BIOLOGICALLY INFORMED DATA AUGMENTATION FOR IMPROVING AI-DRIVEN ENZYME FUNCTION PREDICTION

by

SHREYASH PATEL

(Under the Direction of Adrienne Hoarfrost)

ABSTRACT

Earth harbors vast microbial genetic diversity, yet AI-driven functional prediction remains challenging due to underrepresentation in functional reference databases and severe class imbalance among 2,200 Enzyme Commission (EC) classes. This project tests three data augmentation methods to increase underrepresented EC classes: (1) reverse-complement (doubling 150,000 training samples), (2) synonymous codon substitution (generating 600,000 sequences with 25–70% replacement probability), and (3) conditional GAN generation conditioned on GC content and codon frequency. We created class-balanced training datasets and trained a classifier using a pretrained DNA encoder, LookingGlass, with a 1D convolutional neural network (CNN) decoder. Model performance was evaluated using micro- and macro-averaged F1 scores. Experiments revealed that codon substitution significantly improved macro-F1 (from 0.15 to 0.23) and rare-class recall (from 33.42 to 38%), while reverse complementation degraded performance by introducing label noise. GAN-based augmentation yielded marginal gains without filtering. This work develops a complete training system, evaluation framework, and benchmark

datasets to enhance AI-driven functional annotation of DNA sequences across Earth's diverse microbial communities.

INDEX WORDS: Microbial dark matter, Enzyme Commission (EC) classes, Class imbalance, Data augmentation, Synonymous codon substitution, Reverse-complement generation, LookingGlass embeddings, Rare-class recall, Long-tail distribution, Functional annotation.

BIOLOGICALLY INFORMED DATA AUGMENTATION FOR IMPROVING AI-DRIVEN ENZYME FUNCTION PREDICTION

by

SHREYASH PATEL

B.Tech Savitribai Phule Pune University, India, 2023

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2025

BIOLOGICALLY INFORMED DATA AUGMENTATION FOR IMPROVING AI-DRIVEN ENZYEME FUNCTION PREDICTION

by

SHREYASH PATEL

Major Professor:	Adrienne Hoarfrost
Committee:	Frederick Maier
	Khaled Rasheed

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
August 2025

DEDICATION

This thesis is dedicated to my family, whose unconditional love and support made this possible. I especially thank my dad for always being there for me, even in the worst of times, and my mom, who gave up so much for me. I also appreciate my brother, who has stood by me in pursuing my dreams and nurturing my passion for science. I am deeply grateful to my friends and best friend, whose presence and support have been a true blessing. My sincere appreciation is also extended to my major professor and committee members for their invaluable guidance and mentorship throughout my academic journey. Lastly, I would like to thank lord Shanidev for his blessings.

ACKNOWLEDGEMENTS

I want to thank my major professor, Dr. Adrienne Hoarfrost, for her valuable guidance and help throughout my research. I would also like to thank my committee members, Dr. Khaled Rasheed and Dr. Frederick Maier, for their feedback and support. Special gratitude to my family—my father, Dinesh Patel, for his constant encouragement; my mother, Pravina Patel, for her sacrifices; and my brother, Shirish Patel, for encouraging me to follow my interest in science. I thank my best friend, Shardul Kulkarni, for his friendship and encouragement, and I would also like to mention my special friend Reena Bardeskar, for her guidance and support, and all my friends and extended family for their support. Finally, I would like to thank the University of Georgia and the Institute for Artificial Intelligence for providing a collaborative environment where my work could develop and Lord Shani Dev for his eternal blessings during this project.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

Introduction

1.1 The Challenge of Microbial Functional Annotation

Microbial life represents most of the biological diversity on Earth, spanning every environment from deep oceans to hydrothermal vents. This diversity is reflected not only in taxonomic variation but also in an immense range of functional capabilities encoded in microbial genomes. Yet most microbial organisms remain uncultured, uncharacterized, and functionally unannotated – a phenomenon often referred to as "microbial dark matter" (Lloyd et al., 2018). This hidden functional potential represents a massive untapped resource for biotechnology, medicine, and ecosystem science (Jiao et al., 2021; Cavicchioli et al., 2019).

The rise of untargeted metagenomic sequencing has accelerated the discovery of this hidden diversity. Large-scale global surveys, such as Tara Oceans (Sunagawa et al., 2015), Bio-GO-SHIP (Garcia et al., 2018), and BioGEOTRACES (Biller et al., 2018), have revealed millions of previously uncharacterized genes from environmental DNA. These efforts have expanded the known functional landscape of environmental microbiomes and led to the identification of sequences with biotechnologically relevant properties—such as plastic-degrading enzymes, novel antimicrobial peptides, and marine-encoded variants of already known systems like CRISPR-Cas (Chen et al., 2024). While CRISPR-Cas systems were originally discovered in cultured organisms, these environmental studies extended their presence to marine microbial communities, underlining the depth of uncharted functional diversity in the ocean microbiome.

Despite these discoveries, a critical bottleneck remains in functional annotation. Environmental sequencing has vastly outpaced our ability to assign biological meaning to genetic data. By the mid-2020s, protein sequence databases exceeded 300 million entries, yet fewer than 0.2% of these sequences had been experimentally validated (Bateman et al., 2023).

1.2 Enzyme Class Imbalance and Functional Gaps

Exacerbating this annotation gap is a severe class imbalance among categories of enzyme activities; common functional classes are heavily overrepresented in protein reference databases, whereas many enzyme classes have only a few representatives or exist as orphan classes with only one protein representing that class. The top 20% of Enzyme Commission (EC) numbers cover 90% of enzyme annotations in UniProt, while the remaining 80% represent just 10%, and approximately half of the proteins lack any EC annotation (Silveira et al. 2014; De Ferrari et al. 2012; Figure 1).

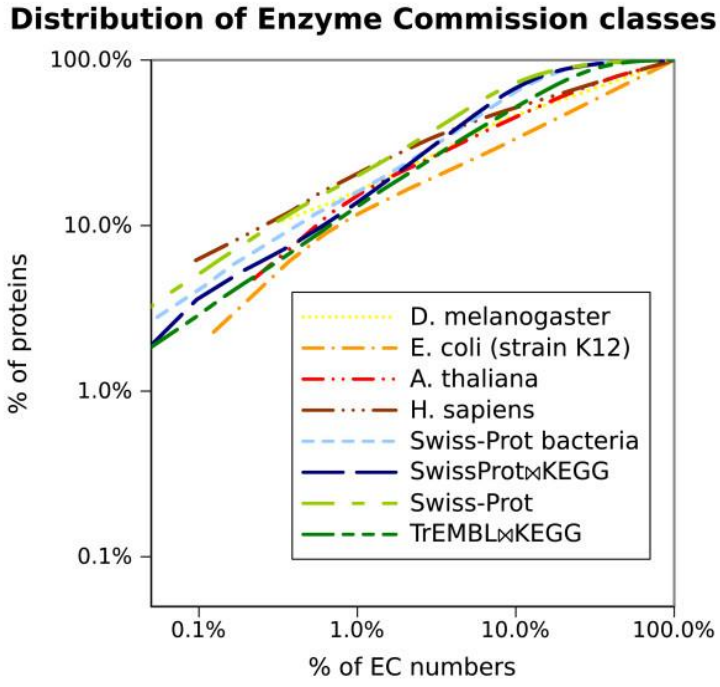

Figure 1: Proteins' Enzyme Commission (EC) numbers are distributed logarithmically in reference databases of proteins and common model organisms, with a few EC classes accounting for the majority of representative proteins. Starting with the most frequent EC number, the distribution is shown as a

cumulative percentage. Note the logarithmic scale on both axes. (Reproduced from De Ferrari et al., 2012.)

Enzyme Commission (EC) numbers categorize enzymes by the chemical transformations they catalyze, where each enzyme receives a four-digit number representing a hierarchical numerical classification, with each digit describing an increasingly specific molecular function (Han et al. 2023) (Appendix A, Figure 8). UniProt, the Universal Protein Resource, is a comprehensive database providing protein sequences, many of which with EC number functional annotations. Its manually curated section, Swiss-Prot, contains sequences with high-quality functional annotations and additional information such as enzyme activity, domains, and catalytic residues (Bateman et al. 2023). The prediction of enzyme function is made difficult by a severe class imbalance in the data, where some EC classes consist of hundreds or thousands of sequences with good representation, and many other EC classes consist of only one or a few gene sequences. An AI classification model trained on this class imbalance will likely perform better on highly represented classes, but worse on underrepresented classes with very few examples.

Despite advances in sequencing and annotation tools, several critical gaps persist in the field of enzyme function prediction. Homology-based annotation methods cannot capture the full diversity of microbial genes, especially those from uncultured organisms or poorly studied environments; portions of this "microbial dark matter" remain uncharacterized due to incomplete genome annotation across large sections of the bacterial tree of life, and low sequence similarity of genes within the same functional (EC) class (Vanni et al., 2022; Hoarfrost et al., 2022; Price et al., 2018).

AI-driven functional classification also faces several obstacles to adequate performance, particularly for rare EC classes. Most enzyme functional classification datasets exhibit a long-tail distribution, where a small number of EC classes dominate the training set, while many rare classes contain only a few examples. Standard classifiers often struggle to learn from these

underrepresented classes, resulting in biased predictions that favor frequent EC numbers (Dalkiran et al., 2018). Additionally, many existing AI-driven classifiers were trained on datasets that split data into training, validation, and test sets randomly without regard to sequence similarity of sequences across sets, causing homologous sequences with near-identical sequences to appear in both training and test sets. This artificially inflates performance by favoring memorization of the training set over generalization to true functional features underlying DNA sequences. Newer dataset such as BioTalk (Zhang et al., 2024) attempt to address this with sequence similarity-aware data splitting (e.g., using UniRef50 clusters such that proteins across sets do not have more than 50% amino acid similarity), offering a more realistic benchmark for functional prediction (Hou et al., 2023).

Moreover, few-shot learning remains underexplored. While recent models such as CLEAN (Yu et al., 2023) and HDMLF (Zhenkun et al., 2023) attempt to address this via contrastive learning and hierarchical structure, systematic frameworks for few- or one-shot function prediction remain limited.

1.3 Addressing Class Imbalance for Deep Learning-Driven Function Prediction with Data Augmentation

Class imbalance in functional annotation presents significant challenges. The dominance of a few EC classes biases computational models and annotation pipelines toward well-represented functions, diminishing prediction accuracy and functional inference for rare or orphan enzyme classes (Yang et al., 2024). Machine learning and homology-based approaches particularly struggle with underrepresented classes, exacerbating the functional annotation gap (Radivojac et al. 2013).

