Supervised Machine Learning Approaches for Predicting Polysaccharide Utilization Loci in Bacterial Genomes

Ben Rosenthal (s3625389)

Supervisors:
Fons Verbeek (LIACS) & Georg Zeller (LUMC)

Daily Supervisor:
Raymund Hackett (LUMC)

Bachelor Thesis– Bioinformatics

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 1, 2026

Abstract

Polysaccharide utilization loci (PULs) are genomic regions involved in the uptake and degradation of complex carbohydrates by bacteria. Despite the important role of gut microbial carbohydrate metabolism in host–microbe interactions and overall human health, experimentally validated PUL annotations remain limited, prompting the development of computational prediction methods. In this study, we investigated supervised machine learning approaches for PUL prediction and explored the impact of dataset construction, feature representation, model architecture, and taxonomic bias. Several models were evaluated on a set of 958 experimentally verified PULs from 288 genomes: the Conditional Random Fields-based model GECCO, the genomic language models GeneCAT and Bacformer, and the protein language model ESMC. We addressed the incomplete nature of experimental annotations by identifying 1225 uncertain genomic regions, termed cryptic PULs, using a rule-based prediction tool and homology-based searches, and excluding them from the negative class during training. Masking cryptic PULs substantially improved performance across the board. The effect of incorporating Carbohydrate-active enzyme annotations to the feature representation used by GECCO and GeneCAT was also assessed, providing modest performance gains for GeneCAT and a small decrease for GECCO. Overall, models based on domain-based genome representations outperformed approaches based on protein sequences or protein language model embeddings. The Transformer-based GeneCAT model showed the strongest results after finetuning (AUPRC 0.55 for experimental PULs, 0.64 for cryptic, 0.70 for both), and outperformed ESMC and Bacformer in a zero-shot setting, (0.44 vs 0.18 and 0.17, respectively). Models trained exclusively on *Bacteroidetes* generalized poorly to other phyla, while the reverse setting yielded substantially better performance. Finally, our models predict novel putative PULs in diverse gut bacteria that currently lack experimental annotations, showing that machine learning can be an effective guiding tool for biological discovery.

Contents

List of Abbreviations

ABC ATP-Binding Cassette

ANI Average Nucleotide Identity

AP Average Precision

AUPRC Area Under the Precision-Recall Curve

AUROC Area Under the Receiver Operating Characteristic Curve

BGC Biosynthetic Gene Cluster

CAZyme Carbohydrate-Active Enzyme

CDS Coding DNA Sequence

CGC CAZyme-containing Gene Cluster

CRF Conditional Random Field

GH Glycoside Hydrolase

gLM Genomic Language Model

gpPUL Gram-positive Polysaccharide Utilization Locus

(p)HMM (Profile) Hidden Markov Model

ML Machine Learning

pLM Protein Language Model

PUL Polysaccharide Utilization Locus

UMAP Uniform Manifold Approximation and Projection

1 Introduction

This bachelor's thesis was written during an internship at the Zeller research group, located at the Leiden University Medical Center (LUMC), under supervision of Georg Zeller and Raymund Hackett. The project was initiated at the Leiden Institute of Advanced Computer Science (LIACS) supervised by Fons Verbeek.

This introduction starts with the main problem statement and provides the motivation for this study. Following that, the main research question and its sub-questions are formulated and the section concludes with a global overview of the thesis.

1.1 Problem Statement and Motivation

The human gut microbiome is a complex ecosystem that plays an important role in overall health and a variety of diseases. Bacteria in the gut normally form a symbiotic relationship with the human host, with some species feeding on nutrients that cannot be broken down in the human digestive tract. Bacterial carbohydrate metabolism in the gut is an important factor in the functional profile of the microbiome. The metabolism of carbohydrates by gut microbes produces beneficial metabolites that may contribute to overall gut health [LTB22, vdHW21]. On the other hand, microbes can also digest host-glycans, compromising the integrity of the gut mucus barrier, often observed in diseases such as colorectal cancer and inflammatory bowel disease [PZS+22, ZTV+14].

Carbohydrate-active enzymes (CAZymes) enable microbes to modify and digest complex polysaccharides [DS08]. Genes coding for or related to CAZYme activity often co-localize into genomic regions where they follow specific organizational structures, forming polysaccharide utilization loci (PULs) [GTD+17, BGB21]. PULs were first characterized in bacteria of the Gram-negative *Bacteroidetes* phylum, many of which are abundant in the human gut. Although comprehensive study of PULs has been mainly aimed at this phylum, CAZyme-rich gene clusters with functional similarities to PULs have been observed in other bacterial phyla, primarily in Gram-positive *Firmicutes* where they are referred to as Gram-positive PULs (gpPULs) [OSML+16, BMB+07, BGB21].

Our understanding of microbiomes has greatly advanced with next generation sequencing, allowing for thorough and accurate taxonomic profiling using a plethora of computational methods. While information on microbial abundances on a species or strain level can provide many insights into the microbiome, functional profiling allows for a better understanding of the overall biological activity [BMF+13]. Accurately identifying PULs in bacterial genomes could contribute to functional microbiome profiling efforts, by adding functional information on bacterial carbohydrate metabolism and providing a foundation for identifying possible metabolic mechanisms underlying microbial signatures. Current methods for PUL identification are inherently limited to bacteria from the *Bacteroidetes* phylum, taking a rule-based approach defined around the canonical PUL structure. These methods (from PULDB and PULpy) rely on identifying a sequential pair of *susC* and *susD* gene homologs, coding for proteins involved in binding and transporting polysaccharides into the bacterial cell, then extending the PUL if any CAZymes are found within a predefined context window [TLGH15, SARW18a]. While this structure is ubiquitous and highly conserved in *Bacteroidetes*, PULs from other phyla can be a lot more diverse. For example, gpPULs do not contain a susC-homolog, instead employing different transport mechanisms [BGB21]. This limits

these methods to bacteria from a single phylum, and reduces their potential for discovering novel PULs and PUL structures.

Machine learning (ML) methods have increasingly been applied to problems in microbial genomics, where supervised learning is commonly used to predict functional properties from sequences and sequence-derived features, such as protein domain composition [LN15]. ML methods have been applied to predicting protein functions from amino-acid sequences [BV20], predicting antimicrobial resistance from whole-genome sequencing data [RCD+22] and for accurately predicting biosynthetic gene clusters from bacterial genomes [CLF+21]. Despite the growing success of ML in microbial genomics, the application of supervised learning to PUL prediction remains unexplored, even though such an approach may provide a more flexible and generalizable alternative to existing rule-based methods. Supervised models can learn patterns directly from annotated genomic regions and infer whether a region is likely to represent a PUL based on genomic and functional characteristics, instead of relying on predefined rules. This could allow models to find PULs that deviate from the canonical *susC-susD* organization observed in *Bacteroidetes*, which could improve generalization across taxa and potentially support identifying promising novel PUL architectures.

1.2 Research Questions

The main goal of this project is to explore and evaluate the application of supervised learning approaches to the prediction of PULs. This constitutes the main research question:

Main RQ: How effective are machine learning techniques for predicting polysaccharide utilization loci?

Since the absence of PUL annotations does not necessarily imply lack of CAZyme activity, careful definition of negatives is critical for the success of the classification task. Therefore, the initial challenge in this project relates to setting up a clear and effective strategy to build a dataset for binary classification with realistic negatives. This leads to the first sub-question:

RQ1: How can a binary classification dataset be constructed despite the incomplete nature of experimental PUL annotations?

Additionally, feature engineering and selection are important aspects in supervised learning, especially when dealing with genomic data, which are inherently structured as long strings of letters. Models can learn from the DNA sequences, from translated protein sequences (in the case of protein language models) or from any other meaningful and consistent genome annotations. Among others, this study evaluates two models that use Pfam protein domain annotations as their feature representation. For these models, we wish to explore whether the inclusion of CAZy annotations in the feature set would enhance performance, as these are directly relevant to carbohydrate metabolism. This can be formulated as the second sub-question:

RQ2: Does the inclusion of CAZy annotations as features improve performance for models which utilize domain annotations?

Model complexity is also an important aspect to explore, with paradigms shifting towards the use of large foundation models and transfer learning. This raises the question of whether increased model complexity contributes to better performance on this task. Additionally, given the large

diversity of genome representations used for ML, we wish to evaluate which representation works best for PUL prediction. Formulated as the third sub-question:

RQ3: How do models with different architectures and genome representations compare in the context of PUL prediction?

Finally, as both experimental data and existing prediction methods are heavily biased towards *Bacteroidetes*, we wish to evaluate whether supervised learning could overcome this limitation. An effective model should be able to generalize properly to other phyla of bacteria, to allow for accurate predictions on a much wider range of taxa. This is formulated as the final sub-question:

RQ4: Given that experimental work is mostly focused on *Bacteroidetes*, can models generalize to other phyla?

1.3 Thesis Overview

This thesis is divided into five main chapters. Section 2 discusses the relevant theoretical background and related work, making a distinction between the biological and computational background. Section 3 specifies the datasets, software and hardware used for this project, followed up by Section 4, that describes the implementation of all methods for the dataset construction, model training, evaluation and interpretation. Then, the results are outlined and thoroughly discussed in Section 5, and the thesis is concluded by Section 6, in which the research questions are answered, and the study's limitations and suggestions for future work are discussed.

2 Background & Related Work

This section is dedicated to providing the background knowledge necessary for the following sections. It is divided into two subsections: the first is a short literature review on the biological background and existing PUL prediction methods. The second relates to the computational background and provides an overview of important machine and deep learning concepts, alongside the models used in this research.

2.1 Biological Background

In this subsection we first introduce microbial carbohydrate metabolism, focusing on carbohydrate active enzymes (CAZymes), followed by an overview of polysaccharide utilization loci (PULs). We conclude with an outline of current computational methods for predicting PULs using genomic sequences.

2.1.1 Microbial Carbohydrate Metabolism

Complex carbohydrates are abundant in the biosphere and serve as energy sources for many microbial communities. This family of molecules is comprised of oligo- and polysaccharides, which are built out of monosaccharide subunits in a variety of configurations. Due to the complexity of these molecules, microbes require a vast arsenal of enzymes to metabolize them, so called **C**arbohydrate-**A**ctive en**Z**ymes, referred to as CAZymes. In the 1990s, Bernard Henrissat initiated a protein sequence-based CAZyme classification system [DS08], which had been developed into the CAZy database [DGD+22]. The database covers five CAZyme classes, each containing numerous different families. The largest class is of Glycoside Hydrolases (GHs), which catalyze hydrolysis and glycosidic bond rearrangement. The other classes include Glycosyl Transferases, Polysaccharide Lyases, Carbohydrate Esterases, Carbohydrate-Binding Modules, and Auxiliary Activity enzymes.

While fibers are abundant in the human diet, the human genome contains a very limited number of CAZymes that allow us to only digest starch, sucrose and lactose [KAG+13]. In contrast, human gut microbiota can effectively utilize a large diversity of dietary fibers, with some bacterial genomes encoding for more than 300 CAZymes (*B. ovatus* and *B. intenstinalis*) [KAG+13]. The metabolic capacity of the gut microbiome plays an important role in the host's health. For example, the fermentation of carbohydrates by gut bacteria produces metabolites such as short-chain fatty acids, contributing to the epithelial barrier integrity and overall gut health [vdHW21]. However, many gut microbes encode for CAZymes that can degrade host-derived polysaccharides. This is most concerning when these CAZymes target mucins, glycans that support the lining of the intestinal epithelium [KAG+13]. When dietary fibers are scarce, bacteria can switch to digesting mucus, potentially compromising the integrity of the gut mucus barrier, which has been correlated with diseases such as colorectal cancer and inflammatory bowel disease [DSK+16, PZS+22, ZTV+14].

2.1.2 Polysaccharide Utilization Loci

In order to efficiently metabolize complex carbohydrates, many bacterial genomes show an organization of CAZyme-encoding and related genes into cohesive genomic regions. These loci often also include transport proteins, carbohydrate-binding proteins and transcriptional regulators that col-

lectively enable microbes to detect, bind, transport and degrade polysaccharides. This organization is most commonly found in Gram-negative bacteria of the *Bacteroidetes* phylum. These bacteria possess a periplasmic space between their inner and outer cellular membranes, that required the evolution of specialized systems for polysaccharide uptake and degradation [BGB21, GTD+17]. The term polysaccharide utilization loci (PULs) was first introduced in 2006 to describe these gene clusters, initially characterized in the abundant human gut symbiont *Bacteroides thetaiotaomicron* [BMG06]. Since then, PULs have been extensively studied in other members of the *Bacteroidetes* phylum, where they appear particularly frequently in the *Bacteroides* genus [BGB21, GTD+17].

The starch utilization system (Sus) of *Bacteroides thetaiotaomicron* is the most well-studied PUL. It is an operon consisting of eight genes, encoding for three glycoside hydrolases (SusA, SusB, SusG), a TonB dependent transporter (susC), three starch-binding outer membrane lipoproteins (susD, SusE, SusF) and a regulator protein (SusR) (see Figure 1). Extracellular starch is initially bound by the outer membrane proteins encoded by susD, SusE and SusF, while the SusG-encoded CAZyme performs partial hydrolysis at the cell surface. The resulting oligosaccharides are transported through the susC-encoded transporter into the periplasm. Afterwards, the glycoside hydrolase CAZymes encoded by SusA and SusB degrade them into smaller sugars, mono- or disaccharides, that can be utilized by the bacterium. SusR regulates the expression of the operon, by responding to the presence of maltose in the periplasm. All gut-residing *Bacteroidetes* have some form of Sus-like systems that are used to target a wide variety of polysaccharides, with some species dedicating up to 18% of their genomes to PULs [FCK16, MLC+11].

The Sus operon formed the basis for formulating a canonical architecture that is commonly shared across PULs in *Bacteroidetes*. Canonical PULs follow a fairly regular structure, starting with a sequential pair of susC and susD homologs, followed by one or more CAZymes, depending on the complexity of the substrates [GTD+17]. CAZymes of the GH family are the most common, but some PULs have also been found to contain enzymes such as proteases or sulfatases. The latter is important for marine bacteria feeding on highly sulfated algae polysaccharides and also for utilizing host glycans such as mucins by gut bacteria [GTD+17, LBB+22].

Although PULs were originally defined in *Bacteroidetes* by the presence of tandem susC-susD homologs and CAZymes, similar gene clusters have also been identified in bacteria of other

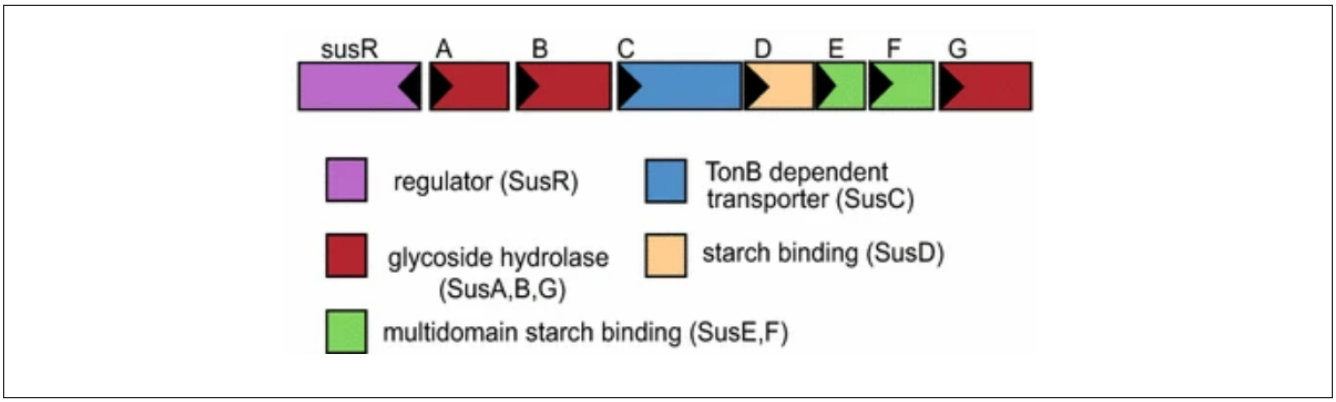


Figure 1: Overview of the starch utilization system (Sus) genes in *Bacteroides thetaiotaomicron* (adapted from [FCK16]).

phyla. In several species, including Gram-negative *Proteobacteria* associated with plant biomass degradation, clusters containing genes encoding for TonB-dependent transporters (TonB-Dependent Transporters) and CAZymes have been observed without the addition of a *susD* homolog or any connected regulatory systems. These clusters nevertheless show coordinated carbohydrate utilization functions and have been described as carbohydrate utilization loci-containing TonB-Dependent Transporter [BMB+07]. Furthermore, carbohydrate utilization systems resembling PULs have also been identified in Gram-positive *Firmicutes* bacteria, despite their fundamentally different cell membrane structure [WLBF14, OSML+16]. For example, the gut-abundant Gram-positive *Roseburia intestinalis* is capable of competing with the PUL-enriched *Bacteroides ovatus* for plant cell wall polysaccharides such as β-Mannans [LRLM+19]. These so-called Gram-positive PULs (gpPULs) do not follow the canonical *Bacteroidetes* PUL structure, but still typically contain CAZymes, carbohydrate transport systems and transcriptional regulators. gpPULs do not encode for TonB-Dependent Transporters and thus do not contain susC-homologs, due to the lack of an outer-membrane structure in Gram-positive bacteria. Instead, gpPULs can make use of three different transport mechanisms: either ATP-binding cassette (ABC) transporters, carbohydrate phosphotransferase system transporters or major facilitator superfamily transporters [BGB21].

The existence of alternative PUL architectures across bacterial phyla shows that PULs are more structurally and functionally diverse than canonical *susC-susD*-based *Bacteroidetes* PULs [GTD+17]. For this reason, the term had been expanded to include gpPULs and other similar gene clusters that lack the susC-susD structure. This project embraces this broader definition.

