjusting the logits, the standard cross-entropy loss is applied.

We conducted experiments using the naive cross-entropy method (CE). We evaluated these methods on a 4:1 cross-validation using the split70 dataset, with performance based on the F1 score, as shown in Table 7. The CE and BA methods proved to be the least effective, with F1 scores of only 0.8903 and 0.8901, respectively. This is likely because these methods did not account for the long-tailed distribution of different classes, leading to poor results. Conversely, the IFR approach outperformed the others with an F1 score of 0.9332 and was thus adopted. The third methodology, LA, while effective in long-tailed image classification tasks, was less effective in our context, yielding an F1 score of 0.9242. We speculate that datasets tailored for long-tailed image classification, often engineered with standardized declining power-law frequencies, might be ill-suited for the multifunctional enzyme classification task. This highlights the need for novel long-tail handling strategies, a promising avenue for future research.

F Related Work

Enzyme Function Prediction. Traditional methods for enzyme function prediction are categorized into three main approaches: sequence-based (Altschul et al.; 1997), homology-based (Steinegger

Table 7: Performance comparison of the proposed methods for solving the long-tailed distribution based on F1 score.

Method	F1 Score
CE	0.8903
BA	0.8901
IFR	**0.9332**
LA	0.9242

et al., 2019), and structure-based (Zhang et al., 2017). While each method has its merits, challenges in obtaining accurate homology information and enzyme structures make sequence-based predictions particularly appealing. Classical sequence-based methods, such as BLASTp (Altschul et al., 1997; Stephen, 1990), predominantly derive function annotations based on sequence similarity. However, these methods can yield unreliable predictions when sequence similarity is low. Moreover, traditional methods often suffer from slow processing speeds and suboptimal accuracy, particularly in the context of large and diverse enzyme datasets. This has driven the need for more efficient and accurate computational strategies in enzyme function prediction.

With advancements in the deep learning community, several innovative methods have emerged for enzyme function prediction. For instance, DeepEC (Ryu et al., 2019) and ProteInfer (Sanderson et al., 2023) utilize convolutional neural networks (CNNs) for this task. However, their performance is hampered by CNNs' inherent limitations in capturing long-range dependencies within protein sequences. More recent innovations, such as DeepECtransformer (Kim et al., 2023) and EnzBert (Buton et al., 2023), employ Transformer Encoders to achieve residue-level granularity and offer a degree of interpretability through attention mechanisms. Despite these advancements, their straightforward approach of managing thousands of binary classifications for different enzymes struggles with the pronounced long-tail distribution of enzyme functions. The recent method CLEAN (Yu et al., 2023) represents the SOTA, applying supervised contrastive learning (Khosla et al., 2020) to global features extracted from ESM-1b (Rives et al., 2021). This method has shown notable improvements, especially in recognizing enzymes from tail classes. However, it only models global representations coarsely and lacks detailed interpretability. Similarly, the recent GRU-based (Chung et al., 2014) method HDMLF (Shi et al., 2023) also models using global features derived from ESM-1b, encountering similar limitations in terms of coarse representation and lack of detailed interpretability.

Our proposed ProtDETR innovatively transforms the multi-label classification challenge into an object detection task. This approach not only effectively addresses the issues inherent in sparse multi-label classification but also significantly enhances the model's granularity and interpretability, marking a paradigm shift in how enzyme functions are predicted.

Detection Transformer. Object detection has long been a fundamental task in computer vision, traditionally characterized by complex, multi-stage pipelines (Ren et al., 2015; Chen et al., 2018; Lin et al., 2018; Fan et al., 2021). The introduction of DETR (Carion et al., 2020) brought about a paradigm shift with its end-to-end Transformer Encoder-decoder architecture, which simplified the detection pipeline and achieved remarkable results. Following DETR's innovations, subsequent works (Dai et al., 2021b;a; Jia et al., 2023; Zhang et al., 2022) have proposed various improvements to enhance its efficiency and accuracy. Given the structural similarities between multi-functional enzyme prediction and object detection — where the total number of EC numbers mirrors the extensive pool of potential bounding boxes in an image, but the actual enzyme functions resemble the sparse bounding boxes found in most images — we adapt the DETR framework to address the inherent challenges of sparse multi-label classification in enzyme function prediction.

Our initial inspiration came from the multi-label image classification task approach in query2label (Liu et al., 2021), where each class is represented by a unique, learnable query token, termed *class queries*, facilitating fine-grained multi-label classification through cross-attention between image patches and class queries. Due to the extensive domain of enzyme functions, directly applying this method to multi-functional enzyme prediction is not feasible because of the prohibitive computational demands and the quadratic complexity involved in attention calculations for lengthy enzyme sequences. Therefore, we adopted an approach analogous to DETR, modifying

it by replacing a large number of class queries with a manageable number of function queries. This adaptation not only reduces computational complexity but also ensures detailed and interpretable modeling suitable for the sparse nature of enzyme functions, thus mirroring the fine-grained task-specific adjustments seen in object detection frameworks.

G DETECTION VS. MULTI-LABEL CLASSIFICATION

The impetus for integrating a detection-oriented methodology within ProtDETR for the annotation of multifunctional enzymes stems from the complex nature of enzyme function prediction, especially when addressing enzymes with multiple functionalities. Inspired by the DETR framework from computer vision, ProtDETR adopts object detection principles to surpass the constraints of conventional methods.

Traditional techniques for enzyme function prediction frequently falter in sparse multi-label learning scenarios. ProtDETR addresses this challenge by re-envisioning the prediction of enzyme functions as a detection task, employing an array of learnable functional query tokens to bolster classification accuracy.

In departure from the class queries of standard multi-label classification (Kim et al., 2023; Shi et al., 2023), ProtDETR utilizes functional queries, each uniquely crafted to discern a specific enzymatic function. This approach not only alleviates computational burdens but also enhances the interpretability of the model. ProtDETR prioritizes enzyme functions pertinent to a given protein sequence, thereby improving prediction efficiency and enabling a more precise EC number-specific analysis, which enriches our comprehension of enzyme functions.

The DeepECtransformer, a recent approach (Kim et al., 2023), employs a traditional multi-classification framework, where a single protein descriptor informs multiple binary classifiers. Its efficacy, however, was found wanting on the NEW-392 and PRICE-149 benchmarks set by CLEAN (Yu et al., 2023), especially due to its non-use of ESM-1b embeddings (Rives et al., 2021). To discern if the multi-classification enhancements credited to CLEAN are linked to the ESM-1b embeddings, we restructured our codebase to a conventional multi-classification layout. In this version, named Protein Multifunctional Enzyme Classification (ProtMC), we removed the decoder from ProtDETR and instead trained a multitude of binary classifiers, each representing an EC number, with the encoder's averaged feature outputs. The focal loss technique was applied to counter the long-tail skew of binary classification (Lin et al., 2017). The comparative performance results for ProtMC are shown in Table 8.

The ProtMC variant's performance lagged significantly on both benchmarks, an outcome attributable to the onerous task of training 5200 independent binary classifiers, as necessitated by the split100 category's pronounced long-tail distribution. The abundance of classes engendered data scarcity for numerous classifiers, thus impeding the model's capacity to forge potent decision boundaries.

Table 8: Comparative results of Detection (ProtDETR) vs. Multi-Label Classification (MLC) on the NEW-392 and PRICE-149 benchmarks.

Method	**NEW-392**			**PRICE-149**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
ProtDETR (Detection)	0.5943	0.6083	0.5773	0.5773	0.5066	0.5078
ProtMC (MLC)	0.4596	0.4414	0.4316	0.3307	0.3224	0.3133