Sequence similarity does not necessarily imply functional equivalence between proteins (Pearson et al. 2013). Proteins with high overall sequence similarity can exhibit distinct biochemical functions due to variations in active site residues, and many proteins within the same EC number exhibit low sequence similarity (Pearson et al. 2013).

Homology-based methods fail to assign functions to proteins that have low sequence similarity to existing representatives in reference databases, and frequently misannotated functions of proteins based on high sequence similarity of functionally distinct enzymes (Schnoes et al. 2009). The assignment of functional labels to proteins thus proves challenging in two cases: proteins without closely homologous relatives, and functional classes with few known instances (Radivojac et al. 2013). Machine learning-based functional prediction methods use computational algorithms to predict protein functions from sequence data. However, class imbalance significantly impacts these methods by skewing training datasets toward dominant functional classes, reducing accuracy and reliability for predicting rare or novel enzyme classes (Yang et al. 2024).

In response to these challenges, deep learning approaches have shown strong promise for functional prediction. Functional annotation refers to the process of predicting the biological role, activity, or localization of nucleotide or protein sequences. Recent advances in biological foundation models such as ESM (Rives et al. 2021), ProtTrans (Elnaggar et al. 2022), and ProteinBERT (Brandes et al. 2022) demonstrate that transformer-based architectures can learn complex biological patterns from hundreds of millions of sequences. These pre-trained models provide powerful sequence embeddings for downstream tasks, including enzyme classification, subcellular localization, and secondary structure prediction. However, these embeddings still

require task-specific classifiers to extract actionable predictions, particularly under class imbalance (Yang et al. 2024; Zou et al. 2019).

To improve model performance in low-data regimes and mitigate class imbalance, data augmentation has emerged as a promising strategy in biological sequence modeling (Wen et al. 2020). In this work, we explore three key augmentation strategies:

First, reverse complement augmentation (Appendix A, Figure 9) leverages the strand symmetry of DNA. Since sequencing technologies can randomly capture either strand, generating reverse complements effectively doubles the dataset size while preserving biological validity and has been previously used for data augmentation in biological deep learning contexts (Cao and Zhang 2019, Hoarfrost et al. 2022). This promotes strand-invariant learning and is particularly useful in metagenomic contexts; however, the complementary strand of a gene coding sequence does not typically code for the same gene, and the utility of reverse complementation for data augmentation in a functional prediction context is unknown.

Second, synonymous codon substitution takes advantage of codon redundancy in the genetic code. By substituting codons that encode the same amino acid, new gene sequences can be generated without changing the encoded protein, thereby enhancing nucleotide diversity while preserving biological function (Rodriguez et al., 2024) (Appendix A, Figure 10,11). This method mimics natural genetic variation and may strengthen model robustness to unseen codon usage patterns.

Third, we employ conditional Generative Adversarial Networks (cGANs) (Appendix A, Figure 14) to synthesize entirely new sequences conditioned on biological features such as EC class labels, GC content, and codon frequency. While GANs (Appendix A, Figure 13) have been successfully used to generate novel enzymes (Repecka et al. 2021) and functionally constrained

protein sequences (Kucera et al. 2022), their use for functional DNA augmentation—particularly conditioned on EC labels remains underexplored. The application of generative models for biological classification is still poorly understood, particularly in low shot learning contexts. Conditional generation enables the enrichment of low-frequency classes with realistic sequences that align with biological constraints (Marouf et al. 2020). Our study applies cGANs to address this gap.

These augmentation methods are evaluated both individually and in hybrid combinations to assess their impact on enzyme classification, with an emphasis on rare EC classes. Our research investigates how training classifiers with synthetically augmented data affects their ability to recognize underrepresented enzyme functions. This study represents a new direction, drawing from protein design concepts and generative modeling to enhance functional prediction in the context of metagenomic enzyme discovery (Hawkins-Hooker et al. 2021; Marouf et al. 2020).

This work is driven by two research questions:

1. Can data augmentation improve the classification of rare enzyme classes from gene sequence?
2. Which augmentation strategies offer the greatest improvements?

To address these questions, we define the following objectives:

1. Develop and evaluate a functional classification pipeline using embeddings derived from a deep learning biological foundation model connected to an enzyme function prediction classifier.

2. Implement multiple augmentation techniques individually and in hybrid combinations to create benchmark augmented datasets that alleviate class imbalance of EC functional classes.
3. Assess the impact of these data augmentation strategies on functional prediction, with an emphasis on rare-class enzymes using performance metrics including macro recall and F1-score.

CHAPTER 2

Methodology

2.1 Functional Prediction Dataset Curation

A dataset of gene sequences with known functional annotations was curated from the SwissProt database (Bairoch et al., 2000), with each gene DNA sequence associated with EC number annotations. A previously described BioTalk dataset (Zhang et al., 2024) associating gene coding DNA sequences with functional annotations (EC numbers) was used for training and evaluation. This dataset is divided into training, validation, and test sets in a stratified manner as described in (Zhang et al., 2024), preserving an approximate 80/10/10 proportion of each enzyme class across splits, and maximizing sequence dissimilarity across training, validation, and test sets. In brief, training, validation, and test splits use UniRef cluster assignments (Suzek et al., 2015) to ensure minimal sequence similarity between partitions. This allows for a more realistic evaluation of generalization performance, particularly on novel or rare enzyme classes. This dataset consists of 151,314 training examples, 19,296 validation examples, and 19,930 test examples spread across 2,228 EC number categories, with a median count of 4 genes per EC class, a 90^{th} percentile of 248, and maximum of 2,288.

This curated dataset addresses several long-standing challenges in biological functional prediction: Realistic class imbalance, reflecting the rarity of many enzyme functions in real-world DNA

sequencing datasets; use of high-quality, manually curated SwissProt entries, ensuring accurate functional prediction labels; and cluster-aware data splitting using UniRef50–100, ensuring that train, validation, and test splits have low sequence similarity, as is expected during deployment.

In this work, we use the 'SwissProt Unbalanced' set (Benchmark 3) from Zhang et al. 2024, and the corresponding Test Set I, which includes EC classes that are also present in the training set. This test set enables a meaningful evaluation of the model's baseline performance for known functional categories. In this study, data augmentation strategies were applied only to the training set, and the validation and test sets remain unaugmented.

Table 1: Summary statistics of enzyme classification dataset (Zhang et al., 2024), showing the number of unique EC classes, total samples, and class distribution metrics.

Total EC classes	**2,228**
Total training examples	151,314
Median genes per EC class	4
EC classes with one example	770
EC classes with more than 10 examples	267
Mean genes per EC class	68

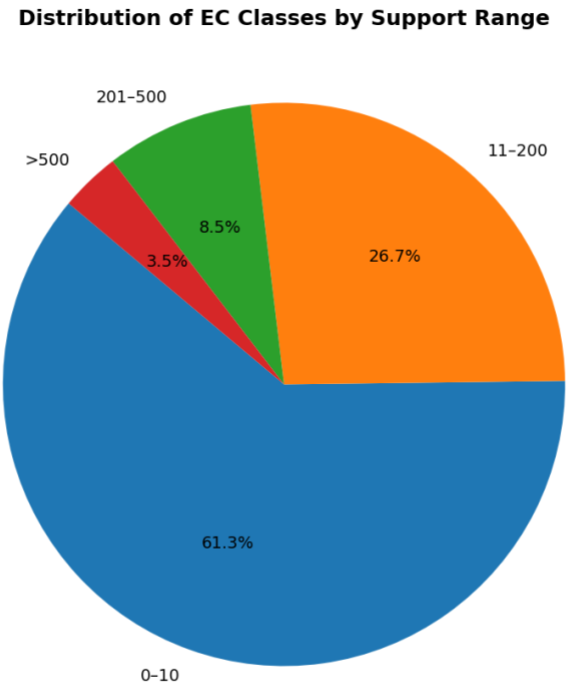

Figure 2: Distribution of EC classes by support range. Most EC classes (61.3%) fall in the lowest support range (0–10 sequences per class), illustrating a highly skewed long-tail distribution. Only 3.5% of EC classes have more than 500 examples, highlighting the underrepresentation of most enzyme functions in the dataset.

2.2 Embeddings and Classifier Architecture for Enzyme Function Prediction

Embedding Generation from LookingGlass

We leveraged LookingGlass 2.0 (Jha and Hoarfrost 2025), a biological foundation model hosted on Hugging Face, to encode baseline and augmented gene sequences as context-rich embedding vectors. LookingGlass captures functional features of DNA sequences by transforming them into dense embeddings that reflect evolutionary and biochemical relationships—information not captured by one-hot encodings or k-mer counts (Hoarfrost et al., 2022). To prevent overfitting, model weights were frozen during training, and the fixed-length embeddings were passed directly to the CNN classifier.

Choice of Classifier: 1D Convolutional Neural Networks (CNNs)

We designed a 1D Convolutional Neural Network (CNN) as the classification head on top of LookingGlass embeddings. CNNs are particularly effective at detecting localized motifs and conserved patterns critical for enzyme function, offering strong inductive biases in biological sequence analysis. Compared to RNNs and Transformer-based decoders, CNNs are computationally efficient, less prone to overfitting, and well-suited for small-to-medium datasets (Zeng et al., 2016; Almagro Armenteros et al., 2019;). Recent work demonstrated CNNs' state-of-the-art performance in the Random Promoter DREAM Challenge (Rafi et al., 2024).

2.3 Data Augmentation Methods

Reverse Complement Augmenation

We tested using reverse complementation for a varying number of randomly selected underrepresented EC classes. We selected N number of classes [10, 25, 50, 70] such that for N percent of underrepresented classes, the reverse complement of each gene within that class was added to the augmented dataset, doubling the number of training examples of genes in the selected classes (Table 2, Figure 3). This resulted in a training set size ranging from 152,739 to 302,628 training examples for RC10-RC100 (corresponding to 10-100% of training examples augmented), relative to a baseline training set size of 151,314. Here RC (100) is Reverse complement of 100% of the original dataset sequences.

Table 2: Reverse Complement Augmentation Impact on Training Size. Summary of total and median training samples after applying varying levels of reverse complement augmentation.