2.1.3 Computational Identification of PULs

Experimentally identifying and verifying the presence of PULs is a complex and labor-intensive process. Experiments typically involve a combination of (meta)genomic, transcriptomic and biochemical analyses of individual CAZymes [GTD+17]. As a result, a number of computational methods have been developed to automatically predict PULs from purely genomic data.

In 2015, Terrapon et al. presented a computational approach to predict PULs in *Bacteroidetes* species, as part of the PULDB database [TLGH15]. The method follows a rule-based strategy centered around the canonical *susC-susD* architecture. First, tandem pairs of *susC* and *susD* homologs are identified across the genome. Next, the surrounding genomic region is iteratively extended if neighboring genes encode CAZymes or if additional *susC* or *susD* homologs are located. The resulting regions are annotated as predicted PULs and integrated into the PULDB database, which contains predictions from 67 *Bacteroidetes* genomes [TLGH15]. One important limitation is that the original PULDB prediction procedure is not fully reproducible, as several implementation details and manual curation steps were not publicly documented. Additionally, the absence of a publicly available software implementation limits the method's practical utility beyond the set of predicted PULs deposited in PULDB [TLGH15].

Another prediction tool called PULpy was introduced in a pre-print from 2018 by Steward et al., as an attempt to reverse-engineer the PULDB prediction method and produce a publicly available tool that can be used on new genomes [SARW18a]. Predictions involve first locating *susC-susD* pairs and then utilizing a sliding window of five genes in both the upstream and downstream directions to search for any CAZyme-encoding genes within that window, while limiting the intergenic distance to 500 basepairs. If at least one CAZyme is found within this genomic context, then the window is

extended by five genes again and the process is repeated until no more CAZymes are identified, or the intergenic distance exceeds 500 basepairs [SARW18a].

As part of the dbCAN2 database, the Python-based tool CGC-Finder was introduced in 2018 to find CAZyme-containing Gene Cluster (CGC). This term was coined by the involved researchers to describe physically linked gene clusters that encode at least one CAZyme, one transporter, one transcriptional regulator, and one signaling transduction protein [ZYH+18, HZW+18]. While this tool does not rely on finding a tandem susC-susD pair, and is therefore not limited to the *Bacteroidetes* phylum, the authors emphasize that the way CGCs are defined is inherently different than PULs in literature. Hence, a large-scale comparison of CGCs with literature-derived PULs is not performed. Moreover, the method still relies on (albeit user-defined) rules, limiting its ability to capture the complexity and structural diversity that characterize PULs.

In conclusion, all previously developed prediction methods for PULs and related clusters such as CGCs rely on a rule-based approach. In the case of PULDB and PULpy, the rules are intrinsically limited to *Bacteroidetes*, meaning that these tools are not capable of predicting PULs in other bacterial phyla. These limitations necessitate the exploration of more flexible and generalizable computational approaches for PUL prediction.

2.2 Computational background

This section focuses on the use of machine learning within the field of genomics. We first introduce this topic, and further discuss the use of Transformer-based models and transfer learning paradigms. Then, we narrow the scope to genomic language models and afterwards provide an overview of the specific models selected for this project.

2.2.1 Machine Learning in Genomics

Machine learning (ML) is a field in AI and data science centered around developing algorithms that learn patterns from data or experience, and then using them to make inferences on new unseen data. ML algorithms are roughly divided into supervised learning, where models learn from labeled data with a ground truth, and unsupervised learning, where patterns are extracted from data without the use of labels. A related paradigm is self-supervised learning, where labels are generated from the input data itself. In this setting, models are trained on tasks such as predicting missing or masked parts of the input from context, in order to learn useful representations from large amounts of unlabeled data. This research project is concerned with using supervised learning for binary classification, a task where the model learns to classify instances into two different classes based on labeled examples.

Deep learning is a sub-field of ML concerned with the application of artificial neural networks. These models consist of interconnected layers of computational units called neurons. Unlike classical ML methods, which often rely on manually engineered features, deep learning models can learn complex features directly from raw input data. The network can build increasingly abstract representations as each layer computes a non-linear transformation of the previous layer's output. Training a neural network involves optimizing the weights of these transformations across layers, most often using backpropagation. Building on this foundation, a plethora of different deep learning model architectures have been developed over the years, often engineered for specific types of tasks. Deep

learning models have achieved unparalleled performance in fields such as computer vision and speech processing, but also in drug discovery and genomics [LBH15].

With vast improvements in next generation sequencing methods and overall computational capabilities, ML and deep learning approaches are becoming common in many biological fields. Specifically in genomics and genetics, where datasets can be very large and complex, ML is highly applicable [RM22] [LN15]. Models are often trained on DNA sequences and/or any other features that can be derived from them for the general task of genome annotation, which involves labeling DNA segments in the genome with a corresponding function, such as gene functions, regulatory elements and more [RM22]. In microbial genomics, supervised learning had been previously applied in predicting functional information from amino-acid sequences, for predicting antimicrobial resistance from full genome sequences of bacteria, and for accurately predicting biosynthetic gene clusters from bacterial genomes [BV20, CLF+21, RCD+22]. Deep learning is also becoming prominent in microbiome research, as highlighted in a 2025 review by Pryzmus et al. [PRMS+25]. Deep learning is capable of automating and improving aspects such as the detection and quantification of functional genomic elements in metagenomes and has also been applied for predicting biosynthetic gene clusters [PRMS+25, HPP+19].

2.2.2 Transformers and Transfer Learning

The introduction of the Transformer architecture in 2017 revolutionized the field of natural language processing and ushered in the current era of AI. Transformers are neural networks that employ attention mechanisms, allowing them to very effectively capture relations and dependencies between individual tokens in sequences [VSP+17]. Unlike previous sequence models, Transformers are capable of capturing long-range interactions while remaining computationally efficient, making them particularly effective for learning representations in extremely large datasets. Although originally gaining the spotlight in natural language processing, Transformers can be applied to a much wider range of domains, including genomics [BYA+25].

Due to the sheer size and computational complexity of the training process, Transformer-based language models are hardly ever trained for specific tasks. Instead, they are pretrained on vast generalized datasets in a self-supervised manner, to learn general purpose representations from as much (unlabeled) data as possible. Such a pretrained model, also referred to as a foundation model, can be finetuned further for a more particular application, adapting these representations to a specific task using a much smaller labeled dataset [BYA+25]. The process of finetuning large pretrained models, called transfer learning, is becoming an increasingly common paradigm in ML.

It is important to distinguish between encoder and decoder architectures. Encoder models are trained to learn contextual representations of input sequences, called embeddings. These models are therefore primarily used for tasks such as classification or regression. Decoder models, in contrast, are trained to generate sequences by predicting the next token given the preceding context. This distinction influences how foundation models are applied in practice. In many cases, foundation models are used in a zero-shot setting, where they are applied to a previously unseen task without any task-specific training. For generative models, this involves performing some tasks directly through prompting, where instructions are provided as part of the input [LCC+25]. In other cases, this is frequently done by extracting the model's learned representations and using them to train a simpler model for a specific task. It is also possible to finetune foundation models on a labeled

dataset for a particular task. This is the most computationally expensive form of transfer learning, since it requires the model to re-optimize its weights to a particular task. However, it often also results in better performance, by allowing the model's learned representations to become specialized for that particular task [BYA+25].

2.2.3 Genomic Language Models

In recent years, Transformer-based deep learning models have been gaining traction in the life sciences. Biological sequence-based language models such as DNA language models or protein language models (pLMs) are pre-trained on tremendous amounts of nucleotide or amino-acid sequences with the goal of learning semantic relationships. This renders them capable of analyzing and presumably "understanding" biological sequences [BYA+25]. Generative models such as regLM and ESM-3 can be used to design novel biological sequences, promising immense potential in fields such as drug discovery [LGBE24, HRA+24]. Many other biological language models, based on encoder architectures, are used to extract informative embeddings from biological sequences rather than generate new sequences. Since the objective of this thesis is binary classification rather than sequence generation, encoder-based models are of primary interest. Some examples include the DNA language models DNABERT and DNABERT-2 [JZLD21, ZJL+24], and ESM Cambrian [Tea24], the pLM evaluated in this thesis.

Genomic language models (gLMs) add an additional layer of complexity to the aforementioned models. Instead of training on short context windows of nucleotide or protein sequences, these models learn to place genes within a broader genomic context. The inputs for gLMs can take the form of very long DNA sequences (up to 1 million base-pairs for Evo 2 [BDK+26]), or more frequently a sequence of genes. In the second case, the way in which genes are represented can vary between models. Since gLMs are capable of producing dense contextual representations of genes within a full genome, such models are well suited for predicting and characterizing larger interconnected genomic regions such as PULs.

2.2.4 Selected Models

This thesis compares four different models for the task of PUL prediction: GECCO, GeneCAT, ESM Cambrian, and Bacformer. These models differ in size, architecture, feature representations, and the way contextual information is incorporated (see Table 1).

GECCO (Gene Cluster prediction with Conditional Random Fields) is a machine learning-based method originally developed in the Zeller group by Larralde et al. for identifying Biosynthetic Gene Clusters (BGC's) in genomic data. The method is based on conditional random fields (CRFs), discriminative graphical models that are often used for sequence segmentation and allow for easy model introspection. GECCO can be trained on a sequence of protein coding genes with protein domain annotations and incorporates local genomic context by modeling dependencies between neighboring genes [CLF+21].

GeneCAT is an encoder-based gLM currently under development within the Zeller group. The model represents genomes in a manner similar to GECCO's: ordered sequences of genes annotated with functional domains. It was pretrained on a large collection of high-quality prokaryotic genomes from the ProGenomes3 database using self-supervised learning with a masked multi-label classification

objective. The model looks at windows of genes and captures genomic context using the attention mechanism. Serving as a foundation model, GeneCAT can be finetuned to specific tasks such as PUL prediction (see Supplementary Figure 1).

ESM Cambrian (ESMC) is a pLM that produces embeddings for individual protein sequences, released by EvolutionaryScale in 2024 and described in a 2026 pre-print [Tea24, CHD+26]. The model was pretrained on sequences from UniRef, MGnify, and the Joint Genome Institute (JGI), using a masked language modeling objective, with the goal of producing protein-level embeddings that reflect the biology of proteins. These embeddings are typically aggregated at the gene level and used as input features for downstream prediction tasks. While ESMC can model relationships between individual amino-acids in a protein sequence, it does not capture any broader genomic context.

Bacformer is a gLM (or "contextualized pLM") that utilizes precomputed embeddings from pLMs, such as ESMC, as its input vocabulary. Instead of directly encoding raw sequences or domain annotations, each gene is represented by the average embedding of its corresponding protein sequence. Bacformer was pretrained on 1.3 million bacterial genomes represented as an ordered sequence of proteins, using a masked language modeling objective. Instead of having to predict exact embeddings, the model learned to predict protein families of masked proteins, based on genomic context. The model can process complete genomes, which presumably allows it to capture genome-wide organizational and evolutionary patterns [WTN+25].

Table 1: Summary of models used in this thesis and their properties, including model architecture, input representation, pretraining data, model scale, and genomic context modeling.

Characteristic	**GECCO**	**GeneCAT**	**ESM-C**	**Bacformer**
Publication	Pre-print 2021 [CLF+21]	Unpublished	Pre-print 2026 [CHD+26]	Pre-print 2025 [WTN+25]
Architecture	Conditional Random Fields	Transformer	Transformer	Transformer
Input representation	Ordered sequences of genes with domain annotations	Ordered sequences of genes with domain annotations	Amino-acid sequences	Ordered sequences of pLM embeddings
Pretraining data	—	ProGenomes3	UniRef MGnify JGI	MGnify v2.0 SPIRE NCBI RefSeq GenBank
Model scale	—	60M params	600M params	300M params
Genomic context	Local (CRF)	Gene window attention	None	Genome-wide attention

3 Material & Methods

3.1 Data

All of the data obtained for this project was taken from open-access databases. First, PUL annotations were obtained through the databases PULDB [TLD+18] and dbCAN-PUL [AZY+21]. Next, nucleotide sequences were downloaded from the NCBI Nucleotide database.

3.1.1 PULDB

The Polysaccharide Utilization Loci database (PULDB) was launched in 2015, containing PULs predicted from genome sequences using a rule-based method, alongside some PULs previously reported in literature [TLGH15]. PULDB was expanded in 2018 to include more literature-derived PUL data, but the database remains limited to PULs from the *Bacteroidetes* phylum [TLD+18]. The exact methods and criteria used by the authors to curate these literature-derived PULs are not stated, besides the mention of "a continuous literature survey".

Due to the lack of a download or external API interface in PULDB, we ran a custom web-scraping script using the Python package Beautiful Soup 4 (version 4.14.3) in February 2026. We extracted genomic coordinates and NCBI sequence identifiers for all PULs in the database and filtered the results to only include literature-derived PULs.

3.1.2 dbCAN-PUL

dbCAN-PUL is an online database for literature-derived PULs, which is connected to the dbCAN3 server, a service used for automated CAZyme annotation [AZY+21, ZGY+23]. In contrast to PULDB, dbCAN-PUL contains PULs across various bacterial phyla. PULs in this database were manually curated from the literature under the condition that experimental methods were used to determine that at least one gene from the PUL is involved in carbohydrate utilization. The database provides information on the PUL sequence, genomic coordinates, PUL substrate, experimental characterization methods and the associated NCBI taxonomy identifier.

The latest version of dbCAN-PUL was downloaded in February 2026, and genomic coordinates and sequence identifiers were extracted from all PUL annotations from the Excel sheet with the most recent upload date (`dbCAN-PUL_Feb-2025.xlsx`, uploaded 27-06-2025).

3.1.3 NCBI Nucleotide

The NCBI nucleotide database is a publicly available collection of genome, gene, and transcript sequences from sources such as GenBank, RefSeq, and PDB. The NCBI sequence identifiers obtained from PULDB and dbCAN-PUL (Sections 3.1.2 & 3.1.1) were used to retrieve the corresponding nucleotide sequences via Entrez (Section 3.2.1). For each PUL, the full contig/genome containing the locus was downloaded. Sequences were retrieved in both Fasta and GenBank formats, and were used downstream for preprocessing.

3.2 Software

This section describes the software tools and packages used for this project. Tools are presented roughly in the order in which they were applied. For each tool, we report the version used and any specific adaptations or deviations from default settings. The specific ways in which these tools were used in this project are described in Section 4.

3.2.1 NCBI Entrez & Biopython

Entrez is a database retrieval system developed and maintained by the National Center for Biotechnology Information (NCBI). It allows for querying sequence data, genomic mapping information, 3D structure data, PubMed and more. Entrez was used in this project to query the NCBI Nucleotide database, through the interface provided by the Biopython python package (version 1.86). Specifically, the `Entrez.efetch` method was used to retrieve nucleotide records in batch.

3.2.2 BLAST

BLASTn is generally used to identify regions of nucleotide similarity between query sequences and a reference database. The algorithm performs local alignments by identifying "words" (short exact matches) and then extending these into high-scoring alignments [AGM+90]. During this project, BLASTn was used for various purposes, both in a remote and local setting (see section 4). We used the Biopython NCBIWWW module for remote BLASTn searches against the *nt* database, and the BLAST+ binaries (version 2.17.0) for searches against locally stored sequences.

3.2.3 Cblaster

Cblaster is a tool used for identifying homologous gene clusters in local or remote databases [GBvW+21]. The search algorithm works by first querying protein sequences against a reference database using either remote NCBI databases with BLASTp, or a local genome database with DIAMOND. Hits are filtered using specific hit quality thresholds and grouped by subject protein, genomic scaffold and organism. Clusters are identified if subjects are sufficiently close (pre-specified intergenic distance) and if the formed clusters contain a pre-specified minimum number of genes. For this project, we used Cblaster (version 1.4) with a local DIAMOND database created using the Cblaster makedb module (see section 4).

3.2.4 OrthoANI

OrthoANI (Average Nucleotide Identity by Orthology), is an algorithm used to compute average nucleotide identity between genomic sequences [LOKPC16]. This algorithm was developed to account for ANI calculations being non symmetrical, by using the concept of orthology. To calculate the orthoANI score between two genomes, the following steps are taken: (i) both sequences are first fragmented into 1020-bp long fragments, (ii) followed by reciprocal BLASTn searches between all fragments to identify orthologous pairs, (iii) after which the average nucleotide identity values are calculated for all reciprocal BLASTn hits. By only taking orthologous fragment pairs into account, the algorithm results in identical reciprocal ANI values and also runs faster than original ANI implementations [LOKPC16]. To calculate pairwise ANI scores for all genomes in the dataset we used PyOrthoANI (version 0.7.0), the Python implementation of the OrthoANI algorithm [LZC25].

3.2.5 PULpy

PULpy is a rule-based PUL prediction pipeline, based on the methodology presented by the prediction tool used in PULDB [SARW18b, TLGH15]. PULpy was released in 2018 as a Snakemake pipeline and the predictions showed significant overlap with the predictions in PULDB.

The latest version of PULpy was downloaded from the GitHub repository [SARW18b]. The tool did not run immediately and some adaptations were made to the code. First, the Snakefile and all files in the `scripts` directory contained mismatching indentations (spaces and tabs), so all tabs were replaced by spaces. An additional custom script was written to automate running PULpy, including all necessary preparations that are provided in the repository documentation. Finally, additional snakemake and perl dependencies were added to the environment definition at `envs/PULpy.yaml`.

3.2.6 GTDB & GTDB-Tk

The Genome Taxonomy Database (GTDB) is a standardized genome phylogeny based microbial taxonomy [PCC+20]. GTDB-Tk is a software toolkit for the classification of new genomes within the GTDB framework [CMHP22]. The classification algorithm is based on placing the query genome in the GTDB reference tree and computing its Average Nucleotide Identity (ANI) with reference genomes. For this project, GTDB-tk (version 2.6.1, with GTDB release 226) was used to generate taxonomic annotations of all genomes in the dataset.

3.2.7 Prodigal & Pyrodigal

Prodigal (PROkaryotic DYnamic programming Gene-finding ALgorithm) is a gene prediction tool for prokaryotic genomic sequences. Prodigal works by identifying start and stop codons throughout the sequence, and then computing scores for each pair of codons using dynamic programming [HCL+10]. It is used frequently in bioinformatics pipelines that require gene-calling, and was found to be one of the best performing ORF finders [KWC+20].

Pyrodigal is a Python module that provides bindings to Prodigal [Lar22]. We used Pyrodigal (version 3.4.0) as part the of data preprocessing (see section 4.1.3).

3.2.8 Hidden Markov Models: Pfam, dbCAN, PyHMMER

Genomic features of interest were annotated with profile Hidden Markov Models (pHMMs). pHMMs are used to model protein sequence families and are seen as the standard method for annotating protein domains. Annotations were generated using the Python bindings of the HMMER package, PyHMMER (version 0.12.0) [LZ23]. Two sets of pHMMs were used: Pfam HMMs (version 37.1, [MCW+21]) for functional protein domain annotations and dbCAN HMMs (version 14) for CAZy annotations.

3.2.9 UMAP

UMAP (Uniform Manifold Approximation and Projection) [MHM20] was used as a dimensionality reduction technique to visualize gene embeddings generated by GeneCAT, ESMC and Bacformer, using the umap-learn Python package (version 0.5.12) and hyperparameters `distance=euclidean`, `n_neighbors=12`. Other settings were kept as defaults.

3.2.10 Machine Learning Models

GECCO was trained with default settings, except for the hyperparameter `select`, which was set to 0.5, to only select most informative 50% of features from the training data using a contingency test. The python implementation gecco-tool[train] was used (version 0.10.3) [CLF+21].

GeneCAT version 0.0.7 (unpublished) was finetuned with the hyperparameter `--middle-focus`, making the model only predict on the middle gene in the context-window during training. To ensure the training data is seen sufficiently by the model we set the maximum number of epochs to 30. In the zero-shot setting, we extracted gene embeddings from the models `model_gene_multilabel_pfam_cazy_april_goycr91w_v0` and `model_gene_multilabel_untied_april_sriqcx3c_v0` located on the LUMC Shark HPC cluster.

We used the Huggingface implementation of ESMC, called ESM++ large (600M parameter), to generate embeddings for the protein sequences encoded by each gene. (Huggingface commit hash `94b8ccf`, [CHD+26]). Specifically, protein embeddings were saved from the output of the Bacformer function `protein_seqs_to_bacformer_inputs`.

Bacformer Large (300M parameters) was used, with the initial (and at the time of writing also latest) release of the corresponding Python package (Huggingface commit hash `589fda4`). Since Bacformer was trained on sequences with a maximum of 6000 proteins, it resulted in issues when generating embeddings of longer genomes. To alleviate this, a sliding window approach was taken with a window size of 3000 and a stride of 2000, and embeddings of overlapping regions were averaged. Additionally, the workaround from the Bacformer GitHub issue 33 was used to properly initialize rotary embeddings buffers, preventing the model from outputting NaN values during inference.

A logistic regression classifier was utilized to classify embeddings in a zero-shot setting (for GeneCAT, ESMC and Bacformer). We used the implementation in the `LogisticRegression` class from the scikit-learn Python package (version 1.8.0) [PVG+11], with the liblinear solver and a maximum of 10000 iterations. All embeddings were normalized using L2-norms before training. The hyperparameters `l1_ratio` (either 0 or 1) and `C` (1, 10 or 500) were optimized using a 3-fold cross-validation gridsearch with AUPRC as the scoring metric.

3.3 Hardware

The majority of the work was performed on the SHARK High-Performance Computing cluster, provided by the LUMC. The cluster runs on Linux Rocky 8, with the Slurm workload manager (version 20.11).

For language model training and inference, compute nodes gpu01 and gpu02 were used, both running on the Nvidia GPU TITAN Xp (3840/12Gb).

4 Implementation

4.1 Dataset Construction

This section outlines all steps taken in the dataset construction process, including dataset curation, preprocessing, cryptic PUL identification, and train-test splits. A graphical overview is provided in Figure 2.

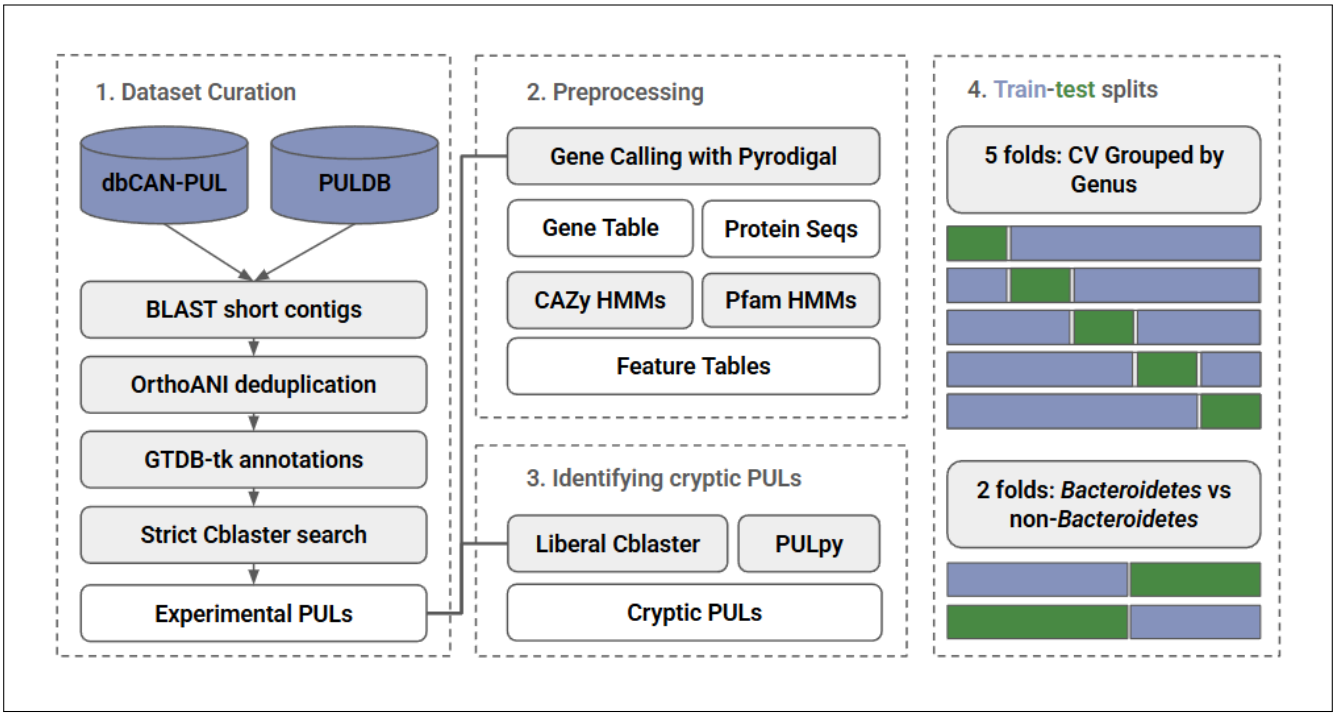


Figure 2: Graphical overview of steps taken during dataset construction. First, a dataset of literature-derived PULs was curated and then preprocessed for model training. Then, cryptic PULs were identified and the data was split in 5-fold cross-validation and two *Bacteroidetes*-based splits.

4.1.1 Dataset Curation

This project started with curating a robust and exhaustive dataset, containing known experimental PULs in their genomic context, across different bacterial taxa. All literature-derived PUL annotations were obtained from dbCAN-PUL and PULDB, and the corresponding sequences were downloaded from NCBI nucleotide (Section 3.1).

Due to inconsistencies in the databases, certain PULs had to be handled manually (see Supplementary Table 1). One genome with PUL annotations was downloaded separately from JGI. Another two PULs spanned over two contigs from a larger whole-genome shotgun sequencing project. Using BLAST on NCBI, we identified these PULs in the full genome of the specific organism. However, the PULs in both contigs, while initially appearing to be connected, mapped to separate subsequences in the full genome. These were then manually added to the dataset as two separate PULs. Three genomes from PULDB had to be removed, as we could not find their sequence identifiers in any publicly available database, and the data was not shared by the authors of the original papers.

Three other contigs from the database had unknown sequence identifiers, but were successfully found by consulting the relevant publication [CRU+18] After these steps, the annotations from dbCAN and PULDB were combined, and overlapping PULs sharing the same parent sequence were merged.

During data exploration it was observed that the dataset contained numerous extremely short contigs containing mainly or exclusively the associated PUL. We defined "truncated sequences" as sequences shorter than 100kbp, which encompassed 94 sequences and 98 corresponding PULs. Since we are evaluating models that rely on genomic context, a large number of very short sequences would create operational challenges and potentially biased results. To avoid removing these PULs, we used BLAST with the MegaBlast setting and default parameters to map all PULs from truncated sequences to full genomes or larger contigs. BLAST results were filtered to remove any self-hits and based on identity percentage (p_iden $\geq 99.5\%$) and the longest hit with the longest subject sequence was selected.

To further strengthen the quality of the dataset we performed an all-versus-all OrthoANI scan using PyOrthoANI to remove duplicate or overly similar genomes [LOKPC16, LZC25]. OrthoANI was selected over other ANI methods due to its faster run-time and robustness to gene rearrangements common in bacterial genomes. A cutoff of 99% identity was selected for deduplication and 98 genomes were deduplicated in the following way: for every identical (or highly-similar) pair, the longer sequence was selected and all PUL annotations from the shorter sequence were re-mapped using a local BLAST search with dc-megablast mode, a minimal percent_identity of 99% and a minimal query coverage of 90%. 19 PULs could not be re-mapped with these filters and were consequently dropped from the dataset.

Finally, to further improve data annotation given the patchy nature of experimental data, highly homologous PULs were identified across all contigs in the dataset using Cblaster [GBvW+21]. The search was done using the following strict filters: `-min_eval 1.0e-9 -min_identity 70 -min_coverage 75 --gap_size 5000 --min_hits 2 --unique 2`. All hits were filtered based on the number of genes from the query that were also in the hit, requiring a minimum of 70% of the genes to be present. With this search, 376 new homologous PULs were identified and added to the dataset.

4.1.2 Identification of Cryptic PULs

In the context of this project, we define the term "cryptic PULs" to refer to any gene clusters that possibly constitute a PUL, without having a literature-derived annotation. The presence of such cryptic PULs in our data presents a substantial challenge for this task, possibly causing several problems. In testing, genes in cryptic PUL regions will likely be classified as positives, and due to the lack of an annotation be thus evaluated as false positives, making the model seem worse than it might be. In training, conflicting annotations of similar PULs, where one is experimentally annotated and the other is cryptic, can mislead optimization attempts and thus also limit performance.

In order to reduce this uncertainty we first leveraged the only existing and openly available prediction tool for PULs, PULpy [SARW18b]. However, due to its rule-based nature, the tool is limited to only annotating *Bacteroidetes* PULs. To account for the lack of predictions in non-*Bacteroidetes*

species we used a homology-based approach, utilizing Cblaster with liberal filtering parameters [GBvW+21]. The default filtering parameters were kept for this search, only changing the following `--min_hits 2 --unique 2`. A threshold of 55% of genes being present in the hit was used, to enforce the two gene minimum, but also include any hits of two genes from longer query PULs. This method identified an additional 68 cryptic PULs.

4.1.3 Data Preprocessing

After curating a robust dataset of genomic sequences with literature-derived PUL annotations, the data was preprocessed in two main steps. First, gene calling was performed using Pyrodigal to generate standardized gene annotations for each contig and all protein-coding (CDS) genes were selected [Lar22]. This produced both a gene coordinate table and translated amino acid sequences for all predicted coding regions. Second, genomic features of interest were annotated using Hidden Markov Models (HMMs) through the PyHMMER package. All predicted proteins were searched against the Pfam and dbCAN databases using profile HMM alignment. Pfam contains curated protein domain families representing broadly conserved structural and functional motifs across proteins, while dbCAN provides HMM profiles specifically designed for the identification of carbohydrate-active enzymes (CAZymes) based on the CAZy classification system [MCW+21, ZGY+23]. Domain hits were filtered based on statistical significance thresholds ($p < 1e{-}9$). This resulted in two separate feature sets: one containing only Pfam domain annotations, and one combining Pfam and CAZyme annotations derived from dbCAN.

4.1.4 Train/Test Split

To minimize phylogenetic bias and prevent data leakage between training and test sets, a taxonomically stratified 5-fold cross-validation split was constructed. Contigs were first taxonomically annotated using GTDB-Tk version 226 [CMHP22, PCC+20], after which contigs were grouped at the genus level. All contigs originating from the same genus were assigned to the same fold, preventing closely related genomic contexts from appearing in both training and evaluation data. The data was stratified to preserve the relative distribution of positive and negative labels across folds.

Given the historical bias towards PULs in the *Bacteroidetes* phylum, we constructed two additional splits grouped by *Bacteroidetes* and non-*Bacteroidetes* phyla. The goal is to evaluate if our models are capable of generalizing across phyla, especially given the fact that the *Bacteroidetes* phylum is overrepresented in the dataset. Additionally, as existing PUL prediction methods work exclusively on *Bacteroidetes*, this evaluation would allow us to investigate whether learned representations remain informative outside this phylum. In the first setting, models were trained exclusively on sequences originating from the *Bacteroidetes* phylum and evaluated on non-*Bacteroidetes*, and the inverse configuration was used in the second setting.

4.2 Model Training and Evaluation

This section outlines all steps taken in training and evaluating the models for PUL prediction. A graphical overview is provided in Figure 3.

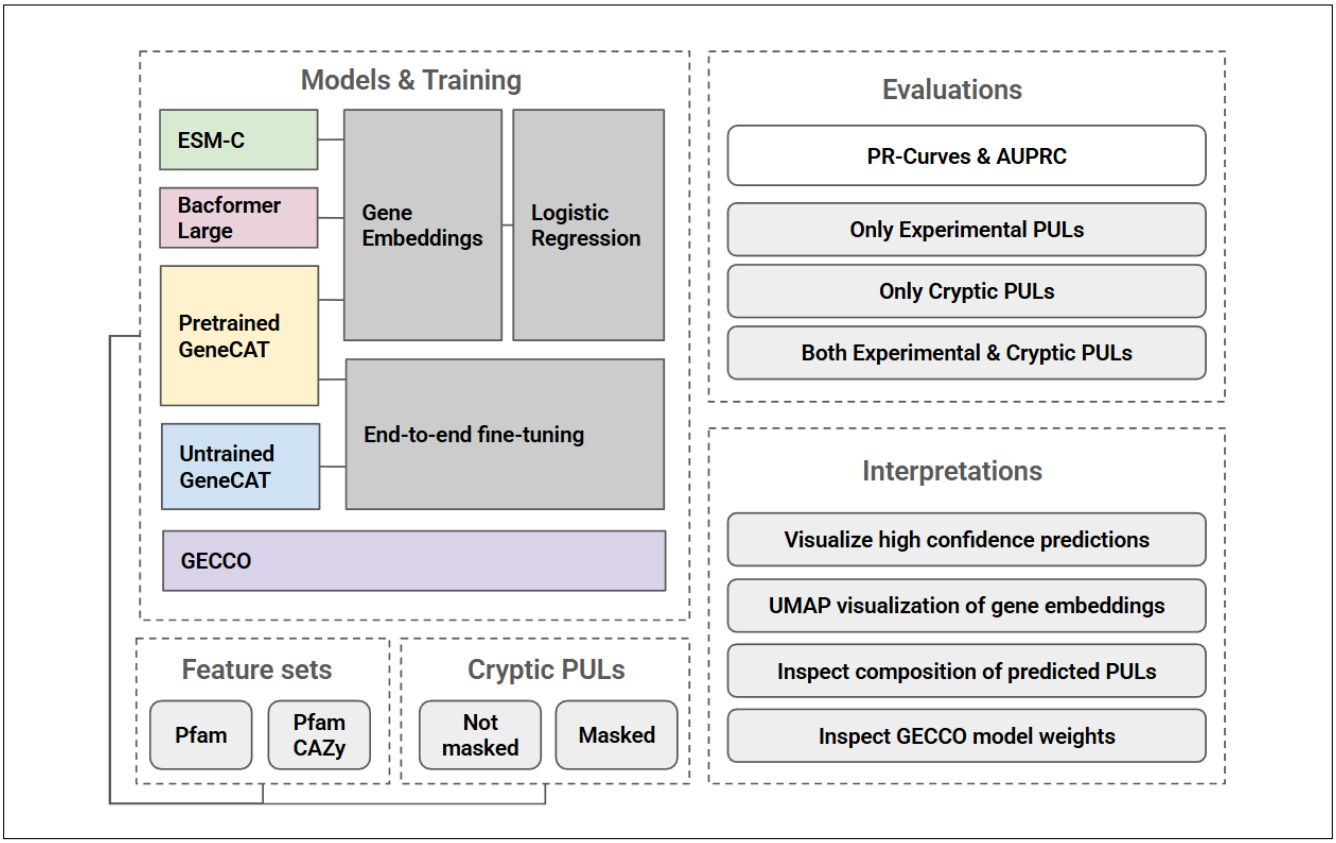


Figure 3: Overview of model training and evaluation process.