Augmenation Variation	Total training sample	Median
RC 10	152,739	5.0
RC 25	152,996	5.0
RC 50	153,694	6.0
RC 70	154,234	7.0
RC 100	302,628	8.0

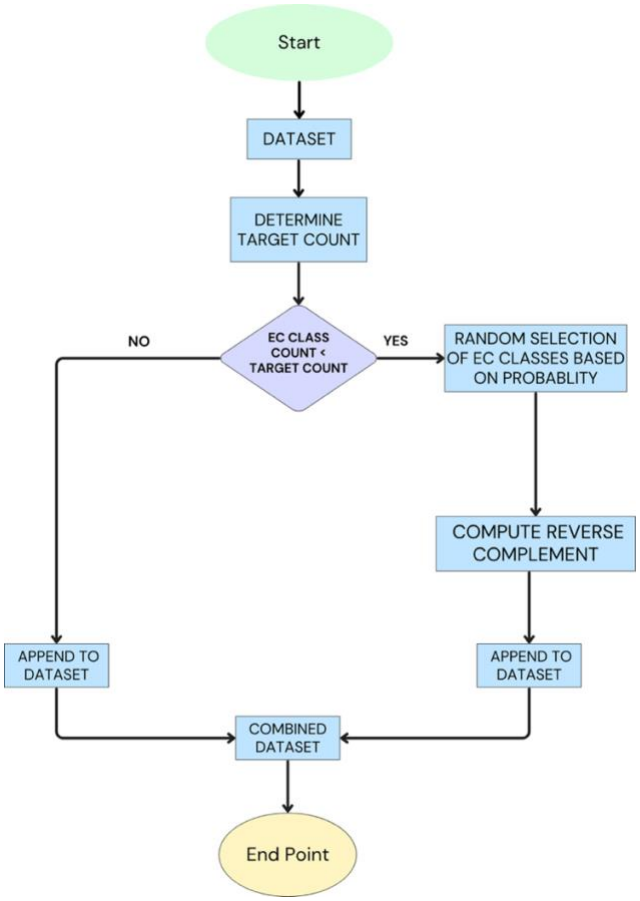

Figure 3: Reverse Complement Probability Based Substituion Flowchart. Workflow for reverse complement-based augmentation. Sequences from underrepresented EC classes are probabilistically selected and augmented using their reverse complements to expand the training dataset.

Synonymous Codon Substitution Augmentation

Codon substitution augmentation was performed using a random selection algorithm for augmentation. For each sequence in the training set, synthetic sequence creation involved codon-by-codon examination, with p representing the probability of selecting a synonymous codon from the original sequence. The primary augmentation runs used p = [0.1, 0.25, 0.5, 0.7], resulting in approximately 10, 25, 50, or 70% of codons in a gene sequence being replaced by different randomly selected synonymous codon (Appendix A, Figure 15).

To balance the dataset, the augmentation method was applied iteratively to EC classes with fewer than 248 training examples, a threshold selected as it represents the 90th percentile of gene sequence counts across all EC classes. This process continued until each EC class contained 248 gene sequences, resulting in a training set of 615,608 total sequences with a median of 248 genes per EC class (Table 3).

Table 3. Codon Substitution Augmentation Impact on Training Size. Summary of total and median training samples after applying varying levels of Codon Substitution augmentation.

Augmenation Variation	Total training sample	Median support
CS 10	615,608	248
CS 25	615,608	248
CS 50	615,608	248
CS 70	615,608	248

Conditional GAN-Based Sequence Augmentation

We trained a conditional Generative Adversarial Network (cGAN) (Appendix A, Figure 14) to generate novel DNA sequences for underrepresented enzyme function classes. To tailor sequence generation toward specific enzyme classes, the cGAN was conditioned on three biological attributes: (i) the EC class label, (ii) codon frequency (Appendix A) vectors, and (iii) GC content (Appendix A, Figure 12). These conditioning components were provided to both the generator (G) and discriminator (D) to guide generation. GC content quantifies the proportion of guanine (G) and cytosine (C) bases characterizing an EC class, which impacts gene stability and expression, while codon frequency captures how synonymous codons are preferentially used within enzyme

classes. Conditioning G on these features ensures that it produces sequences statistically aligned with the characteristics of the original target class.

The generator is a fully connected feedforward network that accepts a 100-dimensional noise vector along with three conditioning vectors: an EC class embedding (projected to 100 dimensions via an embedding layer), a GC content scalar (transformed using a single-layer linear network to a 100-dimensional vector), and a codon frequency vector (64-dimensional, also transformed using a single-layer linear network to a 100-dimensional vector). Each of these three conditioning inputs is projected separately into 100-dimensional vectors. To regulate their influence, we apply a conditioning strength parameter (tested between 0.1 and 1.0), which mixes the learned embedding with random noise to encourage robustness while preserving biological relevance. The noise vector and the three condition vectors are concatenated and passed through a single dense (fully connected) layer activated by a ReLU function. This layer transforms the input into a sequence-length × vocabulary-size matrix, which is reshaped and normalized via a softmax function to approximate one-hot encoded DNA sequences.

The discriminator is a convolutional neural network (CNN) that receives either a real or generated DNA sequence (one-hot encoded), along with the same conditioning inputs: EC label embedding, GC content vector, and codon usage profile. The input sequence is passed through three stacked 1D convolutional layers with kernel size 4 and stride 2, using increasing filter sizes of 64, 128, and 256 (as specified by HIDDEN_DIM), each followed by a LeakyReLU activation function. The output of the final convolutional layer is flattened and concatenated with the conditioning vectors before passing through two fully connected layers: a hidden layer (with LeakyReLU) and a final sigmoid output layer that classifies the input as real or synthetic.

Training was conducted for up to 200 epochs using the Adam optimizer, with learning rates of 0.0005 for G and 0.0002 for D ($\beta_1 = 0.5$, $\beta_2 = 0.999$), on a Google Colab Pro instance equipped with a NVIDIA A100 GPU (40 GB memory). Mini batches (batch size = 64) were stratified by EC class to maintain class distribution during training. We applied label smoothing, assigning real labels a value of 0.9 to stabilize discriminator training. While individual generator and discriminator losses fluctuated during training, as is typical in adversarial training, progress was assessed through the biological plausibility of generated sequences (based on GC content and codon usage profiles), training dynamics, and improvements in downstream classifier performance, especially for rare EC classes.

Effective generator performance was typically observed between 6 to 18 epochs, with earlier epochs showing consistent alignment to target class properties and later stages offering diminishing returns. These observations guided our final training configuration, with 6 to 18 epochs producing high-quality synthetic sequences suitable for augmentation.

This cGAN framework offers a biologically informed strategy for sequence-level data augmentation, particularly for long-tail rare functional classes. By embedding enzyme-specific genomic features into the generative process, the model creates class-consistent synthetic sequences that help balance training datasets and enhance the functional annotation of rare enzymes (Appendix A, Figure 17). This resulted in augmented training sets with rare EC classes boosted to a target threshold of 200 sequences per class (Table 4).

Table 4: cGAN Augmentation Impact on Training Size. Summary of total and median training samples after applying varying levels of cGAN augmentation.

Augmenation Variation	Total training sample	Median support
GAN 30	600,715	248
GAN 50	615,608	248
GAN 70	625,678	249

Hybrid augmentation Techniques:

We also tested additional combinations of augmentation techniques in two datasets:

Codon500: In the original codon frequency augmented set, we used dynamic targeting for generating new sequences up to a the 90th percentile threshold (248 genes per EC class). We additionally create a codon frequency augmented set where the target threshold of number of genes per class is 521 (97th percentile), with substitution probability of 25% (Table 5).

GAN70 +Codon25: We first used the cGAN (with 0.7 conditioning strength) to generate synthetic sequences for rare enzyme classes, up to the 90th percentile class size threshold. To introduce additional variation, we subsequently applied synonymous codon substitution to the synthesized sequences 1x with a 25% probability directly on these GAN-generated sequences. As a result, each rare-class sequence from the GAN had a codon-variant counterpart, effectively doubling the diversity while preserving functional constraints. This resulted in a training set of 1,325,752 sequences with a median of 574 sequences per EC class (Table 5).

Table 5: Hybrid Augmentation Impact on Training Size. Summary of total and median training samples after applying varying types of Hybrid augmentation.

Augmenation Variation	Total training sample	Median support
GAN70 +CODON25	1,325,752	574
CODON 500	1,208,201	521

2.4 Classification Model Architecture and Training

We developed a sequence classification pipeline to predict enzyme functions (EC numbers) trained on our baseline benchmark dataset (see above) and its augmented variants. Each gene sequence was first embedded using LookingGlass 2.0 (Jha and Hoarfrost 2025), a pretrained DNA language biological foundation model, and the resulting fixed-length vectors were used as input to a custom 1D Convolutional Neural Network (CNN) trained for multiclass classification across all 2,228 EC classes in the training set. The embedding dimension was fixed at 512 across all experiments to ensure a uniform input shape for the CNN decoder.

The CNN architecture (Appendix A, Figure 18) consisted of two sequential 1D convolutional layers with 128-dimension filters and a kernel size of 3, using ReLU activation and stride-based down sampling in place of traditional pooling layers. Dropout layers with a rate of 0.5 followed each convolutional block to reduce overfitting. A global average pooling operation was applied to maintain input length invariance and extract compact feature representations, followed by a 256-unit dense layer and a softmax output layer aligned to the EC label space. This design enabled efficient extraction of sequence-level motifs and functional signals from the embedded input.

The model was trained using the Adam optimizer with categorical cross-entropy loss. Training ran for up to 200 epochs, and early stopping was applied if validation loss failed to improve for 25

consecutive epochs. A batch size of 64 was used throughout, and training was conducted on a single NVIDIA A100 GPU (80 GB RAM).

To ensure reproducibility, a fixed random seed (42) was applied across all components, including dataset shuffling and model initialization. Training and validation loss and accuracy were recorded at each epoch during training. For the final evaluation of the test set, we selected the best-performing model checkpoint based on validation loss using the early stopping criteria described above.

2.5 Implementation and Code Availability

The CNN classifier and conditional GAN (Paszke et al., 2019) were trained using PyTorch (with CUDA support), Biopython was used for sequence manipulation including reverse complementation and codon translation (Cock et al., 2009), and fastBio's API was used to incorporate pre-trained LookingGlass 2.0 embeddings (Hoarfrost et al., 2022, ref LGv2). Other tools included scikit-learn for performance analysis, NumPy for vector projections, and SciPy for statistical comparisons and GC content calculation. Large-scale GAN and CNN training was made possible by the use of an NVIDIA A100 GPU (40 GB) for both training and experimentation on Google Colab Pro. This hardware and software infrastructure enabled our classification pipeline and evaluated proposed augmentation methods.

All code developed and used in this study, including the preprocessing scripts, synonymous codon substitution pipeline, conditional Generative Adversarial Network (cGAN) model, and the enzyme classification framework, is publicly available on GitHub.(https://github.com/Hoarfrost-Lab/DataAugmentation).

2.6 Data Availability

The datasets used for training and evaluating the models, including the original, augmented, and benchmark-ready versions, are hosted on Hugging Face and freely available for academic and research use. The dataset repository includes metadata, class distributions, and augmentation settings to enable full reproducibility of results presented in this work (https://huggingface.co/datasets/HoarfrostLab/Augmented_Dataset_for_EC_Class_Prediction).

2.7 Model Evaluation and Logging

To evaluate the performance of our classification pipeline, particularly under conditions of severe class imbalance, we employed several standard classification metrics. Each metric provides insight into a specific aspect of model behavior, especially in distinguishing well-represented EC classes from rare ones (defined here as EC classes with fewer than five training examples). These metrics were computed over the held-out test set to assess generalization beyond the training distribution. These metrics were computed over the held-out test set to assess generalization beyond the training distribution.

Test accuracy is the most basic metric. It reflects how many predictions the model got right overall, regardless of class. Mathematically, it is defined as:

TP (True Positives) are cases where the model correctly predicted the right enzyme class.

TN (True Negatives) refer to all the non-target class instances correctly identified as not belonging to the predicted class.

FP (False Positives) occur when the model incorrectly predicts a sequence to belong to a certain EC class when it does not.

FN (False Negatives) happen when the model fails to detect a class that is present.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision quantifies the reliability of a positive prediction. In other words, when the model assigns a sequence to a specific class, precision measures how often that assignment is correct. Mathematically defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall: asks how well the model captures all relevant instances of a class. It answers: "Of all the true examples belonging to a class, how many did the model successfully retrieve?

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 score combines both precision and recall into a single number by taking their harmonic mean. It balances the trade-off between being accurate and being complete.

$$\text{F1 SCORE} = 2 \times \frac{\text{Precision} + \text{Recall}}{\text{Precision} \times \text{Recall}}$$

This metric gives us a robust sense of per-class performance, especially when both false positives and false negatives are of concern.

To evaluate performance across the entire EC classification space, we report two types of averaged F1 scores:

Macro F1: calculates the F1 score separately for each EC class and then averages them without weighting. Every class, whether frequent or rare, is treated equally.

$$\text{Macro F1} = \frac{1}{C}\sum_{i=1}^{C} F1_i$$

Where: C is the total number of EC classes. $F1_i$ is the F1 score computed for class i.

Weighted F1: also averages F1 scores across classes but assigns higher weight to frequent classes based on the number of samples per class.

$$\text{Weighted F1} = \sum_{i=1}^{C} \frac{n_i}{N}.F1_i$$

Where n_i is the number of samples in class i, N is the total number of samples across all classes, and $F1_i$ is again the F1 score for class i.

To further characterize the impact of augmentation strategies on rare enzyme functions, we introduce additional evaluation metrics:

Predicted Classes (PC) refers to the number of unique EC classes the model predicted at least once in the test set.

Rare Predicted Classes (RPC) is the subset of PC that includes only those rare EC classes (support < 5) for which the model made at least one correct prediction.

Rare Class Coverage (%) quantifies how much of the rare class space was captured:

$$\text{Rare Coverage} = \frac{\text{Rare Class Predicted}}{\text{Total Number of Rare Class}} \times 100$$

This metric directly assesses whether our augmentation strategies help recover functional diversity from the long tail of the EC class distribution.

By combining these metrics, we can comprehensively evaluate the effectiveness of the model, not only in predicting dominant classes, but more importantly, in uncovering rare and novel enzyme functions a key focus of this work.

CHAPTER 3

Results

This chapter presents a comparative evaluation of different data augmentation strategies for enzyme commission (EC) classification, relative to baseline performance on a model trained on the training dataset with no augmentation. Our evaluations were conducted with a test set comprising 19,929 DNA sequences having 2,228 EC classes, out of which 1,638 classes are classified as rare (defined as having five or fewer examples in the test set without augmentation).

3.1 Baseline Performance

The baseline classifier was trained on the unaugmented dataset, which showed 24.74% test accuracy, with a macro-averaged F1 score of 0.15 and a weighted F1 score of 0.25. The model predicted 740 different EC classes in the test set out of a total of 2,228 enzyme classes in the dataset. Among these, 545 belonged to the rare class subset, showing a rare class coverage of 33.27 percent. This indicates that, the classifier is biased toward dominant classes. For instance, EC 2.7.7.7 (DNA polymerase, with 513 genes representing that category) had an F1 score of 0.92, whereas EC 4.1.1.101 (malolactic enzyme, with 3 genes representing that category) had an F1 score of 0.00 (undetected). This disparity underscores the severe class imbalance present in known functional annotation data.

3.2 Synonymous Codon Substitution (CS) Augmentation

Augmentations produced by synonymous codon substitution at 10–70% substitution rates of the training data showed significant improvement in model performance at all substitution rates,

effectively oversampling underrepresented classes without introducing noise (Table 6, Figure 4). The highest test accuracy performance was observed at a 25% substitution rate, although performance was comparable at 10, 25, and 50% substitution rates.

Table 6: Codon Substitution Augmentation Performance Metrics. Comparison of models trained with varying levels of codon substitution probability, evaluating accuracy, F1 scores, predicted class counts, and rare class coverage.

Augmentation	Test Acc	Macro-F1	Weighted F1	PC	RPC	Rare Coverage
Baseline	24.75%	0.15	0.25	740	545	33.27%
CS10	43.0%	0.23	0.41	852	629	38.4%
CS25	**44.0%**	**0.23**	**0.41**	**856**	**628**	**38.34%**
CS50	43.86%	0.23	0.41	851	625	38.15%
CS70	31.0%	0.08	0.28	788	580	35.4%

Notably, rare classes such as EC 4.1.1.101 improved from F1 = 0.00 in the baseline to F1 = 0.55 (CS50), while EC 4.1.1.103 improved from F1 = 0.00 to F1 = 0.40 (CS25).

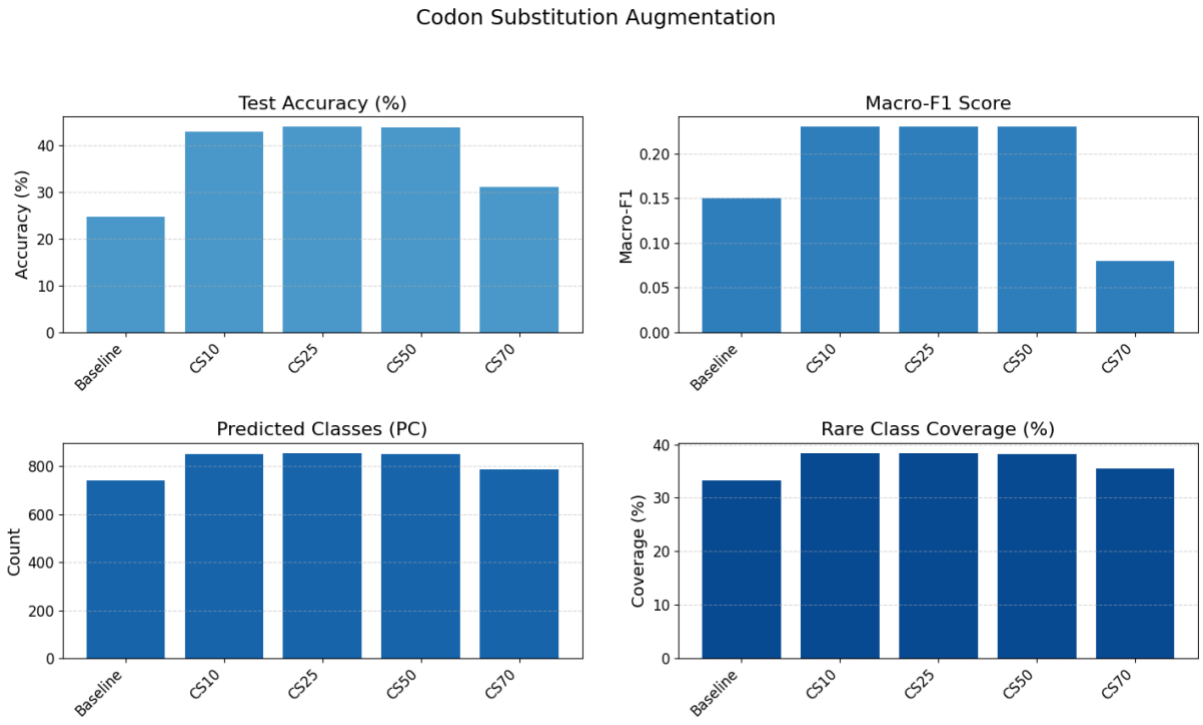

Figure 4: Performance Metrics for Codon Substitution Augmentation. Charts display changes in test accuracy, macro-F1 score, predicted class count (PC), and rare class coverage (%) compared to the baseline.

3.3 Reverse Complement (RC) Augmentation

Reverse Complement augmentation at varying levels from 10% to 100% of the training data. The performance declined sharply with higher augmentation rates (Table 7, Figure 5). Since reverse complement sequences typically do not encode the same protein, and may not encode any protein at all, this strategy introduced label noise, confusing the model and reducing both rare and common class accuracy.

Table 7: Reverse Complement Augmentation Performance Metrics. Comparison of models trained with varying levels of reverse complement probability, evaluating accuracy, F1 scores, predicted class counts, and rare class coverage.