4.2.1 Model Implementations

In total, four different base models were compared: (1) GECCO, (2) GeneCAT, (3) the pLM ESM Cambrian and (4) the gLM Bacformer. GECCO was trained using the Python implementation with specific details discussed in section 3.2.10. ESMC and Bacformer were only used in a zero-shot setting, by extracting embeddings for every gene and using a logistic regression classifier to predict PUL genes based on these embeddings. GeneCAT was evaluated both in a zero-shot setting and with end-to-end finetuning. Additionally, we finetuned an untrained GeneCAT model with randomly initialized weights, to explore whether pretraining is beneficial for this task.

As both GECCO and GeneCAT rely on a sequence of genes with functional domain annotations, both were trained on either only Pfam annotations or a combination of Pfam and CAZy annotations. In order to include CAZy annotations in the vocabulary of GeneCAT, it was pretrained again with CAZy domains as part of the pretraining dataset. Moreover, for all models except for GECCO, we explored masking cryptic PULs in training, to ensure the models do not fit to a dataset where such regions are explicitly stated as negatives. In zero-shot settings, cryptic PUL genes were kept in the context window when extracting embeddings and later removed from the training data

for downstream logistic regression. For finetuning GeneCAT, the genes were kept in the context window, but were assigned the special label `UNKNOWN`, ensuring their logits and loss were masked during finetuning, so that no backpropagation occurs. In total, this results in a combination of 15 different models (see supplementary table 4 for an overview).

4.2.2 Evaluations

As stated above, all models were evaluated using 5-fold cross-validation and two additional *Bacteroidetes*-based splits. All genes present in the previously identified cryptic PULs that did not overlap with experimental annotations were excluded from these evaluations, to produce a more robust set of negative examples. The models were also separately evaluated on all exclusively cryptic PULs, excluding experimentally annotated PULs from the dataset, and on both cryptic PULs and experimental PULs.

Models were compared using Precision-Recall (PR) curves and Receiver operating characteristic (ROC) curves. A PR-curve shows how precision and recall values vary based on the choice of probability thresholds, calculated from binary labels at a given threshold using equations 1 and 2 respectively. Generally, for high cutoffs, precision is high and recall is low, and for low cutoffs precision is low and recall is high. A ROC-curve shows the true positive rate (TPR), which is the same as recall, against the false positive rate (FPR) at different thresholds. These metrics are calculated using equations 2 and 3 respectively.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (1) \qquad \mathrm{R} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (2) \qquad \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \quad (3)$$

The area under the PR-curve (AUPRC) and the area under the ROC-curve (AUROC) both serve as performance metrics that can be computed from raw model probabilities. The AUPRC can be approximated with discrete thresholds using the average precision (AP) score. This is calculated as a weighted average of the precisions at each threshold, using the recall increase from the previous threshold as weight (see equation 4). Due to the highly imbalanced nature of the dataset, AUPRC was selected as a performance metric over AUROC. This is because the FPR is directly biased by a large number of true negatives (TN). When evaluating models on imbalanced datasets, the ROC-curve, and by extension the AUC, can therefore be misleading with regards to classification performance [SR15].

$$\mathrm{AUPRC} = \int_0^1 \mathrm{P}(\mathrm{R}) d(\mathrm{R}) \approx \sum_n P_n \left(R_n - R_{n-1}\right) \quad (4)$$

In order to generate predictions of full PULs rather than only PUL-genes, an optimal probability threshold per model was selected using a held-out validation set of 10% during training, and calculating the threshold at which the maximal Matthews Correlation Coefficient is achieved. Then genes were clustered together using a simple heuristic: any directly adjacent genes with probability above the threshold were joined together to form a PUL, regardless of the intergenic distance. For GECCO, the built-in cluster predictions were used instead.

4.2.3 Interpretation

The results were further interpreted primarily through visualization, using the methods outlined in this section. To further explore the structure of learned representations, dimensionality reduction was performed on gene embeddings generated by GeneCAT, ESMC, and Bacformer. Uniform Manifold Approximation and Projection (UMAP) was used to project gene embeddings from every model (520, 960 and 960 dimensions, respectively) into a two-dimensional space for visualization and to observe any clustering patterns [MHM20]. In addition, the functional composition of predicted PULs was analyzed by computing the proportion of predicted PULs containing susC/susD homologs and carbohydrate-active enzymes (CAZymes). For GECCO, due to its inherent interpretability, feature importance was investigated by extracting model weights from the trained model state file. Finally, high-confidence predictions from GECCO and GeneCAT were visualized together with genomic annotations extracted from GenBank files and CAZyme annotations that were generated during preprocessing.

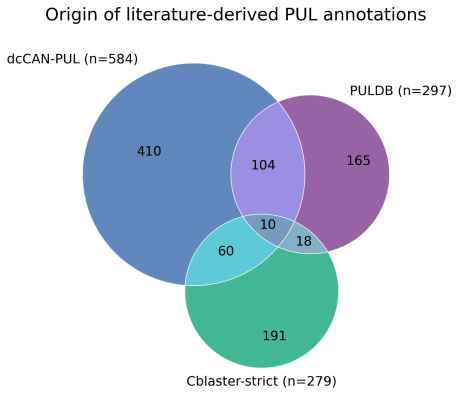
5 Results & Discussion

This section presents and discusses the results of this study in the context of the research questions defined in Section 1.2. We first examine the dataset and evaluate the impact of masking cryptic PULs during training. Subsequently, we investigate the effect of different feature representations and compare the performance of all evaluated models. We then assess the ability of these models to generalize beyond the *Bacteroidetes* phylum. Finally, we present several previously unannotated putative PULs identified by the models and discuss their biological relevance.

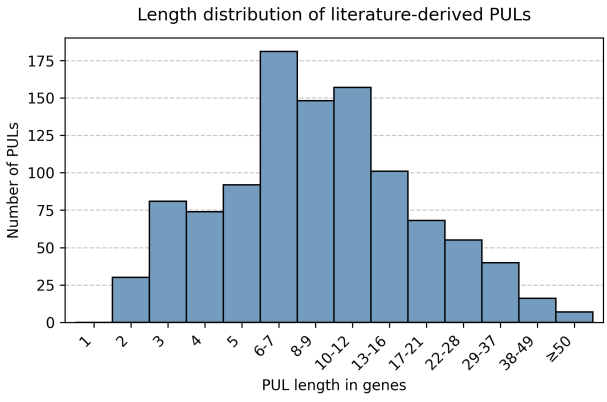
5.1 Data Exploration

The PUL annotation data for this project was obtained from the databases dbCAN-PUL and PULDB. To curate a high-quality dataset for this task, we first replaced truncated sequenced with longer contigs identified through a BLAST search, in order to ensure sequences are sufficiently long for methods incorporating genomic context. Subsequently, we deduplicated all contigs based on pairwise OrthoANI scores and then we used Cblaster to identify highly homologous PULs across genomes and added them to the annotations. After this procedure, the dataset contained 958 unique PULs in 288 contigs. Figure 4a shows the distribution of PUL origin, being either from dbCAN-PUL, PULDB, or from the strict Cblaster search. It is notable that many PULs that were found by Cblaster already had entries in the databases, but also a large number of new highly homologous PULs were found. The aforementioned historical bias in PUL research towards *Bacteroidetes* is also reflected in the dataset, with the *Bacteroidota* phylum (the GTDB equivalent of *Bacteroidetes*), accounting for nearly 69% of the dataset (see Supplementary Figure 3).

Figure 4b shows the length distribution of PULs, represented by the number of genes they contain. We note some very short PULs, consisting of only two or three genes, and very large PULs with more than 50 genes. The median PUL length is eight genes, and the mean is approximately eleven, meaning the data is skewed towards longer PULs. The longest PULs in the dataset belong to the gut



(a) Overlap in PUL origin, showing PULs from dbCAN-PUL, PULDB and Cblaster.



(b) Length distribution of all PULs in the dataset, quantified by the number of genes PULs span.

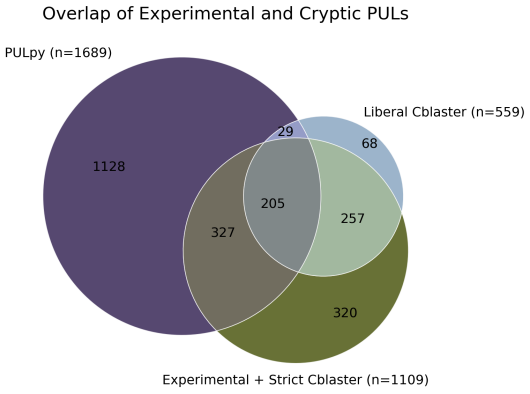
Figure 4: Overview of obtained dataset, showing PUL origin and PUL length distributions.

microbes *Bacteroides thetaiotaomicron* and *Bacteroides xylanisolvens* and to *Gilvimarinus chinensis*, a marine agar-hydrolyzing microbe [GTD+17, YTZ+22]. We further observe that literature-derived annotations are more permissive in terms of which genes are included in a PUL. PULDB often combines multiple co-localized PULs with the same substrate into a single literature-derived PUL. One example is Literature-Derived PUL 4 in *Bacteroides xylanisolvens*, where one PUL encompasses two susC-susD pairs located on separate strands with a distance of 2.000 base-pairs (see Supplementary Figure 2).

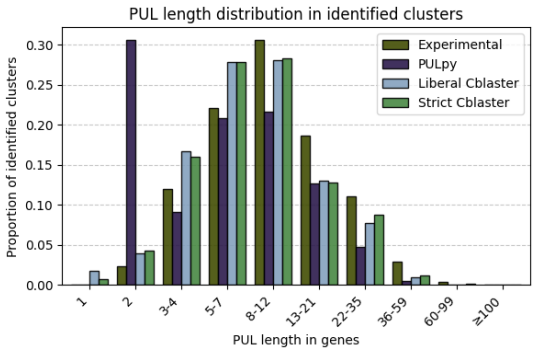
After preprocessing, a total of 532.674 protein-coding genes were called by Pyrodigal, with 11.691 annotated as belonging to PULs (2.19% of all genes). Due to this extreme imbalance, we use AUPRC as the main performance metric in the following sections.

5.2 Masking Cryptic PULs in Training

To address the issue of missing annotations and uncertain negatives, we identified a set of 1225 cryptic PULs by utilizing an existing rule-based and *Bacteroidetes*-specific prediction tool (PULpy) and a homology-based search using Cblaster with liberal parameters, referred to as liberal Cblaster. Most of the cryptic PULs were identified by PULpy, with a small fraction also identified by the liberal Cblaster search (see Figure 5a). Therefore, the overall set of cryptic PULs is predominantly from *Bacteroidetes* and follows the rules set in PULpy. Figure 5b shows the length distributions of PULs from experimental annotations, compared to distributions of PULs identified by PULpy and the liberal and strict Cblaster searches. It seems that PULpy predicts more short PULs, compared to the other methods. This is possibly due to PULpy enforcing a maximal intergenic distance of 500 base-pairs, while the distance between some genes in many of the literature-derived PULs can be much larger [SARW18b, AZY+21, TLD+18]. The PULs identified by the liberal and strict Cblaster searches show similar lengths distributions, with the liberal Cblaster predicting slightly shorter PULs, due to the difference in filtering parameters.



(a) Overlap between PULs identified by PULpy, a liberal Cblaster search and experimental PULs.



(b) Experimental and Cryptic PUL length distributions, quantified by number of genes.

Figure 5: Overview of experimental and cryptic PULs, showing PUL overlap, and length distributions.

To determine whether the presence of cryptic PULs in the dataset affects model performance, we trained our models on a dataset where cryptic PULs are masked, and compared them to models trained on the full dataset in which cryptic PUL genes are labeled as negatives. The models were evaluated on three sets of positives: one with only literature-derived (experimental) PULs, one with only cryptic PULs and one containing both. The negatives were kept constant across all evaluations, as genes belonging to neither experimental nor cryptic PULs. Model predictions from the 5-fold cross-validation were combined and evaluated as one test-set. For all models, masking cryptic PULs in training resulted in better performance on both experimental and cryptic PULs (see Figure 6). Even though the models were only trained on experimental PULs, masking cryptic PULs resulted in a greater performance increase when testing on cryptic PULs alone, compared to experimental PULs. This increase is most drastic for the end-to-end finetuned GeneCAT model, which without masking, in contrast to other models, performed substantially worse on cryptic PULs is than on experimental PULs.

An interesting aspect revealed by this evaluation is that cryptic PULs appear to be easier to predict than experimental PULs. One possible explanation relates to the fact that the majority of our cryptic PULs were predicted by the rule-based PULpy, in which the rules are inherently based on the patterns observed across *Bacteroidetes* PULs [TLD+18, SARW18a]. If cryptic PULs are not masked in training, instead receiving a negative label, then this becomes detrimental to any optimization method: regions with similar patterns would be in some cases labeled as positives and in others as negatives. Through masking, models likely learn the aforementioned patterns from experimental annotations alone, which explains the improved performance. Finally, the performance on a combined set of cryptic and experimental PULs is consistently the highest, because we keep the set of negative examples constant across all evaluations, while the total number of true-positives increases, inherently improving precision.

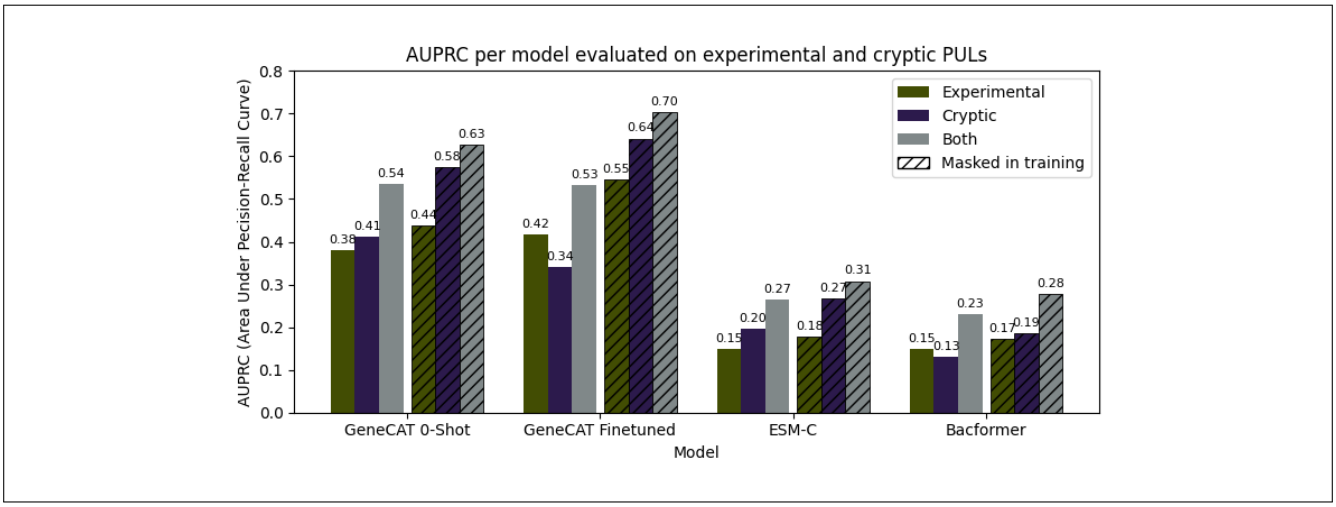
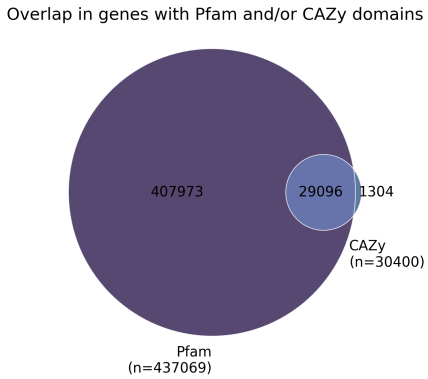


Figure 6: AUPRC scores per model. For each model, solid bars (left) show training without masking cryptic PULs and hatched bars (right) show training with masking. Bar color indicates the test set used for evaluation: experimental PULs (green), cryptic PULs (purple), or both (gray).

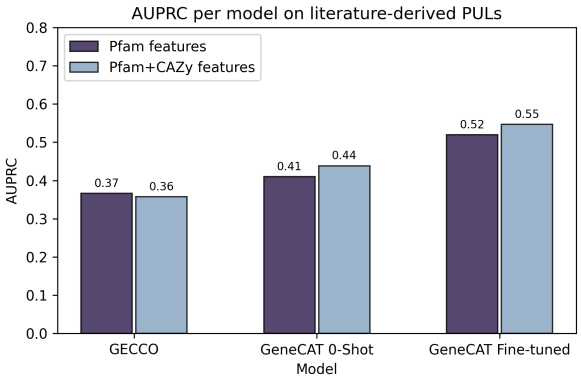
5.3 Comparing Feature Representations

Both GECCO and GeneCAT train on similar feature representations: genes with associated Pfam protein domain annotations. However, since PULs are enriched with CAZymes, we explored whether the addition of CAZy annotations to the feature representation can improve model performance. Initially, we determined how much information is added to the feature space by including CAZy annotations. Figure 7a shows the number of genes in the dataset annotated by either Pfam, CAZy or both. Most genes with CAZy annotations also have some Pfam annotation, but there is a small subset of 1304 genes that only received CAZy annotations. Out of these genes, we find 199 (15.3%) in literature-derived PULs, a relatively large portion considering that only 2.19% of all genes in the dataset belong to PULs. These genes contain 212 unique CAZyme families with various subfamilies of the glycoside hyrolases GH5, GH43 and GH16 appearing most frequently.

Adding CAZy annotations indeed has a positive, albeit small effect on model performance for GeneCAT, resulting in a slight increase of AUPRC (0.03). For GECCO on the other hand, adding CAZy annotations to the features led to a very slight decrease in performance (see Figure 7b). Given the fact that an overwhelming majority of CAZy domains are already captured by Pfam annotations, the minimal effect of including them in the dataset is fairly reasonable. GeneCAT's improved performance may be a result of pretraining on a vocabulary that incorporates CAZy annotations, improving its representation of genes related to carbohydrate metabolism. However, this slight benefit may come at the expense of performance on other unrelated tasks, a trade-off we did not investigate in this study. For subsequent comparisons, we use the models trained with the feature set that resulted in the best performance: GECCO with Pfam features and GeneCAT with both Pfam and CAZy features.