Augmentation	Test Acc	Macro-F1	Weighted F1	PC	RPC	Rare Coverage
Baseline	24.75%	0.15	0.25	740	545	33.27%
RC10	25.01%	0.15	0.25	737	540	32.96%
RC25	23.23%	0.14	0.23	716	519	31.67%
RC50	24.6%	0.14	0.25	724	529	32.96%
RC70	23.77%	0.14	0.24	714	521	31.80%
RC100	13.06%	0.08	0.13	505	375	22.89%

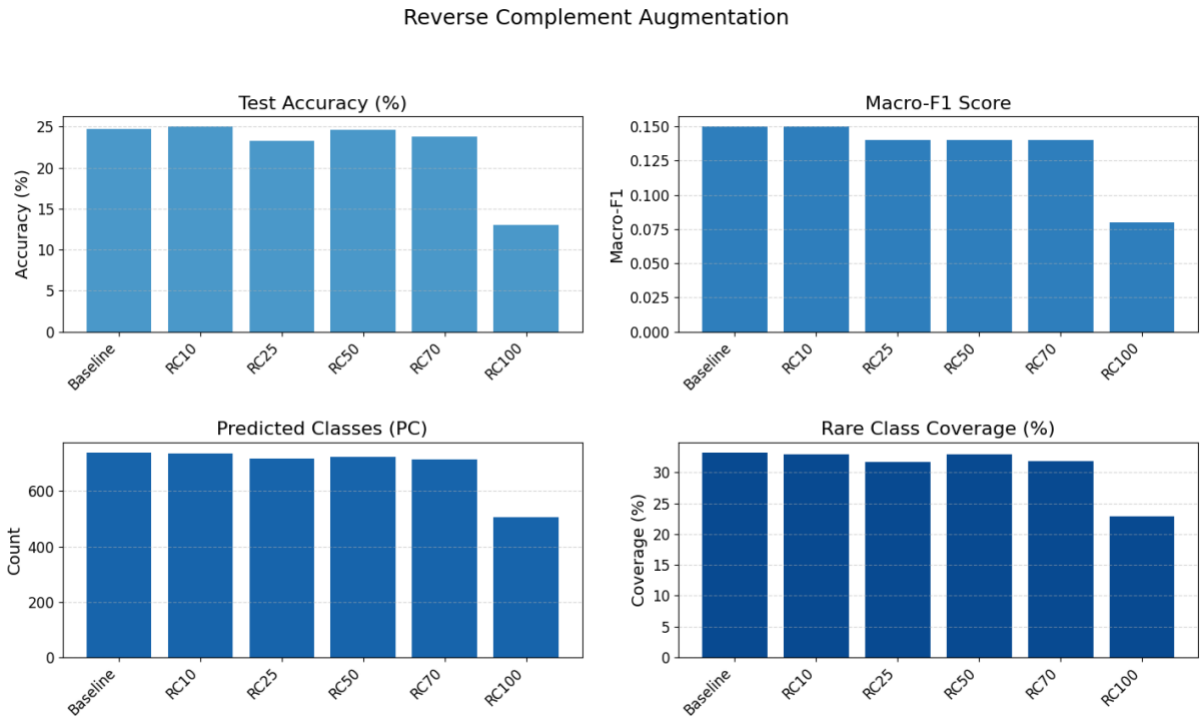

Figure 5: Performance Metrics for Reverse Complement Augmentation. Charts display changes in test accuracy, macro-F1 score, predicted class count (PC), and rare class coverage (%) compared to the baseline.

3.4 cGAN-Based Augmentation

We integrated class-specific synthetic sequences at 30%, 50%, and 70% of the training size using a conditional GAN. Although the overall test accuracy was marginally improved, GAN-based augmentation decreased macro-F1 and rare-class recovery (Table 8, Figure 6). This suggests that random noise may have been introduced by synthetic sequences, particularly for rare classes, resulting in a decline in data quality.

Table 8: cGAN Augmentation Performance Metrics. Comparison of models trained with varying levels of cGAN conditioning factor, evaluating accuracy, F1 scores, predicted class counts, and rare class coverage.

Augmentation	Test Acc	Macro-F1	Weighted F1	PC	RPC	Rare Coverage
Baseline	24.75%	0.15	0.25	740	545	33.27%
GAN30	26.37%	0.05	0.22	359	268	16.37%
GAN50	30.34%	0.07	0.27	398	296	18.07%
GAN70	31.34%	0.08	0.28	432	325	19.84%

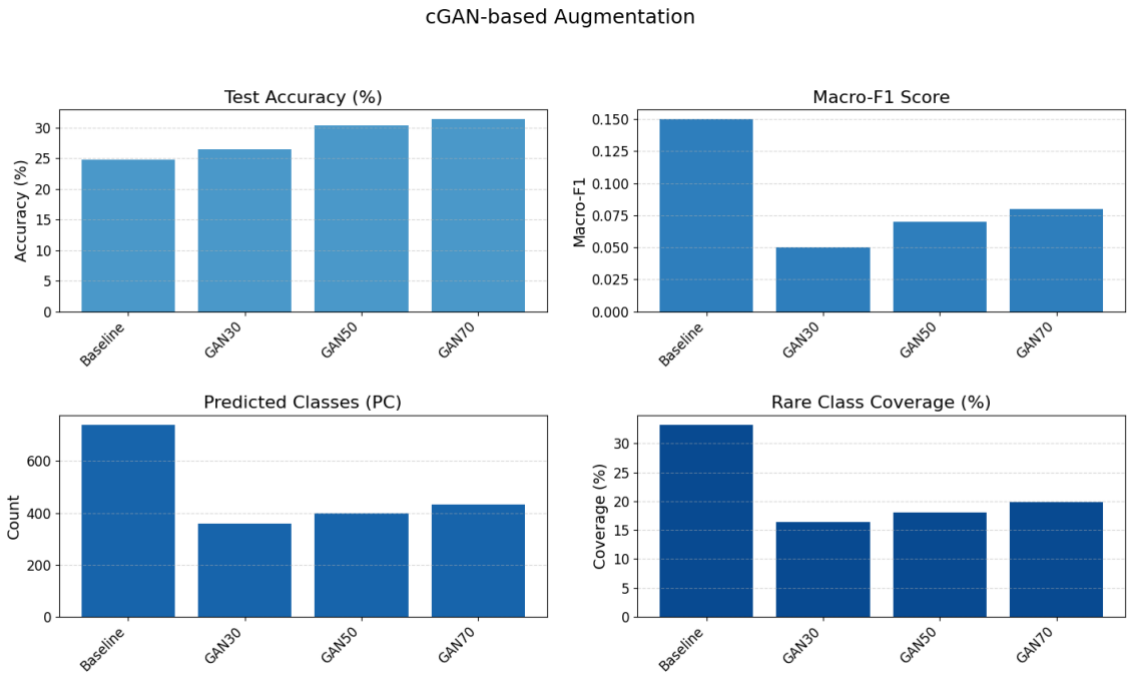

Figure 6: Performance Metrics for cGAN Augmentation. Charts display changes in test accuracy, macro-F1 score, predicted class count (PC), and rare class coverage (%) compared to the baseline.

3.5 Hybrid and Fixed-Size Augmentation

We also evaluated two composite approaches: a combination of GAN generation and codon substitution at a class target threshold of an increased 521 genes per class (Table 5). The hybrid GAN70+CS25 approach performs better than GAN30 in both test accuracy (29.24% vs. 26.37%) and rare class coverage (18.06% vs. 16.37%), indicating that mixing GAN and codon substitution may have the benefits but marginal. Codon25-500 matches Codon25 in accuracy (43.46% vs. 44.0%) but slightly improves weighted F1 (0.44 vs. 0.41) and maintains rare class coverage (38.34%), suggesting that combining GAN and codon substitution may have some advantages, albeit slight ones. Codon25-500 maintains rare class coverage (38.34%) and slightly improves weighted F1 (0.44 vs. 0.41) while matching Codon25 in accuracy (43.46% vs. 44.0%), suggesting more consistent gains are provided by fixed per-class targets. According to these findings, codon

substitution augmentation consistently improves model performance, and more research is necessary to determine the ideal target threshold of labels for each category.

Table 9: Hybrid Augmentation Performance Metrics. Comparison of models trained with varying levels of hybrid augmentation (along with baseline of individual techniques) evaluating accuracy, F1 scores, predicted class counts, and rare class coverage.

Augmentation	Test Acc	Macro-F1	Weighted F1	PC	RPC	Rare Coverage
Baseline	24.75%	0.15	0.25	740	545	33.27%
GAN70+CS25 (Hybrid)	29.24%	0.06	0.25	394	298	18.06
Codon25-500	43.46%	0.22	0.44	856	628	38.34%
GAN30	26.37%	0.05	0.22	359	268	16.37%
CS25	44.0%	0.23	0.41	856	628	38.34%

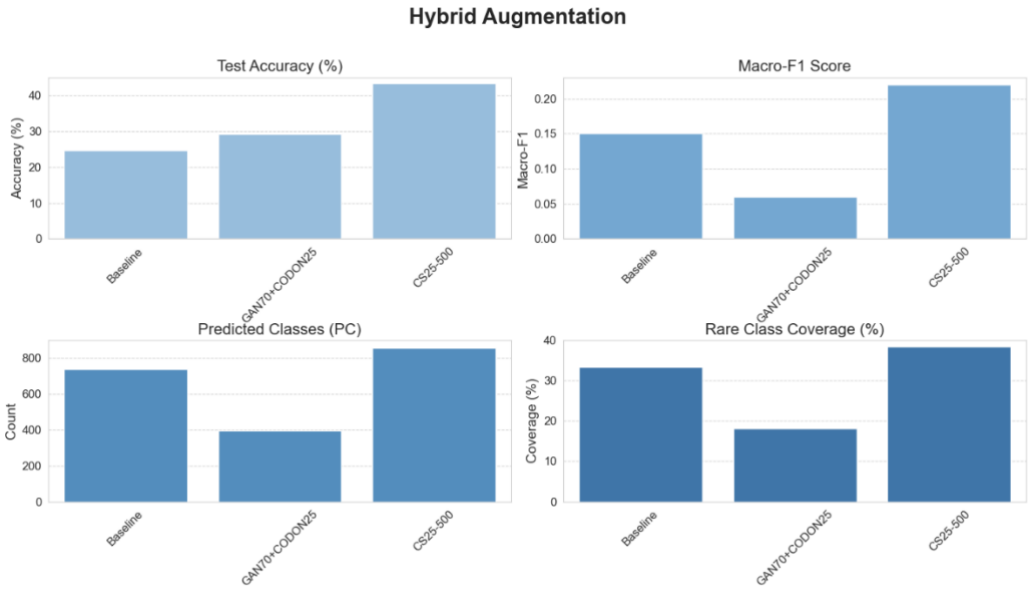

Figure 7: Performance Metrics for Hybrid Augmentation. Charts display changes in test accuracy, macro-F1 score, predicted class count (PC), and rare class coverage (%) compared to the baseline.

CHAPTER 4

Discussion and Conclusion

4.1 The Role of Biologically Informed Augmentation in Rare-Class Prediction

This study demonstrates the significant improvement in enzyme function prediction under severe class imbalance that can be achieved through biologically based augmentation techniques. The most successful approach among the methods assessed was synonymous codon substitution (CS). In contrast, reverse complement augmentation degraded performance in most cases, while GAN-based synthetic sequences marginally improved overall performance but resulted in especially poor performance in rare functional classes.

Codon substitution introduces variation at the nucleotide level while maintaining protein-level semantics by taking advantage of the genetic code's redundancy. Instead of learning sequence-level artifacts, this motivates the classifier to learn domain-level features. Codon substitution increased test accuracy by up to 1.8x (from 24.75% to 44%), macro F1 by 1.5x (from 0.15 to 0.23), and weighted F1 by 1.6x (from 0.25 to 0.41) across all experiments. Moreover, the number of rare class categories predicted at least once by the classifier increased from 545 to 629, and rare-class coverage improved by approximately five percentage points (33.27% to 38.4%). Even well-represented classes such as EC 2.7.7.7 (DNA polymerase) saw performance gains, with

F1 rising from 0.92 to 0.94—demonstrating the robustness of CS across both abundant and sparse classes (Table 5, Figure 3)

4.2 Pitfalls of Reverse Complement and Synthetic GAN-Based Augmentation

On the other hand, reverse complement (RC) augmentation continuously degraded model performance. At higher integration levels (e.g., RC100), accuracy dropped to 13.06%, macro-F1 dropped to 0.08, and rare-class coverage dropped to 22.89% (Table 6, Figure 4). This degradation result is in line with biological expectations, in particular the fact that reverse strands typically do not encode the same protein as the original strand, rendering the reverse-complemented sequences functionally invalid. The inclusion of these sequences introduced noise that hampered the learning process and label integrity. Reverse complementation can be a helpful augmentation strategy in self-supervised learning settings (Hoarfrost et al. 2022), where these sequences are biologically valid and labels are not present. This strategy, however, is inappropriate in a labeled classification context, where it is unlikely that the reverse complement strand will have the same label.

GAN-based augmentation presented a different set of challenges. Although conceptually attractive for balancing rare classes, cGAN-generated sequences struggled to maintain functional validity, particularly for rare classes. This is likely due to the severe data limitation in the rare EC classes, many of which have only single digit training examples, which was insufficient to provide enough context to produce viable synthetic sequences for those classes. While minor gains in test accuracy were observed at some GAN integration levels, particularly for more abundant EC classes, macro-F1 scores dropped to as low as 0.05, and rare-class coverage sank to around 16–20% (Table 7, Figure 5). These results stress the importance of sequence validation in generative

pipelines and the challenges of deploying raw synthetic data in biological tasks with severe class imbalance, particularly for rare classes with very low support.

4.3 Evaluating Combined and Fixed-Size Augmentation Strategies

The evaluation included both Hybrid augmentation strategies GAN70+CS25 and fixed-threshold codon substitution (Codon500). The hybrid GAN70+CS25 method failed to achieve better results than the individual codon substitution approach. The GAN70+CS25 model achieved 29.24% accuracy and 18.06% rare-class coverage, while Codon500 reached 43.46% accuracy and 38.34% rare-class coverage (Table 8, Figure 6). The results indicate future research should explore adaptive augmentation strategies that adjust their approach based on the level of class scarcity and augmentation response metrics.

4.4 Implications for Enzyme Annotation in Novel Microbial Contexts

The enhanced detection of rare classes has substantial effects on future applications. The majority of functions remain poorly documented in reference databases, yet the environment contains numerous sequences with rare or new functions which scientists have yet to discover (Griesemer et al., 2021; Steen et al., 2023). A 5% improvement in rare-class coverage would allow researchers to detect thousands of new enzyme candidates from metagenomic datasets that contain tens of millions of unknown genes (Sunagawa et al. 2015). These sequences may encode novel biocatalysts, ecological biomarkers, or metabolic intermediates with important ecological function and/or industrial potential, especially relevant in biodiversity-rich biomes.

4.5 Conclusion and Future Directions

This research shows how specific augmentation methods decrease the functional annotation gap in enzyme classification. The most effective method to enhance annotation capabilities at scale while maintaining biological soundness involves codon substitution. This approach delivered significant improvements in rare class recall through its biological augmentation methods without requiring extensive architectural changes. The model performance suffered from poorly managed or biologically invalid augmentations when unfiltered GANs and reverse complement augmentation approaches were used, demonstrating the importance of domain-aware augmentation. Together these results suggest that bioinformatic pipelines require function-preserving task-specific augmentation techniques when working with insufficient annotations.

Future research should explore more sophisticated functional discovery methods to advance these findings. The performance of GAN and other generative pipelines may improve through post-generation quality control procedures, such as embedding similarity thresholds and validated conserved motifs to ensure biological plausibility. Finally, the development of adaptive augmentation frameworks using AutoML or reinforcement learning to dynamically determine augmentation rates per class may produce improved per-class performance. Over time, additional experimental annotation with real-world unannotated genes coupled with iterative fine-tuning of functional prediction models will further reduce reliance on synthetic data while increasing trustworthiness and utility in functional prediction tasks for rare functional classes.

REFERENCES

Almagro Armenteros, J. J., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2019). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics*, 35(21), 4049–4056.

Bairoch A., & Apweiler R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1), 45–48. https://doi.org/10.1093/nar/28.1.45

Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., ... & UniProt Consortium. (2023). UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1), D523–D531.

Biller, S. J., Berube, P. M., Dooley, K., Williams, M., Satinsky, B. M., Hackl, T., ... & Chisholm, S. W. (2018). Marine microbial metagenomes sampled across space and time. *Scientific Data*, 5, 180176.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., & Linial, M. (2022). ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8), 2102–2110.

Cao, Y., & Zhang, Y. (2019). Reverse-complement parameter sharing improves deep learning models for genomic sequence classification. *Bioinformatics*, 35(21), 4056–4064.

Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., ... & Webster, N. S. (2019). Scientists' warning to humanity: microorganisms and climate change. *Nature Reviews Microbiology*, 17, 569–586.

Chen, J., Jia, Y., Sun, Y. et al. (2024). Global marine microbial diversity and its potential in bioprospecting. *Nature*, 633, 371–379. https://doi.org/10.1038/s41586-024-07891-2

Dalkiran, A., Rifaioglu, A. S., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Dogan, T. (2018). ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, 19(1), 334.

De Ferrari, L., Aitken, S., van Hemert, J., & Goryanin, I. (2012). Enzyme function classification: the interplay of sequence, structure and function. *Briefings in Bioinformatics*, 13(4), 401–412.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... & Rost, B. (2022). ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE TPAMI*, 44(10), 7112–7127.

Garcia, C.A., Baer, S.E., Garcia, N.S. et al. (2018). Nutrient supply controls particulate elemental concentrations and ratios in the low latitude eastern Indian Ocean. *Nature Communications*, 9, 4868. https://doi.org/10.1038/s41467-018-06892-w

Han, S. R., Park, M., Kosaraju, S., Lee, J., Lee, H., Lee, J. H., Oh, T. J., & Kang, M. (2023). Evidential deep learning for trustworthy prediction of enzyme commission number. *Briefings in Bioinformatics*, 25(1), bbad401.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., Bikard, D., & Marks, D. S. (2021). Generating functional protein variants with variational autoencoders. *Nature Machine Intelligence*, 3, 651–662.

Hoarfrost, A., Apffel, A., Farfañuk, G., Wirbel, J., Crabtree, J., Fierer, N., ... & Pollard, K. S. (2022). Global metagenomic survey reveals antibiotic resistance gene diversity in environmental microbiomes. *mBio*, 13(3), e00593-22.

Hou, J., Wei, W., Zhang, Z., & Song, J. (2023). CARE: Cluster-Aware Representation Learning for Enzyme Function Prediction. *bioRxiv*. https://doi.org/10.1101/2023.04.12.536718

Jha, R., & Hoarfrost, A. (2025). *LookingGlass-2 (Revision 270a2dc)*. Hugging Face.

Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., & Kattner, G. (2021). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Communications*, 12(1), 2973.

Kucera, M., Nagpal, A., Batra, R., & Marks, D. S. (2022). Generative protein design for programmable specificity. *Nature Biotechnology*, 40, 1332–1340.

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically novel uncultured microbial cells dominate Earth microbiomes. *Nature Reviews Microbiology*, 16, 447–460.

Marouf, M., Machart, P., Bansal, V., Kilian, C., Magruder, D. S., Krebs, C. F., & Bonn, S. (2020). Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*, 11(1), 166.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, Chapter 3: 3.1.1–3.1.8.

Price, M. N., Wetmore, K. M., Waters, R. J., Callaghan, M., Ray, J., Liu, H., ... & Deutschbauer, A. M. (2018). Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, 557(7706), 503–509.

Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., ... & Friedberg, I. (2013). A large-scale evaluation of computational protein function prediction. *Nature Methods*, 10(3), 221–227.

Rafi, A. M., Nogina, D., Penzar, D. et al. (2024). A community effort to optimize sequence-based deep learning models of gene regulation. *Nature Biotechnology*.

Repecka, D., Jauniskis, V., Karpus, L., Rembeza, E., Rokaitis, I., Zrimec, J., ... & Zvirbliene, A. (2021). Expanding functional protein sequence spaces using generative adversarial networks. *Science Advances*, 7(30), eabi7626.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15), e2016239118.

Rodriguez, A., Diehl, J. D., Wright, G. S., Bonar, C. D., Lundgren, T. J., Moss, M. J., ... & Clark, P. L. (2024). Synonymous codon substitutions modulate transcription and translation of a divergent upstream gene by modulating antisense RNA production. *PNAS*, 121(36), e2405510121.

Schnoes, A. M., Brown, S. D., Dodevski, I., & Babbitt, P. C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Computational Biology*, 5(12), e1000605.

Silveira, S. d. A., de Melo-Minardi, R. C., da Silveira, C. H., Santoro, M. M., & Meira Jr, W. (2014). ENZYMAP: Exploiting Protein Annotation for Modeling and Predicting EC Number Changes in UniProt/Swiss-Prot. *PLoS ONE*, 9(2), e89162.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., ... & Bork, P. (2015). Ocean plankton. Structure and function of the global ocean microbiome. *Science*, 348(6237), 1261359.

Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., & UniProt Consortium. (2015). UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6), 926–932.

Vanni, C., Schechter, M. S., Acinas, S. G., Barberán, A., Buttigieg, P. L., Casamayor, E. O., ... & Fernàndez-Guerra, A. (2022). Unifying the known and unknown microbial coding sequence space. *eLife*, 11, e67667.