(a) Overlap in genes annotated with Pfam and/or CAZy domains.

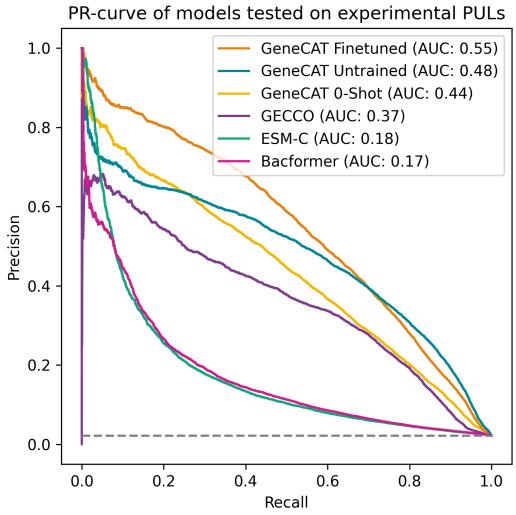


(b) AUPRC-scores of models trained on only Pfam domains or Pfam+CAZy domains.

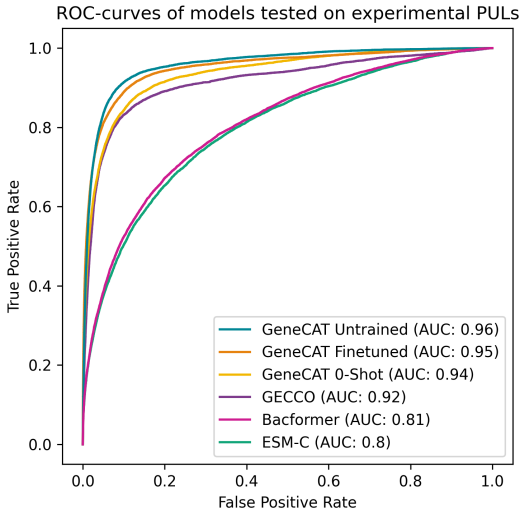
Figure 7: Feature representation comparison, showing (a) overlap between Pfam and CAZy annotations and (b) differences in performance for models trained on both representations.

5.4 The Promise of Foundation Models

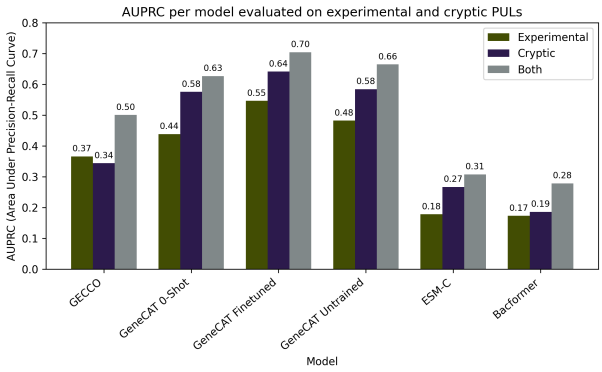
As foundation models become more common in computational biology, it is important to evaluate whether added model complexity and extensive pretraining are beneficial for a given task. In this study, we compared a baseline classical-ML model (GECCO, based on conditional random fields) with the deep learning-based foundation models GeneCAT, ESMC and Bacformer. All foundation models were evaluated in a zero-shot setting, using logistic regression to classify generated gene-embeddings. We also performed end-to-end finetuning on GeneCAT, both with a pretrained and an untrained model, to study the potential benefits gained by pretraining. Cryptic PULs were masked in training for all models except for GECCO. Overall, the pretrained-finetuned GeneCAT model performed the best, achieving the highest AUPRC-score on experimental PULs, followed



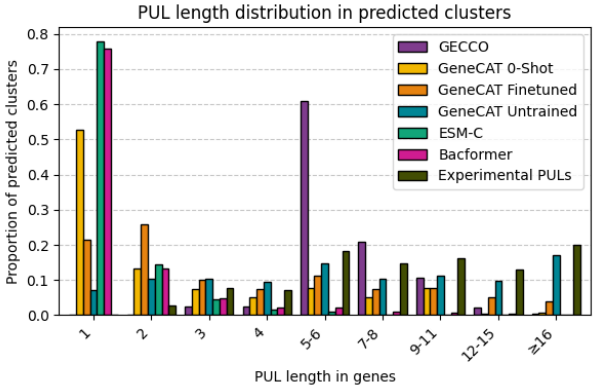
(a) PR-curves of all models, tested on experimental PULs.



(b) ROC-curves of all models, tested on experimental PULs.



(c) AUPRC-scores of models, evaluated on experimental PULs, cryptic PULs or both.



(d) Lengths of predicted PULs per models, quantified in gene count.

Figure 8: Global comparison of all tested models, showing PR-curves, ROC-curves, performance on experimental and cryptic PULs and predicted PUL length distribution.

up by the untrained-finetuned GeneCAT (AUPRC 0.55 and 0.48 respectively, see Figure 8a). Interestingly, the untrained-finetuned GeneCAT shows slightly better precision at high recall values. In accordance with this observation, the untrained-finetuned GeneCAT model also shows a slightly better ROC-curve and AUROC score, meaning it has a better TPR (equivalently recall) at a low FPR (see Figure 8b). Nevertheless, AUROC-scores should be interpreted with caution within this context, due to the extreme class imbalance. Moreover, unlike other models, GECCO shows a dip in precision at very low recall values, indicating that the model predicts many false-positives at very high probability thresholds. These may constitute unannotated putative PULs predicted with high confidence, which is further explored in section 5.6. ESMC and Bacformer both performed the worst on this task, while being the largest of the evaluated models, by number of parameters (600M and 300M, respectively).

The evaluations with cryptic PULs from section 5.2 were repeated on all models (see Figure 8c). We note that GECCO performs slightly worse on cryptic PULs, while all other models show a moderate if not substantial performance boost, which is likely caused by our treatment of cryptic puls as negatives, since GECCO does not allow for masking. Another interesting point is that the performance of the untrained-finetuned GeneCAT and the zero-shot GeneCAT models on cryptic PULs is identical, while the pretrained-finetuned GeneCAT outperforms both. Since all evaluations were done on a gene level, we visualized the lengths of predicted PULs in Figure 8d. Many predicted PULs, especially the ones predicted by ESMC and Bacformer, were quite fragmented, resulting in many single-gene PULs. GECCO mostly predicts PULs that are 5-6 genes long and does not predict any clusters with less than three genes, due to the specific built-in way the model defines clusters. Interestingly, the untrained-finetuned GeneCAT model predicts PULs of lengths that most closely resemble the length distribution of experimental PULs, which are typically longer (also see Figure 4b).

To better understand the benefits of pretraining, we visualized training loss and validation metrics during finetuning for the pretrained and untrained GeneCAT models on the first cross-validation fold (see Figure 9). The pretrained model converged in fewer epochs and reached high validation performance more rapidly than the untrained model. Training loss exhibited fluctuations for both models, but these were more pronounced for the untrained model. The improved training stability and faster convergence indicate that pretraining is advantageous for this task. Nevertheless, as noted above, the untrained-finetuned GeneCAT achieves competitive performance on both experimental and cryptic PULs, suggesting that the combination of functional domain feature representations and a Transformer architecture is well suited to this task even without extensive pretraining.

In order to further explore the differences between foundation models, two-dimensional UMAP projections were calculated and visualized for each set of gene embeddings, shown in Figure 10. For all GeneCAT models (Figures 10a, 10b, 10c) there are visible patterns of experimental and cryptic PUL genes clustering to some degree. As expected, this is most pronounced with the pretrained and finetuned model. Interestingly, genes from cryptic PULs cluster more closely than genes from experimental PULs. This may be attributed to the majority of cryptic PULs being predicted by PULpy, which relies on susC-susD homologs and CAZymes. The UMAPs for ESMC and Bacformer embeddings (Figures 10d & 10e) show more clusters overall, with more distinct islands for Bacformer. However, PUL genes are not clustered in any meaningful way, which is further corroborated by the subpar performance of these models (see Figure 8).

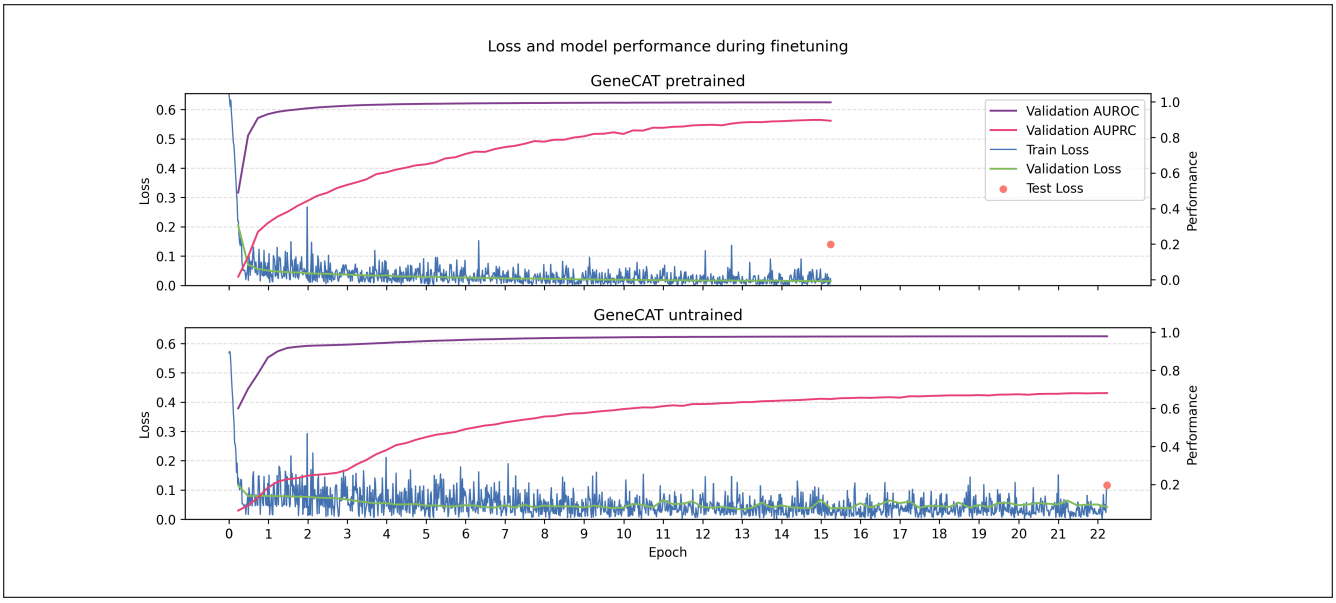
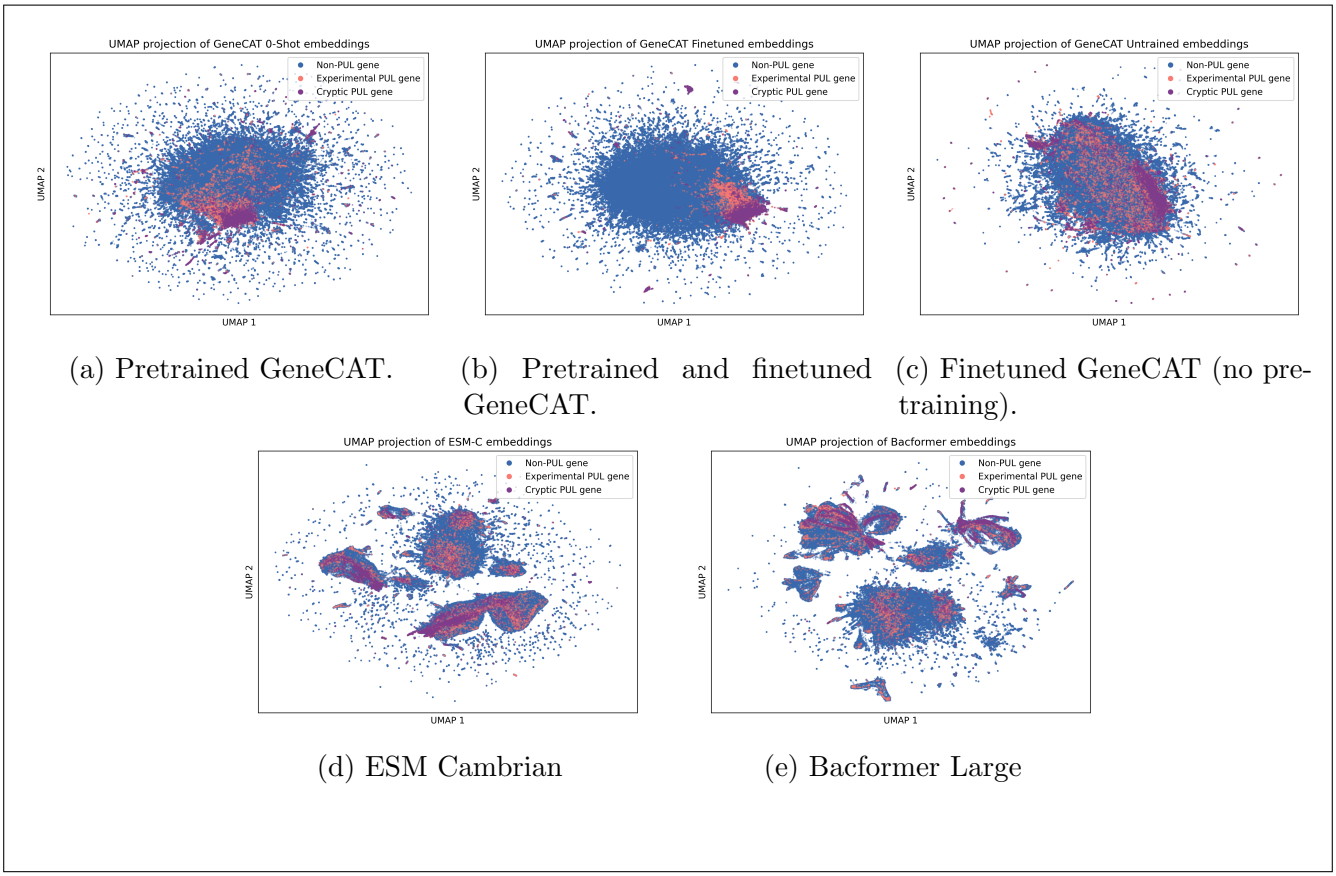


Figure 9: Train and validation loss and other performance metrics (AUPRC, AUROC) during finetuning the pretrained and untrained GeneCAT models, plotted over training epochs.



(a) Pretrained GeneCAT.

(b) Pretrained and finetuned GeneCAT.

(c) Finetuned GeneCAT (no pretraining).

(d) ESM Cambrian

(e) Bacformer Large

Figure 10: UMAP projections of gene embeddings generated by GeneCAT, ESMC and Bacformer, highlighting genes in literature-derived PULs (pink) and genes in cryptic PULs (purple).

One possible explanation for the relatively poor performance of ESMC and Bacformer is a mismatch between their pretraining objectives and this task. PUL prediction relies on the identification of specific functional structures. Investigations of the ESMC latent space by Candido et al. reveal that information on protein function is best embedded in the middle layers of the model, whereas we (and Bacformer) make use of the embeddings from the final layer [CHD+26]. As a result, information relevant for distinguishing PUL genes may not be sufficiently emphasized in the pretrained ESMC embeddings to allow for zero-shot classification. Similarly, although Bacformer incorporates genomic context, its embeddings are optimized for general genome-level structures rather than specifically for identifying functional elements. As evident from the evaluations, these representations may not be directly suitable for zero-shot PUL prediction. We hypothesized that these models might be better at capturing more general taxonomic signatures, which we confirmed by re-coloring the UMAP projections based on phylum. This revealed a better separation between phyla for ESMC, and some distinct clusters for Bacformer (see Supplementary Figure 4). In contrast, Pfam and CAZy domains convey functional information that is directly relevant to carbohydrate utilization, from which GeneCAT inherently learns. While the UMAP projections of this model's embeddings do not show explicitly distinct clusters, genes associated with PULs tend to be closer to each other, even in a zero-shot setting (Figure 10a).

5.5 Generalizing Beyond *Bacteroidetes*

Model generalization is an important aspect of ML, especially when data tends to be biased in one way or another. Due to the historic experimental focus on *Bacteroidetes* in PUL research, PULs that were characterized in *Bacteroidetes* are overrepresented in dbCAN-PUL and PULDB (see Supplementary Figure 3a). For this reason, we explored how well models generalize beyond

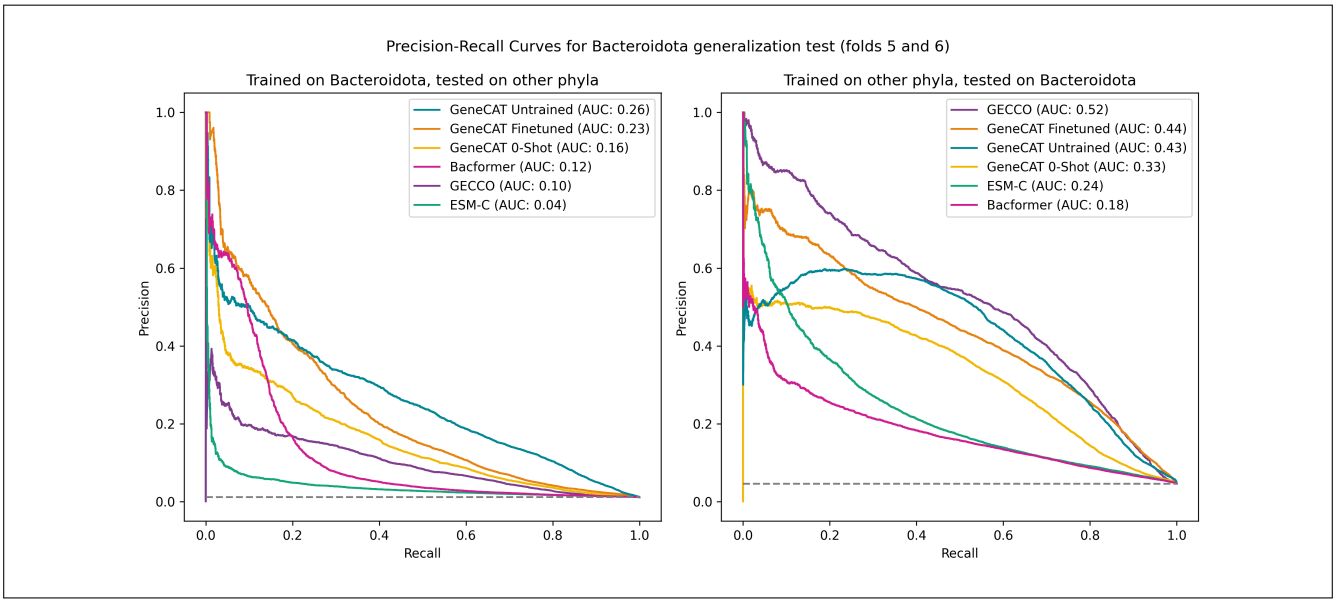


Figure 11: Precision-Recall curves for all models for two specific train-test splits, in order to assess model generalization across phyla. Models were either trained on only *Bacteroidetes* and tested on all other phyla (left), or vice-versa (right).

this phylum. As mentioned in section 4, aside from the ordinary 5-fold cross-validation splits, two additional *Bacteroidetes* vs non-*Bacteroidetes* train/test splits were created and models were evaluated on experimental PULs using PR-curves and the associated AUPRC (see Figure 11). Models trained on only *Bacteroidetes* PULs were overall less capable of generalizing to non-*Bacteroidetes*, with the untrained-finetuned GeneCAT model achieving the best performance (AUPRC 0.26). In contrast, models trained on PULs from non-*Bacteroidetes* could generalize to *Bacteroidetes* fairly well, with GECCO achieving better performance compared to the 5-fold cross-validation (AUPRC 0.52 vs 0.37). One possible explanation for this asymmetric generalization is that PULs in *Bacteroidetes* have a much more conserved structure than PULs in other phyla, particularly with respect to the canonical susC/susD pair. On the other hand, Gram-positive PULs (gpPULs) require no mechanism to overcome the double membrane bilayer of Gram-negative species and therefore lack susC homologs, instead using a variety of different transport mechanisms [BGB21]. As a result, models trained exclusively on *Bacteroidetes* may learn to rely on a relatively narrow set of highly predictive features that do not generalize well to other phyla, while models trained on phylogenically more diverse non-*Bacteroidetes* data are also exposed to a more diverse set of PULs and may therefore learn broader patterns that generalize better to *Bacteroidetes*.

Due to the nature of GECCO, it offers a clear and easy method to inspect the weights assigned to each feature. To support the explanation above, we inspected the weights the model assigned to susC- and susD-associated Pfam domains, averaged across the 5-fold cross-validation and for the two separate *Bacteroidetes* splits (see Figure 12 and Supplementary Table 3 for the specific Pfam domains). We see that GECCO assigns higher weights to susD-associated features, compared to susC-associated features, with a more pronounced difference when the model is trained on non-*Bacteroidetes* PULs.

To further explore the above hypothesis, we looked at the number of predicted PULs in *Bacteroidetes* that contained susC/susD homologs and CAZy domains, for predictions made by either models in the 5-fold cross-validation and or models trained on non-*Bacteroidetes* PULs (see Figure 13). A discernible trend is that models predict less PULs with susC- or susD-homologs when learning

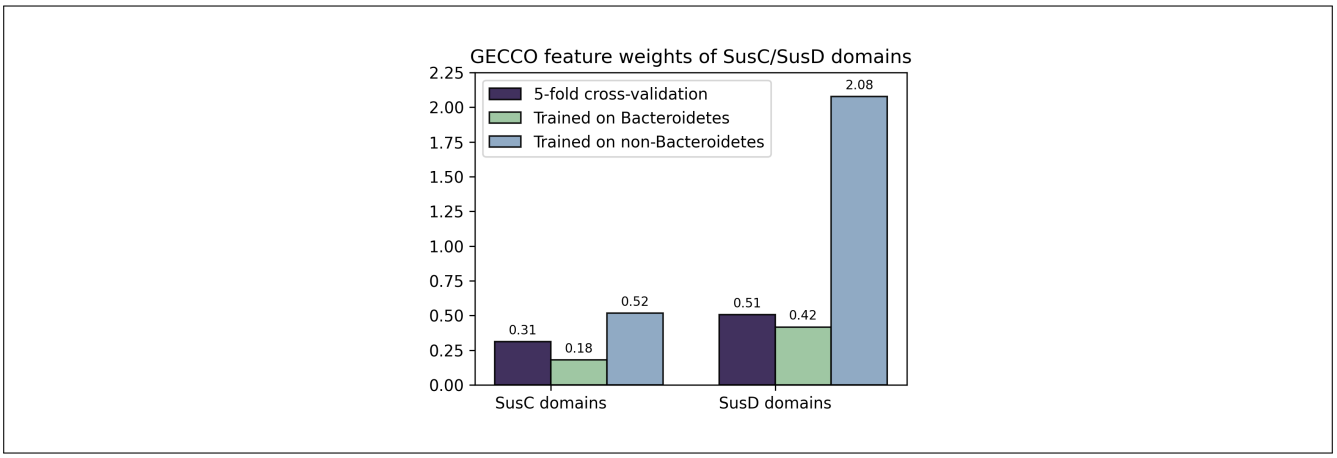


Figure 12: Weights assigned by GECCO to susC- and susD-associated Pfam domains when trained in 5-fold cross-validation, compared to when trained on only *Bacteroidetes* or non-*Bacteroidetes* PULs.

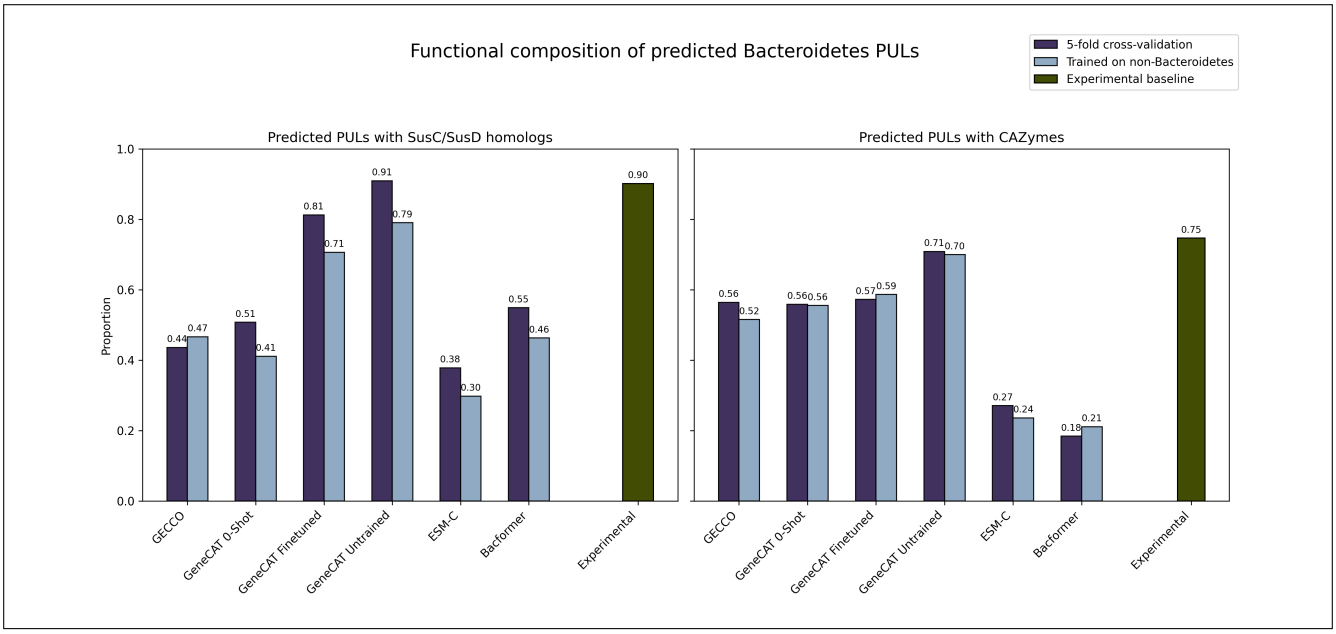


Figure 13: Proportions of predicted PULs containing susC/susD homologs and predicted PUL genes with CAZy domains, for predictions made by models trained in the 5-fold cross-validation and models trained on non-*Bacteroidetes*.

only from non-*Bacteroidetes* PULs, which is logical given the absence of this specific structure in other phyla. Still, a substantial number of such PULs are predicted, which could be explained by these genes co-localizing with CAZymes, or by the fact that susD homologs are often present in gpPULs. GECCO even predicts slightly more PULs with susC/susD homologs, likely due to the aforementioned increased weight assigned to susD domains (see Figure 12). In contrast, the number of predicted PUL genes containing CAZy domains remains more consistent across models, indicating that models learn to predict (partially) based on CAZymes regardless of phylum.

5.6 Newly Predicted Putative PULs

One advantage of supervised learning is the potential to guide the discovery of new PULs, particularly in bacteria where existing rule-based methods cannot be used, such as bacteria of the Gram-positive *Firmicutes* phylum (also called Bacillota). This phylum is one of the most dominant of the human gut microbiota and its members are critical in metabolizing various polysaccharides in the gut [WLBF14, KAG+13]. To further demonstrate this point, we investigated putative novel PULs that were predicted in Firmicutes genomes with high confidence. In this section we discuss hand-picked examples of such putative PULs from *Roseburia* and *Blautia* species commonly found in the human gut microbiome, a *Clostridium* species used for biotechnological applications. We also include one predicted *Bacteroides* PUL that PULpy failed to identify.

Figure 14 shows a region predicted as a PUL by both GECCO and GeneCAT in the *Roseburia intestinalis* L1-82 genome, an abundant human gut microbe [LRLM+19]. Previous research on the *R. intestinalis* L1-82 strain experimentally identified two gpPULs that are involved in metabolizing β-mannans, a type of hemicelluloses from plant cell walls, common in human diets [LRLM+19]. In

a different *R. intestinalis* strain (XB6B4), Sheridan et al. identified 33 gpPULs through manual curation, and two PULs in the L1-82 strain, with homology to these XB6B4 PULs, are explicitly discussed in the context of arabinoxylan, xylan and pectin utilization [OSML+16]. A more recent paper on *R. intestinalis* presents numerous manually curated gpPULs in the genome of the L1-82 strain, but provides no experimental verification [MMS+25]. To our knowledge, the PUL predicted by our models (Figure 14) was not among the gpPULs annotated by the aforementioned papers. Nevertheless, we are confident in this prediction, given the co-localization of CAZymes, transporter-encoding genes and a transcriptional regulator. The PUL contains ten CAZymes, out of which nine with GH domains and one CE, and an AraC family transcriptional regulator, a known binder of L-arabinose in E-coli [Sch00]. Further upstream, there are two ABC-transporter-encoding genes, which the models exclude from the PUL. Nevertheless, given the co-localization with CAZymes and gpPULs often containing ABC-transporters, it is likely that these genes are part of the putative PUL [BGB21]. Based on Cayman annotations, the potential substrates for the CAZymes in this PUL are gums, pectins, glycoproteins and hemicelluloses [DKG+26].

An additional putative PUL was predicted by GECCO in the *Blautia parvula* genome (see Figure 15). Bacteria of the *Blautia* genus are commonly found in mammalian gut microbiomes and are highlighted for potential probiotic functions [LMG+21]. For the species *B. parvula* (NBRC 113351 strain), only a single PUL-like cluster had been experimentally characterized and recorded in dbCAN-PUL, implicated in degrading DFA-III, a cyclic fructo-disaccharide [YHK+24]. The PUL predicted by GECCO contains three CAZymes of the GH2_13 family, which encode β-mannosidases. The CAZymes are preceded by three genes related to ABC-transporters, further supporting this prediction. While one of the transporter-related genes is excluded from the predicted PUL, the location and function of this gene provide a strong indication that it is part of the cluster. The predicted cluster also contains an aldo/keto reductase, possibly involved in catalyzing a reaction

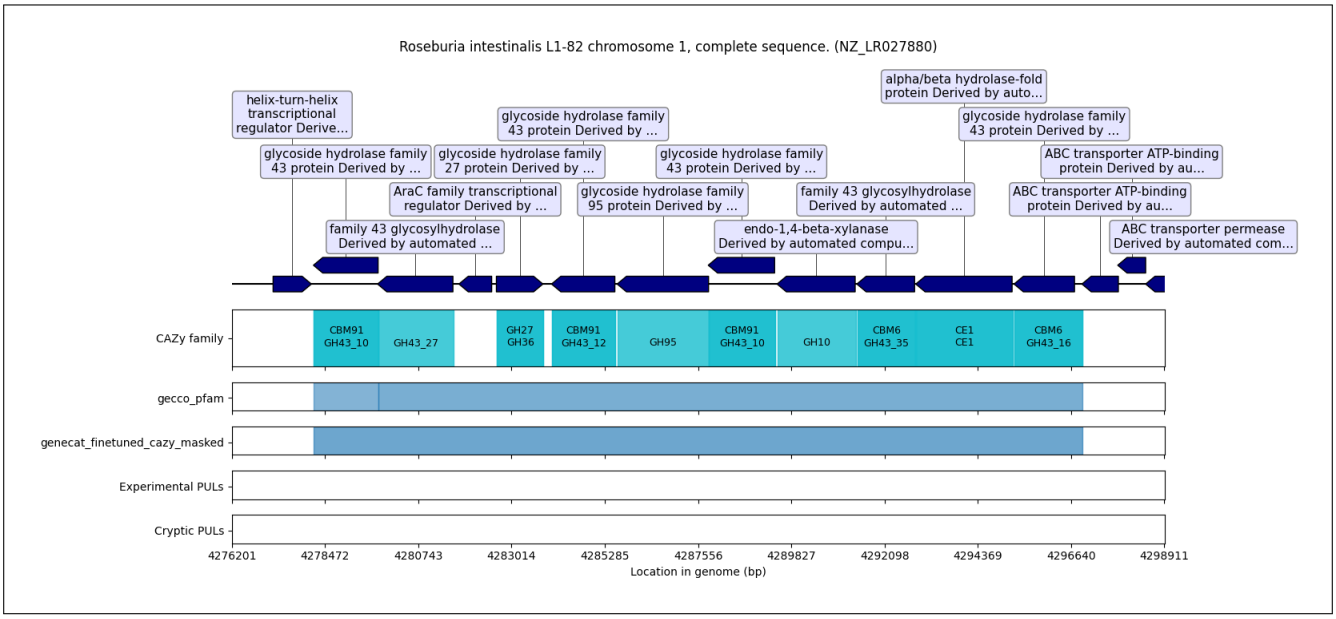


Figure 14: Putative PUL predicted by GECCO and GeneCAT in the *Roseburia intestinalis* genome, without known experimental annotations.

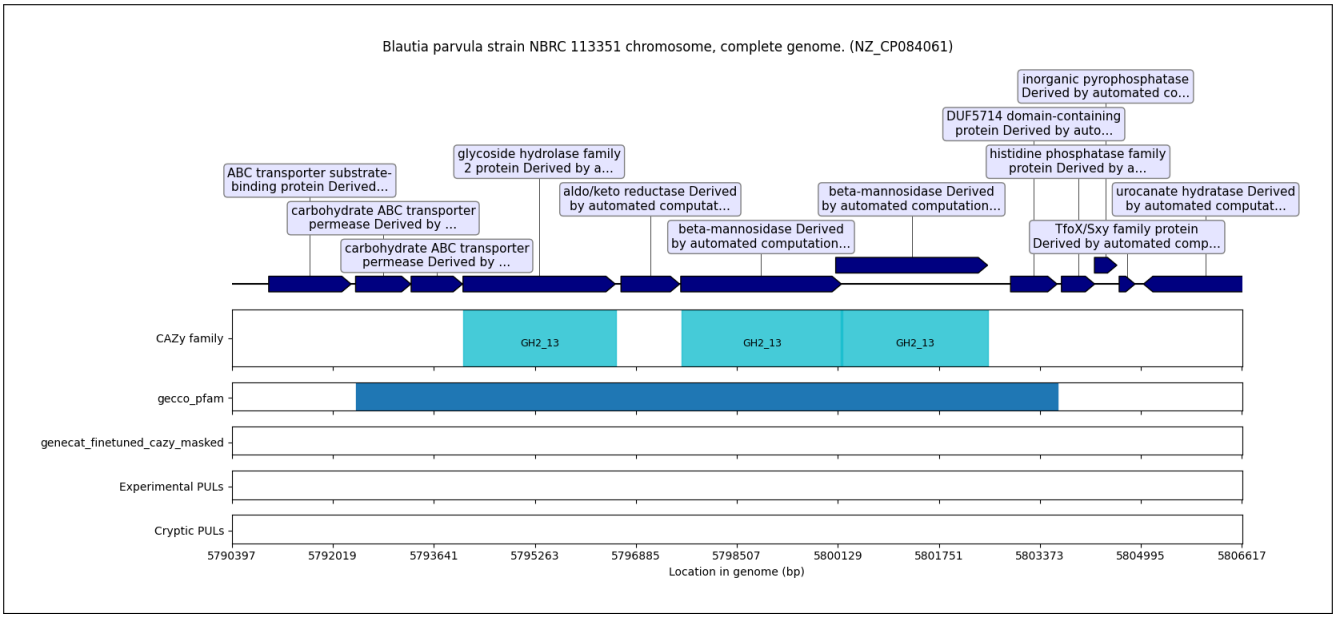


Figure 15: Putative PUL predicted by GECCO in the *Blautia parvula* genome, without known experimental annotations.

with one of the downstream products. GECCO also includes one gene encoding for a protein of unknown function as part of the prediction, which is likely not part of the cluster, as the intergenic distance from the last CAZyme (∼400bp) is larger than the distance with the preceding CAZymes (∼20bp).