Wen, T., Cai, H., Hu, Y., Yan, Z., He, X., Wu, X., & Lv, Y. (2020). Data augmentation for sequence-based deep learning models in bioinformatics: a review. *Briefings in Bioinformatics*, 21(6), 2217–2230.

Yang, Y., Jerger, A., Feng, S. et al. (2024). Improved enzyme functional annotation prediction using contrastive learning with structural inference. *Communications Biology*, 7, 1690.

Yu, T., Cui, H., Li, J. C., Luo, Y., Jiang, G., & Zhao, H. (2023). Enzyme function prediction using contrastive learning. *Science*, 379(6639), 1358–1363.

Zeng, H., Edwards, M. D., Liu, G., & Gifford, D. K. (2016). Convolutional neural network architectures for predicting DNA–protein binding. *Bioinformatics*, 32(12), i121–i127.

Zhang, Z., Wei, W., Hou, J., & Song, J. (2024). Enzyme function prediction in metagenomic data with class-imbalanced long-tail distributions. *Bioinformatics*, 40(1), 105–115.

Zhenkun Shi, Rui Deng, Qianqian Yuan, Zhitao Mao, Ruoyu Wang, Haoran Li, Xiaoping Liao, & Hongwu Ma. (2023). Enzyme Commission Number Prediction and Benchmarking with

Hierarchical Dual-core Multitask Learning Framework. *Research*, 6, 0153. https://doi.org/10.34133/research.0153

Zou, Z., Tian, S., Gao, X., & Li, Y. (2019). mlDEEPre: Multi-Functional Enzyme Function Prediction With Hierarchical Multi-Label Deep Learning. *Frontiers in Genetics*, 9, 714.

APPENDIX A

Background and Flowcharts

EC Number

Enzyme Commission (EC) number describes the molecular function of a protein and chemical reaction it catalyzes. It has a hierarchical organization, with four digits which describe the molecular function of the protein with increasing specificity for each subsequent digit (Figure 8)

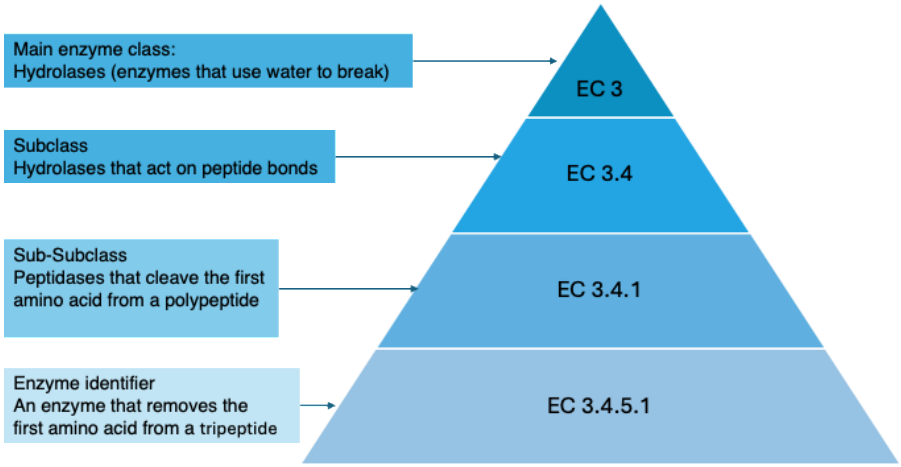

Figure 8: Enzyme Commission (EC) Number Hierarchy. Visual representation of the Enzyme Commission (EC) number system.

Reverse Complement

The DNA sequence augmentation process involved generating reverse complements of DNA sequences. DNA is double stranded and is read from the 5' to 3' direction on both strands, so the reverse complement of a DNA sequence represents the opposing strand of a protein-coding gene

at that site. The reverse complement of DNA sequence data can be derived from sequence reversal followed by nucleotide complements exchanges (A ↔ T, C ↔ G). For example, the reverse complement of the sequence 5'-ATG CCG-3' is 5'-CGG CAT-3'.

Figure 9: Example of DNA Reverse Complement. Illustration of DNA base-pair complementarity used to generate reverse complement sequences. The top strand (5'–3') is paired with its reverse complement strand (3'–5') using standard base-pairing rules.

The augmentation process adds reverse complements of each sequence while preserving their original labels. While these sequences are not generally protein-coding regions that code the same protein, sequencing technologies are equally likely to sequence the reverse strand as the protein-coding strand. This technique produces one deterministic reverse complement for each original sequence, effectively doubling the sample size for every class.

Synonymous Codon Substitution

This augmentation technique utilizes the redundancy of the genetic code by substituting different codons that encode the same amino acids (Figure 10,11). Each 3-mer of DNA encodes one of 20 amino acids; since there are 64 possible codons for the 4 nucleotides in DNA, this results in redundancy in the translation of different codons into the same amino acid. By introducing synonymous mutations into a gene's DNA sequence substituting individual nucleotides to produce different codons that produce the same amino acid this codon substitution process creates new DNA sequences through changes that preserve the encoded protein and its resulting function. The

generation of synonymous variants produces sequence diversity that mirrors potential natural variations in biodiversity.

1st Base	2nd Base: T	2nd Base: C	2nd Base: A	2nd Base: G
T	Phe (TTT, TTC) Leu (TTA, TTG)	Ser (TCT, TCC, TCA, TCG)	Tyr (TAT, TAC) Stop (TAA, TAG)	Cys (TGT, TGC) Stop (TGA) Trp (TGG)
C	Leu (CTT, CTC, CTA, CTG)	Pro (CCT, CCC, CCA, CCG)	His (CAT, CAC) Gln (CAA, CAG)	Arg (CGT, CGC, CGA, CGG)
A	Ile (ATT, ATC, ATA) Met (ATG)	Thr (ACT, ACC, ACA, ACG)	Asn (AAT, AAC) Lys (AAA, AAG)	Ser (AGT, AGC) Arg (AGA, AGG)
G	Val (GTT, GTC, GTA, GTG)	Ala (GCT, GCC, GCA, GCG)	Asp (GAT, GAC) Glu (GAA, GAG)	Gly (GGT, GGC, GGA, GGG)

Figure 10: Codon Table for Amino Acid Translation. Standard genetic codon table showing all 64 nucleotide triplets (codons) and their corresponding amino acids or stop signals. Each codon is composed of three DNA bases.

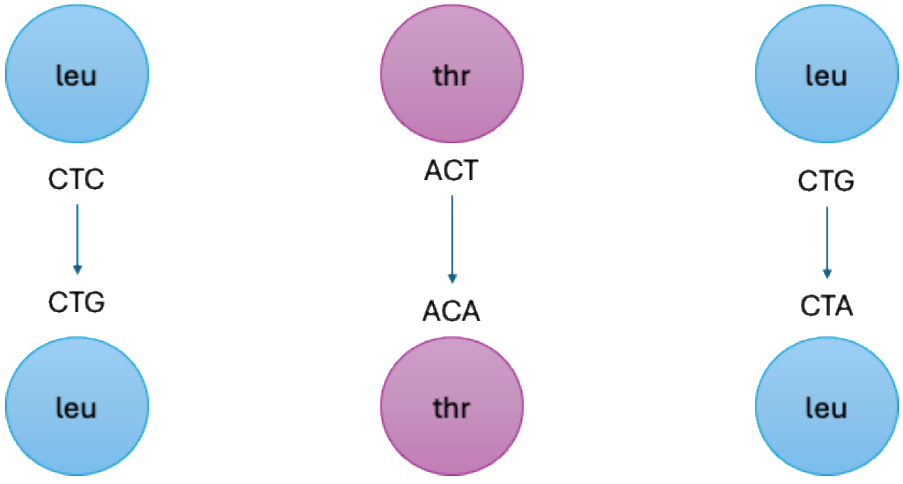

Figure 11: Example of Synonymous Codon Substitutions. Illustration of synonymous substitutions that preserve amino acid identity.

Codon Frequency

Codon frequency refers to how often each of the 64 possible codons is used to encode amino acids in each organism's genome.

GC content

GC content refers to the percentage of cytosine (C) and guanine (G) bases in a DNA molecule. DNA is made up of four nucleotides adenine (A), thymine (T), guanine (G), and cytosine (C). (Figure 12)

CTAGCGCGATCGCAC

$$\text{GC content} = \frac{G}{C}/100$$

$$\text{GC content} = \frac{10}{15}/100 = 66.67\%$$

Figure 12: GC Content Calculation Example. An illustrative example showing how GC content is calculated from a DNA sequence.

GAN

The architecture of the GAN consists of two neural networks, the generator (G) and the discriminator (D). G and D operate via iterative competition, in which G generates artificial data

and D predicts whether data belong to the training set or are produced by G. Over training iterations, G learns to generate more realistic data (Figure 13).

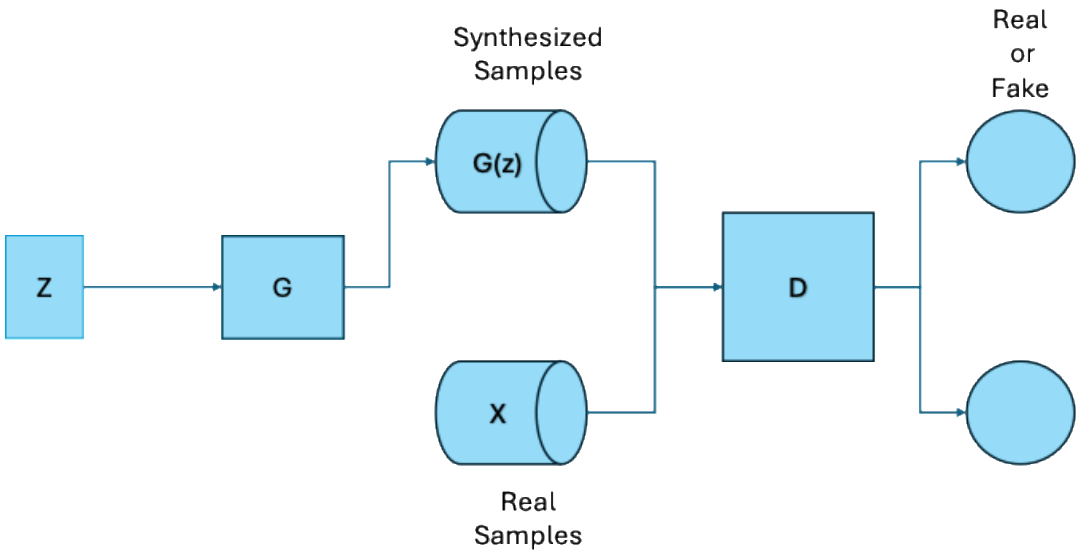

Figure 13: Architecture of a Standard GAN.

cGAN

The architecture of the conditional GAN (cGAN) consists of two neural networks, the generator (G) and the discriminator (D). Similarly to a traditional GAN, G and D operate via iterative competition, however, in a conditional GAN G generates artificial data conditioned on specific inputs such as class labels, GC content, or codon frequency and D predicts whether a given data instance and its associated condition originate from the training distribution or were produced by G. Over successive training iterations, G learns to generate increasingly realistic data that not only resemble true samples but also reflect the desired biological or structural conditions of specific classes. This enables controlled data synthesis, especially useful in domains with class imbalance or limited annotated examples (Figure 14).