Both GECCO and GeneCAT predicted a putative PUL in the *Clostridium bijerinckii* genome, a bacterial species with bio-industrial applications. *C. bijerinckii* has been historically used for producing acetone and butanol in a process called acetone-butanol-ethanol (ABE) fermentation and is being explored for biofuel production, particularly bio-butanol from biomass [RRL99, LYL+10]. To our knowledge, the only experimentally verified PUL in this species was found to be responsible for sucrose utilization, consisting of a phosphotransferase system (PTS) transporter, a transcriptional repressor, an ATP-dependent fructokinase and a sucrose hydrolase [RRL99]. The predicted PUL is composed of six CAZymes, a sigma-I transcription factor, a BgIG family transcription antiterminator (regulatory protein) and a PTS transporter (see Figure 16). The prediction also encompasses an incomplete transposase and a hypothetical protein, both very short and likely irrelevant to the cluster's function. The presence of a PTS transporter in the experimentally verified sucrose utilization operon is also in agreement with this predicted PUL.

In some cases, our models picked up potential PULs in *Bacteroidetes* that were not predicted by PULpy: a total of 150 by the finetuned GeneCAT model and 67 by GECCO, all with varying confidence. We picked an example of a putative PUL from *Bacteroides fragilis* YCH46, predicted by both GECCO and GeneCAT (see Figure 17). The predicted PUL contains two putative sulfatases in addition to susC/susD homologs and multiple CAZymes. Sulfatases in gut microbes are commonly associated with mucin-degradation, since mucins often contain heavily sulfated oligosaccharide side chains. It is therefore plausible that this cluster encodes for a mucin utilization locus (MUL), a term coined in 2023 to describe co-localized genes involved in the use of mucin [DMV+23]. Further

supporting this hypothesis, the three identified CAZy domains (GH36, CBM32 ad GH29) are all involved in mucin degradation, based on Cayman annotations [DKG+26]. Moreover, a similar PUL was predicted in another *B. fragilis* strain (NCTC 9343; see Supplementary Figure 5).

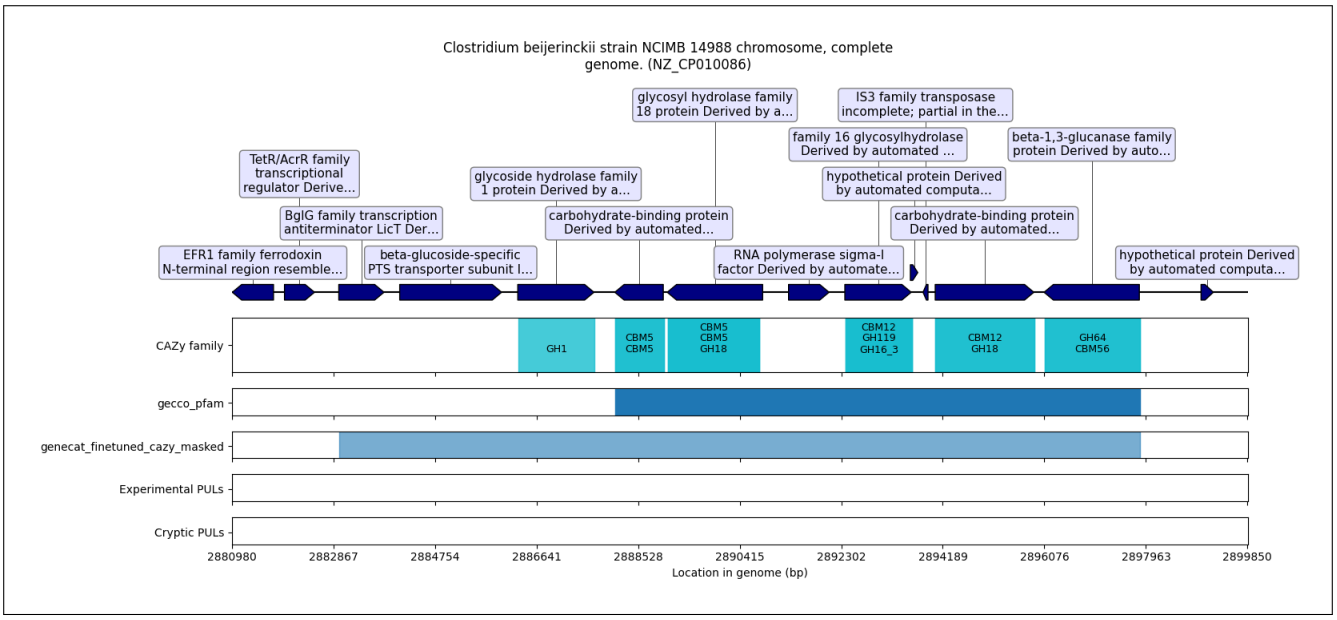


Figure 16: Putative PUL predicted by GECCO and GeneCAT in the *Clostridium bijenrinckii* genome, without known experimental annotations.

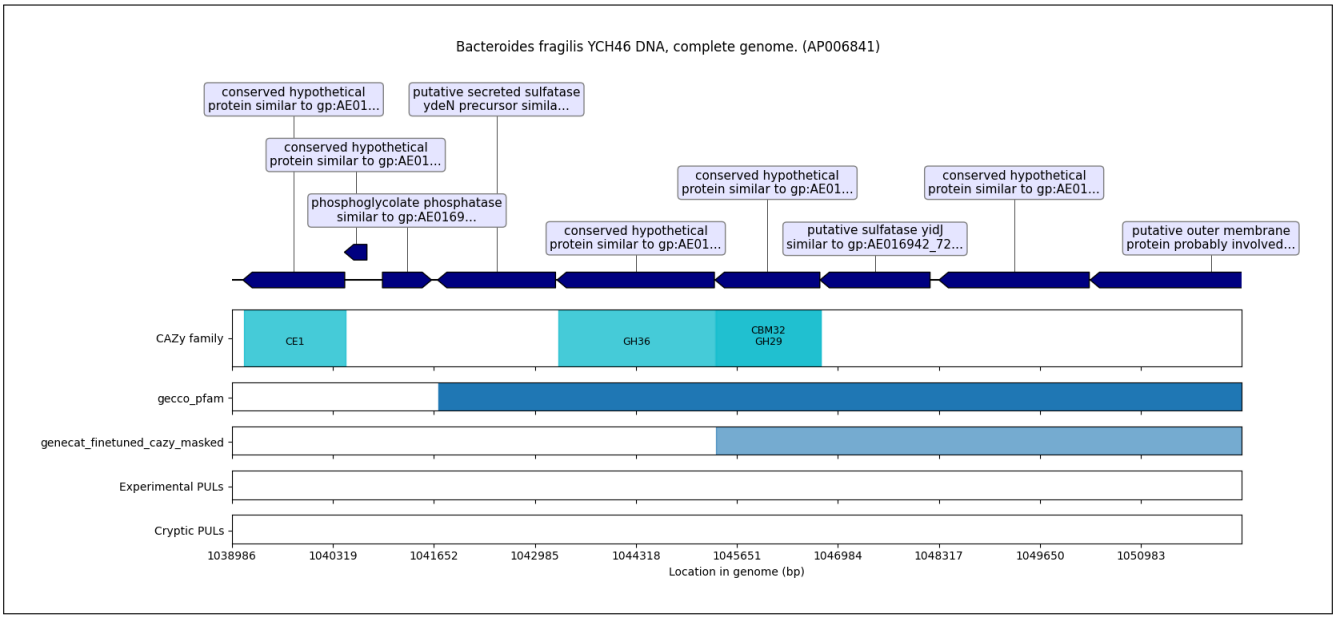


Figure 17: Putative PUL predicted by GECCO and GeneCAT in the *Bacteroides fragilis* genome, without known experimental annotations and not predicted by PULpy.

5.7 Software Implementation

The full data curation, preprocessing, and modeling pipeline was implemented as a series of Python scripts with additonal shell scripts managing job submission on the LUMC SHARK HPC cluster. Following the procedures described in Section 4 these scripts handle data collection, contig extension, deduplication, Cblaster-based annotation, GTDB-tk taxonomic annotation, gene/feature annotation, and the train-test splits. Model training is handled separately for each model using the same cross-validation cluster tables, with scripts for training GECCO, finetuning GeneCAT, extracting embeddings from GeneCAT, ESMC and Bacformer, and subsequently training logistic regression models on the embeddings. Evaluation and visualization were wrapped in a shared interface across all models.

The full pipeline, including a README describing environment set-up and the intended execution order, is stored in a private GitHub repository. Access can be granted upon request.

6 Conclusions & Outlook

6.1 Conclusions

During this research, the main goal was to explore whether supervised machine learning techniques can be effectively leveraged for the prediction of polysaccharide utilization loci. This section aims to answer the research questions formulated in section 1, starting with the sub-questions RQ1-4 and concluding with the main research question.

RQ1: How can a binary classification dataset be constructed despite the incomplete nature of experimental PUL annotations?

To handle the absence of confident negatives in our data, we introduced the term "cryptic PULs" to refer to regions that might constitute PULs but lack experimental validation. We identified cryptic PULs by using the existing prediction method PULpy and an additional homology-based search. Cryptic PULs were then masked in training and handled separately in evaluation, instead of being naively labeled as negatives. Our results show that this is an effective strategy to improve model performance on experimental PULs. For the pretrained and finetuned GeneCAT model, the AUPRC increased substantially, from 0.42 to 0.55. Furthermore, performance on predicting cryptic PULs increased even more, from 0.34 to 0.64, showing that even without directly learning from cryptic PULs, models are capable of learning the patterns based on which the rules were set for PULpy.

RQ2: Does the inclusion of CAZy annotations as features improve performance for models which utilize domain annotations?

Initial exploration showed that while CAZy and Pfam annotations mainly overlap, there are 1304 genes with only CAZy annotations, with 15.3% of these genes being in PULs. For GeneCAT in both the zero-shot setting and finetuning, including CAZy domains in the feature representation slightly boosted performance, from 0.41 to 0.44 and 0.52 to 0.55, respectively. GECCO showed a small decrease in performance, from 0.37 to 0.36. This indicates that while Pfam annotations are mostly sufficiently informative on their own, the addition of CAZy annotations can slightly improve model performance.

RQ3: How do models with different architectures and genome representations compare in the context of PUL prediction?

Models that learn from genomes represented as a sequence of functional domain annotations (GECCO and GeneCAT) outperformed models that learn from other genome representations on this task (ESMC and Bacformer), with the Transformer-based GeneCAT models outperforming the classical ML CRF-based GECCO. ESMC and Bacformer, which respectively represent genomes as protein sequences and pLM embeddings, were less successful in predicting PULs in a zero-shot setting (AUPRC 0.18, 0.17). The information captured by these representations is possibly too broad to be useful for classifying PULs. Two-dimensional UMAP projections of gene embeddings generated by Bacformer corroborate this by showing discernible patterns and clusters, but none that correspond to experimental or cryptic PULs. Finally, the effect of pretraining is not as substantial as expected when finetuning GeneCAT: the pretrained model showed an AUPRC of 0.55 on experimental PULs, compared to 0.48 by the untrained model. However, the finetuned untrained

GeneCAT only marginally outperforms the zero-shot GeneCAT model (AUPRC 0.48 vs. 0.44), despite being more computationally intensive to train (see Supplementary Table 4).

RQ4: Given that experimental work is mostly focused on *Bacteroidetes*, can models generalize to other phyla?

We evaluated the models' generalization capabilities by training and testing on one additional data partition, where models were trained on only *Bacteroidetes* PULs and tested on non-*Bacteroidetes*, and vice-versa. When trained on only *Bacteroidetes* PULs, models predict PULs in non-*Bacteroidetes* with much lower accuracy across the board. However, the converse does not hold. Most models were capable of learning patterns from non-*Bacteroidetes* PULs and generalizing them to *Bacteroidetes* fairly well. Interestingly, GECCO generalizes best in this setting. By inspecting the model weights for every Pfam domain, we observed that GECCO assigns much higher weights to susD domains when trained on non-**Bacteroidetes**, potentially because gpPULs can contain susD homologs but no susC homologs. By Comparing the functional composition of predicted PULs in *Bacteroidetes* from this data partitioning with the 5-fold cross-validation, we observed that all models except for GECCO predict less PULs with susC/susD domains when trained on non-*Bacteroidetes*. We also see that all models continue to rely on CAZymes to predict PULs, as these are consistently present across predicted PULs in both settings.

Main RQ: How effective are machine learning techniques for predicting polysaccharide utilization loci?

To conclude, the results show that ML is a viable approach for PUL prediction. Although none of the evaluated models manage to achieve very high accuracies, this should be considered in the context of limited training data and the incomplete nature of experimental annotations. Model performance improved substantially after accounting for uncertain negatives in training, indicating the importance of constructing a robust and complete dataset for training and evaluation. Among the evaluated approaches, models learning from functional domain-based genome representations consistently achieved the best performance, with the Transformer-based GeneCAT model outperforming the evaluated classical machine learning method GECCO and the protein sequence-based foundation models ESMC and Bacformer in a zero-shot setting. Finally, as several models identified putative PULs in diverse Gram-positive bacteria that currently lack experimental annotations, we show that ML is effective as a guiding tool for biological discovery and hypothesis generation.

6.2 Limitations

Despite the promising results presented in this thesis, several limitations should be considered. First and foremost, despite our efforts to curate a robust dataset, the quality of the data and the incomplete nature of experimental annotations still form the main limitations of this study. Even though the best performing models were overall capable of predicting PULs and even identifying novel putative PULs, predictive performance remained moderate, reaching at best an AUPRC score of 0.55 on experimental PULs and 0.64 on cryptic PULs.

Additional limitations are related to the models we trained. In the comparisons with GeneCAT and GECCO, we evaluated the models ESMC and Bacformer exclusively in a zero-shot setting. While this provides insight into the applicability of the learned representations for this task, it

does not fully represent the capabilities of these models, and end-to-end finetuning may result in improved performance. On a similar note, all zero-shot evaluations were performed with logistic regression on gene embeddings, while it is possible that more complex methods, such as support vector machines or multilayer perceptrons, may be able to capture additional non-linear structures in the embedding space and thereby improve performance.

Another general limitation of this study is that evaluation was performed primarily at the gene level. Since PULs are gene clusters, gene-level performance does not necessarily translate to accurate identification of complete clusters. Indeed, the results show that many models tended to produce fragmented predictions, identifying only subsets of genes within known PULs.

Finally, while the applicability of ML for PUL prediction was explored and evaluated, we did not produce a fully robust PUL annotation framework. We show that our models are capable of identifying PUL-associated genes to a certain degree, but no dedicated pipeline was developed to reconstruct and validate full-length PULs from these predictions.

6.3 Future Work

Building upon the results and limitations of this thesis, several directions for future research can be identified. First, finetuning Bacformer on PUL prediction datasets may improve performance compared to the evaluated zero-shot approach, providing a better comparison with GeneCAT. Similarly, a follow-up project could explore using non-linear classifiers such as multilayer perceptrons, instead of only relying on logistic regression in a zero-shot setting. We could also test a variety of other gLMs for this task and eventually develop this PUL prediction setting into a more comprehensive benchmark for such models.

As we previously observed that GECCO predicted less fragmented PULs (Figure 8d), possibly due to using a conditional random field (CRF), we could explore combining GeneCAT with a CRF-based layer, to potentially improve cluster-level predictions and reduce fragmentation. To properly test these additions, we should address the aforementioned limitation regarding gene-level evaluation and extend our evaluation framework to the cluster level. For example, we could look at the extent of overlap between experimental and predicted PULs, in a manner similar to the "segment overlap" metric proposed in the GECCO pre-print [CLF+21].

More broadly, the application of explainable AI techniques may provide insight into the biological motivations behind predictions and improve interpretability for Transformer-based models [BSSZ25]. In addition, positive-unlabeled learning approaches could offer an alternative solution to the challenge of accurately defining negative examples, by explicitly accounting for uncertainty in the negative class [YLC+14]. Another approach to handle this issue is extending the set of cryptic PULs with CAZyme-containing Gene Clusters (CGCs) found by the CGC-finder tool, which would help us identify more cryptic PULs in non-*Bacteroidetes* [HZW+18]. With that, we could also explore different training procedures regarding cryptic PULs, such as labeling them as positives.

Finally, we could apply these models to other gut bacteria with poorly characterized PULs, and possibly even archaea, where PUL-like systems remain largely unexplored. This could allow us to discover additional putative PULs and help guide future experimental validation efforts.

References

[AGM+90] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, October 1990.

[AZY+21] Catherine Ausland, Jinfang Zheng, Haidong Yi, Bowen Yang, Tang Li, Xuehuan Feng, Bo Zheng, and Yanbin Yin. dbCAN-PUL: a database of experimentally characterized CAZyme gene clusters and their substrates. *Nucleic Acids Research*, 49(D1):D523–D528, January 2021.

[BDK+26] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Mohsen Naghipourfar, Michael Poli, Gwanggyu Sun, Greg Brockman, Daniel Chang, Alison Fanton, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Jonathan C. Schmok, Ali Taghibakhshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Michael T. Pearce, Elana Simon, Etowah Adams, Zachary J. Amador, Euan A. Ashley, Stephen A. Baccus, Haoyu Dai, Steven Dillmann, Stefano Ermon, Daniel Guo, Michael H. Herschl, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, John St. John, Jeremy Sullivan, Joseph Tey, Ben Viggiano, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Sudarshan Pinglay, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modelling and design across all domains of life with Evo 2. *Nature*, 652(8112):1349–1361, April 2026.