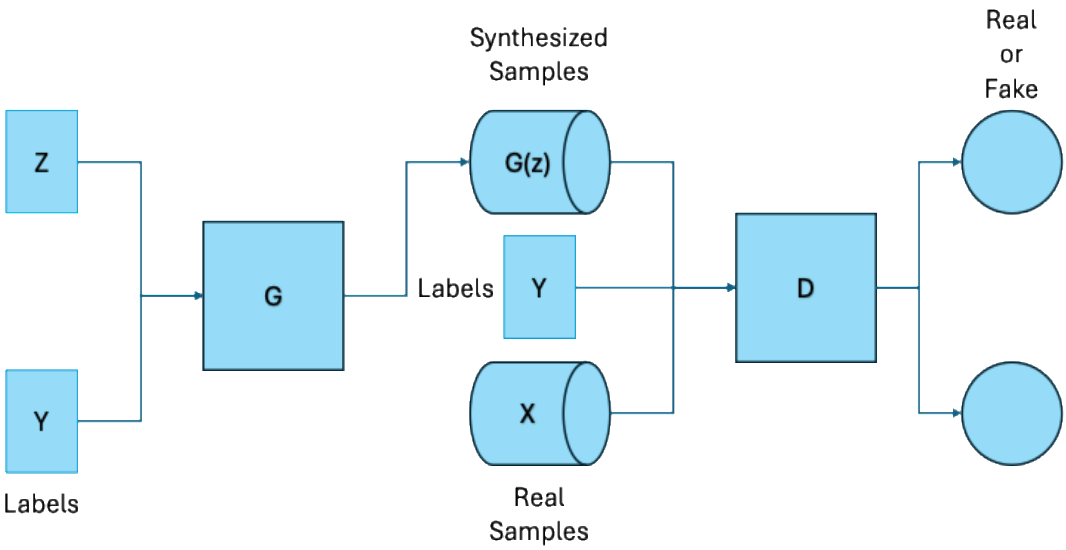

Figure 14: Architecture of a Standard cGAN.

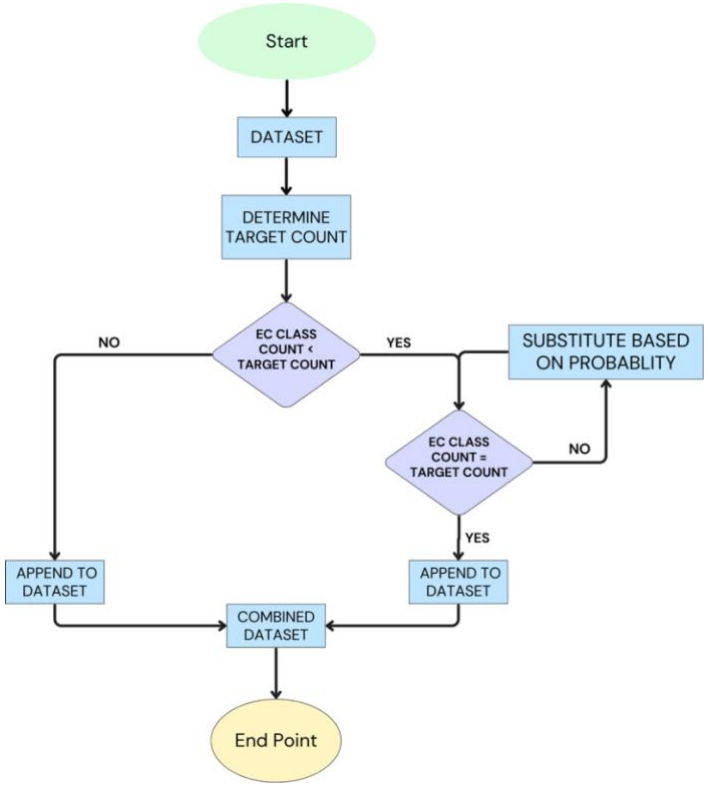

Figure 15: Codon Substitution Augmentation Workflow. Flowchart representing the codon substitution augmentation process. For EC classes with fewer than the target number of samples, additional sequences are generated using probabilistic synonymous codon substitution and appended to the dataset until target support is reached.

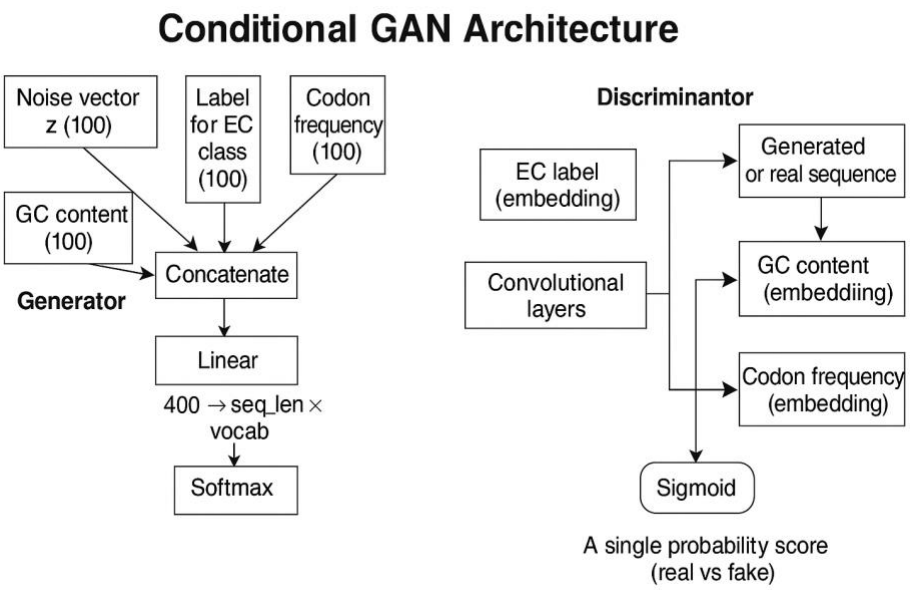

Figure 16: Conditional GAN Architecture. Schematic representation of the conditional Generative Adversarial Network (cGAN) used for sequence generation. The generator takes a noise vector and three conditioning inputs−EC class label, GC content, and codon frequency−all projected into 100-dimensional embeddings and concatenated before transformation into nucleotide sequences. The discriminator receives either a real or generated sequence, alongside the same conditioning inputs, and outputs a probability score indicating whether the sequence is real or synthetic.

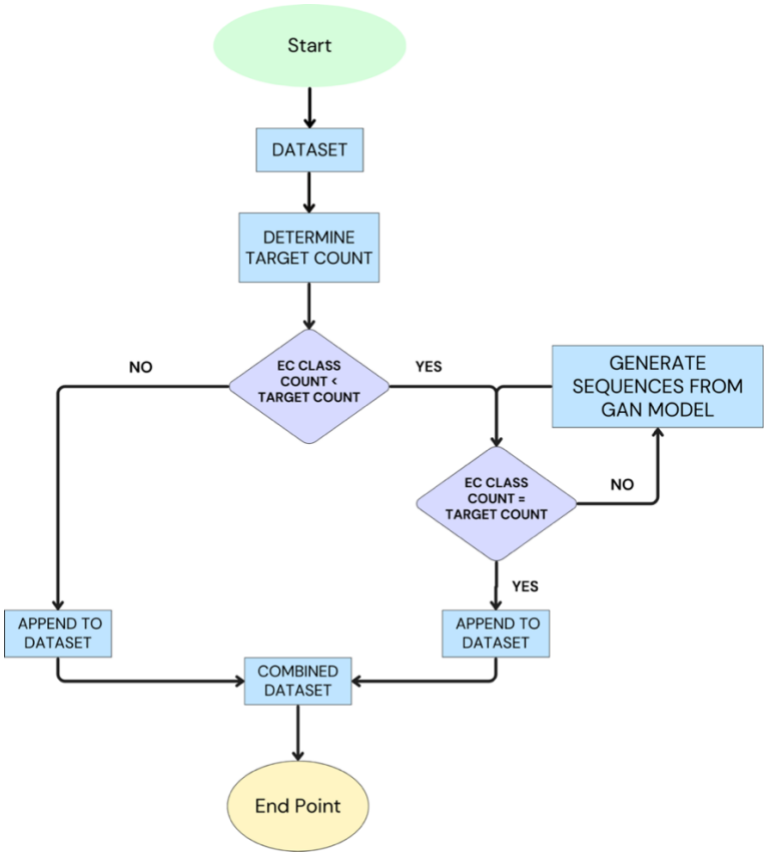

Figure 17: cGAN-Based Sequence Augmentation Flow. Flowchart illustrating the cGAN-based data augmentation process. For each enzyme class (EC) with a sample count below the defined target threshold, synthetic DNA sequences are generated using the trained GAN model.

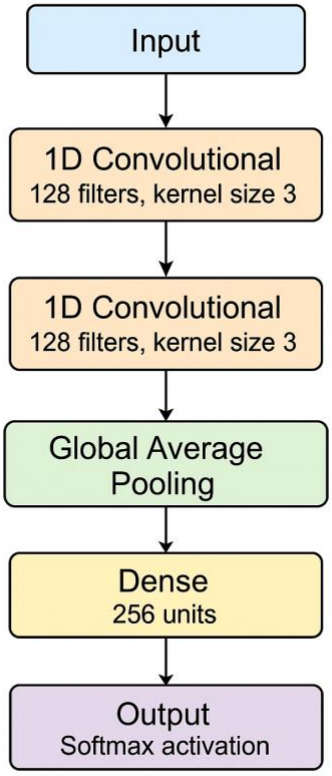

Figure 18: CNN-Based Sequence Classification Model. Architecture of the 1D convolutional neural network used for enzyme function prediction.