[BGB21] Jonathon A. Briggs, Julie M. Grondin, and Harry Brumer. Communal living: glycan utilization by the human gut microbiota. *Environmental Microbiology*, 23(1):15–35, January 2021.

[BMB+07] Servane Blanvillain, Damien Meyer, Alice Boulanger, Martine Lautier, Catherine Guynet, Nicolas Denancé, Jacques Vasse, Emmanuelle Lauber, and Matthieu Arlat. Plant Carbohydrate Scavenging through TonB-Dependent Receptors: A Feature Shared by Phytopathogenic and Aquatic Bacteria. *PLOS ONE*, 2(2):e224, February 2007.

[BMF+13] Daniela Börnigen, Xochitl C. Morgan, Eric A. Franzosa, Boyu Ren, Ramnik J. Xavier, Wendy S. Garrett, and Curtis Huttenhower. Functional profiling of the gut microbiome in disease-associated inflammation. *Genome Medicine*, 5(7):65, July 2013.

[BMG06] Magnus K. Bjursell, Eric C. Martens, and Jeffrey I. Gordon. Functional Genomic and Metabolic Studies of the Adaptations of a Prominent Adult Human Gut Symbiont, *Bacteroides thetaiotaomicron*, to the Suckling Period*. *Journal of Biological Chemistry*, 281(47):36269–36279, November 2006.

[BSSZ25] Aishwarya Budhkar, Qianqian Song, Jing Su, and Xuhong Zhang. Demystifying the black box: A survey on explainable artificial intelligence (XAI) in bioinformatics. *Computational and Structural Biotechnology Journal*, 27:346–359, January 2025.

[BV20] Rosalin Bonetta and Gianluca Valentino. Machine learning techniques for protein function prediction. *Proteins: Structure, Function, and Bioinformatics*, 88(3):397–413, 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/prot.25832.

[BYA+25] Gonzalo Benegas, Chengzhong Ye, Carlos Albors, Jianan Canal Li, and Yun S. Song. Genomic language models: opportunities and challenges. *Trends in Genetics*, 41(4):286–302, April 2025.

[CHD+26] Salvatore Candido, Thomas Hayes, Alexander Derry, Roshan Rao, Zeming Lin, Robert Verkuil, Bryan Z. Wu, Jin Sub Lee, Elise S. Bruguera, Jehan A. Keval, Mykhailo Kopylov, John E. Pak, Wesley Wu, Neil Thomas, Samson Mataraso, Alvin Hsu, Ashton C. Trotman-Grant, Kilian Fatras, Allan dos Santos Costa, Rohil Badkundri, Halil Akın, Deniz Oktay, Jonathan Deaton, Elizabeth Montabana, Hrishita Sitwala, Yue Yu, Marius Wiggert, Dylan Alexander Carlin, Anthony W. Goering, Tomasz Blazejewski, McCullen Sandora, Michael Hla, Tina Z. Jia, Leon H. Kloker, Nicholas J. Sofroniew, Masatoshi Uehara, Jassi Pannu, Sharrol Bachas, Daniel S. Liu, Tom Sercu, and Alexander Rives. Language Modeling Materializes a World Model of Protein Biology, June 2026. ISSN: 2692-8205 Pages: 2026.06.03.729735 Section: New Results.

[CLF+21] Laura M. Carroll, Martin Larralde, Jonas Simon Fleck, Ruby Ponnudurai, Alessio Milanese, Elisa Cappio, and Georg Zeller. Accurate de novo identification of biosynthetic gene clusters with GECCO, May 2021. Pages: 2021.05.03.442509 Section: New Results.

[CMHP22] Pierre-Alain Chaumeil, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*, 38(23):5315–5316, December 2022.

[CRU+18] Jing Chen, Craig S. Robb, Frank Unfried, Lennart Kappelmann, Stephanie Markert, Tao Song, Jens Harder, Burak Avcı, Dörte Becher, Ping Xie, Rudolf I. Amann, Jan-Hendrik Hehemann, Thomas Schweder, and Hanno Teeling. Alpha- and beta-mannan utilization by marine Bacteroidetes. *Environmental Microbiology*, 20(11):4127–4140, 2018. _eprint: https://enviromicro-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/1462-2920.14414.

[DGD+22] Elodie Drula, Marie-Line Garron, Suzan Dogan, Vincent Lombard, Bernard Henrissat, and Nicolas Terrapon. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Research*, 50(D1):D571–D577, January 2022.

[DKG+26] Quinten R. Ducarmon, Nicolai Karcher, Samir Giri, Hanne L. P. Tytgat, Omar Delannoy-Bruno, Selin Pekel, Fabian Springer, Patrick Wörz, Christian Schudoma, Athanasios Typas, and Georg Zeller. Cayman enables large-scale analysis of gut microbiome carbohydrate-active enzyme repertoires. *Nature Microbiology*, pages 1–15, April 2026.

[DMV+23] Lauren E. Davey, Per N. Malkus, Max Villa, Lee Dolat, Zachary C. Holmes, Jeff Letourneau, Eduard Ansaldo, Lawrence A. David, Gregory M. Barton, and Raphael H. Valdivia. A genetic system for Akkermansia muciniphila reveals a role for mucin

foraging in gut colonization and host sterol biosynthesis gene expression. *Nature Microbiology*, 8(8):1450–1467, August 2023.

[DS08] Gideon J. Davies and Michael L. Sinnott. The sequence-based classifications of carbohydrate-active enzymes Sorting the diverse. *Biochemist*, 30(4):26–32, 2008.

[DSK+16] Mahesh S. Desai, Anna M. Seekatz, Nicole M. Koropatkin, Nobuhiko Kamada, Christina A. Hickey, Mathis Wolter, Nicholas A. Pudlo, Sho Kitamoto, Nicolas Terrapon, Arnaud Muller, Vincent B. Young, Bernard Henrissat, Paul Wilmes, Thaddeus S. Stappenbeck, Gabriel Núñez, and Eric C. Martens. A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell*, 167(5):1339–1353.e21, November 2016.

[FCK16] Matthew H. Foley, Darrell W. Cockburn, and Nicole M. Koropatkin. The Sus operon: a model system for starch uptake by the human gut Bacteroidetes. *Cellular and Molecular Life Sciences*, 73(14):2603–2617, July 2016.

[GBvW+21] Cameron L M Gilchrist, Thomas J Booth, Bram van Wersch, Liana van Grieken, Marnix H Medema, and Yit-Heng Chooi. cblaster: a remote search tool for rapid identification and visualization of homologous gene clusters. *Bioinformatics Advances*, 1(1):vbab016, January 2021.

[GTD+17] Julie M. Grondin, Kazune Tamura, Guillaume Déjean, D. Wade Abbott, and Harry Brumer. Polysaccharide Utilization Loci: Fueling Microbial Communities. *Journal of Bacteriology*, 199(15):e00860–16, July 2017.

[HCL+10] Doug Hyatt, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1):119, March 2010.

[HPP+19] Geoffrey D Hannigan, David Prihoda, Andrej Palicka, Jindrich Soukup, Ondrej Klempir, Lena Rampula, Jindrich Durcak, Michael Wurst, Jakub Kotowski, Dan Chang, Rurun Wang, Grazia Piizzi, Gergely Temesi, Daria J Hazuda, Christopher H Woelk, and Danny A Bitton. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Research*, 47(18):e110, October 2019.

[HRA+24] Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model, December 2024. Pages: 2024.07.01.600583 Section: New Results.

[HZW+18] Le Huang, Han Zhang, Peizhi Wu, Sarah Entwistle, Xueqiong Li, Tanner Yohe, Haidong Yi, Zhenglu Yang, and Yanbin Yin. dbCAN-seq: a database of carbohydrate-active enzyme (CAZyme) sequence and annotation. *Nucleic Acids Research*, 46(D1):D516–D521, January 2018.

[JZLD21] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15):2112–2120, August 2021.

[KAG+13] Abdessamad El Kaoutari, Fabrice Armougom, Jeffrey I. Gordon, Didier Raoult, and Bernard Henrissat. The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, 11(7):497–504, July 2013.

[KWC+20] Deepank R Korandla, Jacob M Wozniak, Anaamika Campeau, David J Gonzalez, and Erik S Wright. AssessORF: combining evolutionary conservation and proteomics to assess prokaryotic gene predictions. *Bioinformatics*, 36(4):1022–1029, February 2020.

[Lar22] Martin Larralde. Pyrodigal: Python bindings and interface to Prodigal,an efficient method for gene prediction in prokaryotes. *Journal of Open Source Software*, 7(72):4296, April 2022.

[LBB+22] Ana S. Luis, Arnaud Baslé, Dominic P. Byrne, Gareth S. A. Wright, James A. London, Chunsheng Jin, Niclas G. Karlsson, Gunnar C. Hansson, Patrick A. Eyers, Mirjam Czjzek, Tristan Barbeyron, Edwin A. Yates, Eric C. Martens, and Alan Cartmell. Sulfated glycan recognition by carbohydrate sulfatases of the human gut microbiota. *Nature Chemical Biology*, 18(8):841–849, August 2022.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[LCC+25] Andrew K. Lampinen, Arslan Chaudhry, Stephanie C. Y. Chan, Cody Wild, Diane Wan, Alex Ku, Jörg Bornschein, Razvan Pascanu, Murray Shanahan, and James L. McClelland. On the generalization of language models from in-context learning and finetuning: a controlled study, November 2025. arXiv:2505.00661 [cs.CL].

[LGBE24] Avantika Lal, David Garfield, Tommaso Biancalani, and Gokcen Eraslan. regLM: Designing Realistic Regulatory DNA with Autoregressive Language Models. In Jian Ma, editor, *Research in Computational Molecular Biology*, pages 332–335, Cham, 2024. Springer Nature Switzerland.

[LMG+21] Xuemei Liu, Bingyong Mao, Jiayu Gu, Jiaying Wu, Shumao Cui, Gang Wang, Jianxin Zhao, Hao Zhang, and Wei Chen. Blautia—a new functional genus with potential probiotic properties? *Gut Microbes*, 13(1):1875796, January 2021. _eprint: https://doi.org/10.1080/19490976.2021.1875796.

[LN15] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, June 2015.

[LOKPC16] Imchang Lee, Yeong Ouk Kim, Sang-Cheol Park, and Jongsik Chun. OrthoANI: An improved algorithm and software for calculating average nucleotide identity. *International Journal of Systematic and Evolutionary Microbiology*, 66(2):1100–1103, 2016.

[LRLM+19] Sabina Leanti La Rosa, Maria Louise Leth, Leszek Michalak, Morten Ejby Hansen, Nicholas A. Pudlo, Robert Glowacki, Gabriel Pereira, Christopher T. Workman,

Magnus Ø Arntzen, Phillip B. Pope, Eric C. Martens, Maher Abou Hachem, and Bjørge Westereng. The human gut Firmicute Roseburia intestinalis is a primary degrader of dietary -mannans. *Nature Communications*, 10(1):905, February 2019.

[LTB22] Jee-Yon Lee, Renée M. Tsolis, and Andreas J. Bäumler. The microbiome and gut homeostasis. *Science*, 377(6601):eabp9960, July 2022.

[LYL+10] Ziyong Liu, Yu Ying, Fuli Li, Cuiqing Ma, and Ping Xu. Butanol production by Clostridium beijerinckii ATCC 55025 from wheat bran. *Journal of Industrial Microbiology and Biotechnology*, 37(5):495–501, May 2010.

[LZ23] Martin Larralde and Georg Zeller. PyHMMER: a Python library binding to HMMER for efficient sequence analysis. *Bioinformatics*, 39(5):btad214, May 2023.

[LZC25] Martin Larralde, Georg Zeller, and Laura M Carroll. PyOrthoANI, PyFastANI, and Pyskani: a suite of Python libraries for computation of average nucleotide identity. *NAR Genomics and Bioinformatics*, 7(3):lqaf095, July 2025.

[MCW+21] Jaina Mistry, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik L L Sonnhammer, Silvio C E Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, Robert D Finn, and Alex Bateman. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49(D1):D412–D419, January 2021.

[MHM20] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, September 2020. arXiv:1802.03426 [stat.ML].

[MLC+11] Eric C. Martens, Elisabeth C. Lowe, Herbert Chiang, Nicholas A. Pudlo, Meng Wu, Nathan P. McNulty, D. Wade Abbott, Bernard Henrissat, Harry J. Gilbert, David N. Bolam, and Jeffrey I. Gordon. Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts. *PLOS Biology*, 9(12):e1001221, December 2011.

[MMS+25] Indrani Mukhopadhya, Jennifer C. Martin, Sophie Shaw, Martin Gutierrez-Torrejon, Nikoleta Boteva, Aileen J. McKinley, Silvia W. Gratz, and Karen P. Scott. Novel insights into carbohydrate utilisation, antimicrobial resistance, and sporulation potential in Roseburia intestinalis isolates across diverse geographical locations. *Gut Microbes*, 17(1):2473516, December 2025. _eprint: https://doi.org/10.1080/19490976.2025.2473516.

[OSML+16] Paul O. Sheridan, Jennifer C. Martin, Trevor D. Lawley, Hilary P. Browne, Hugh M. B. Harris, Annick Bernalier-Donadille, Sylvia H. Duncan, Paul W. O'Toole, Karen P. Scott, and Harry J. Flint. Polysaccharide utilization loci and nutritional specialization in a dominant group of butyrate-producing human colonic Firmicutes. *Microbial Genomics*, 2(2):e000043, 2016.

[PCC+20] Donovan H. Parks, Maria Chuvochina, Pierre-Alain Chaumeil, Christian Rinke, Aaron J. Mussig, and Philip Hugenholtz. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology*, 38(9):1079–1086, September 2020.

[PRMS+25] Piotr Przymus, Krzysztof Rykaczewski, Adrián Martín-Segura, Jaak Truu, Enrique Carrillo De Santa Pau, Mikhail Kolev, Irina Naskinova, Aleksandra Gruca, Alexia Sampri, Marcus Frohme, and Alina Nechyporenko. Deep learning in microbiome analysis: a comprehensive review of neural network models. *Frontiers in Microbiology*, 15, January 2025.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[PZS+22] Ana Maria Porras, Hao Zhou, Qiaojuan Shi, Xieyue Xiao, JRI Live Cell Bank, Randy Longman, and Ilana Lauren Brito. Inflammatory Bowel Disease-Associated Gut Commensals Degrade Components of the Extracellular Matrix. *mBio*, 13(6):e02201–22, November 2022.

[RCD+22] Yunxiao Ren, Trinad Chakraborty, Swapnil Doijad, Linda Falgenhauer, Jane Falgenhauer, Alexander Goesmann, Anne-Christin Hauschild, Oliver Schwengers, and Dominik Heider. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*, 38(2):325–334, January 2022.

[RM22] Etienne Routhier and Julien Mozziconacci. Genomics enters the deep learning era. *PeerJ*, 10:e13613, June 2022.

[RRL99] Sharon J. Reid, M. Suhail Rafudeen, and Neil G. Leat. The genes controlling sucrose utilization in Clostridium beijerinckii NCIMB 8052 constitute an operon. *Microbiology*, 145(6):1461–1472, 1999.

[SARW18a] Rob D. Stewart, Marc D. Auffret, Rainer Roehe, and Mick Watson. Open prediction of polysaccharide utilisation loci (PUL) in 5414 public Bacteroidetes genomes using PULpy, September 2018. Pages: 421024 Section: New Results.

[SARW18b] Rob D. Stewart, Marc D. Auffret, Rainer Roehe, and Mick Watson. Open prediction of polysaccharide utilisation loci (PUL) in 5414 public Bacteroidetes genomes using PULpy, September 2018. Pages: 421024 Section: New Results.

[Sch00] Robert Schleif. Regulation of the l-arabinose operon of *Escherichia coli*. *Trends in Genetics*, 16(12):559–565, December 2000.

[SR15] Takaya Saito and Marc Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, March 2015.

[Tea24] ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning, December 2024.

[TLD+18] Nicolas Terrapon, Vincent Lombard, Élodie Drula, Pascal Lapébie, Saad Al-Masaudi, Harry J Gilbert, and Bernard Henrissat. PULDB: the expanded database of Polysac-

charide Utilization Loci. *Nucleic Acids Research*, 46(Database issue):D677–D683, January 2018.

[TLGH15] Nicolas Terrapon, Vincent Lombard, Harry J. Gilbert, and Bernard Henrissat. Automatic prediction of polysaccharide utilization loci in Bacteroidetes species. *Bioinformatics*, 31(5):647–655, March 2015.

[vdHW21] Bart van der Hee and Jerry M. Wells. Microbial Regulation of Host Physiology by Short-chain Fatty Acids. *Trends in Microbiology*, 29(8):700–712, August 2021.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[WLBF14] Bryan A. White, Raphael Lamed, Edward A. Bayer, and Harry J. Flint. Biomass Utilization by Gut Microbiomes*. *Annual Review of Microbiology*, 68(Volume 68, 2014):279–296, September 2014.

[WTN+25] Maciej Wiatrak, Ramon Viñas Torné, Maria Ntemourtsidou, Adam Dinan, David C. Abelson, Divya Arora, Maria Brbić, Aaron Weimann, and R. Andres Floto. A contextualised protein language model reveals the functional syntax of bacterial evolution, August 2025. ISSN: 2692-8205 Pages: 2025.07.20.665723 Section: New Results.

[YHK+24] Ting Ye, Ayako Horigome, Hiroki Kaneko, Toshitaka Odamaki, Kanefumi Kitahara, and Kiyotaka Fujita. Degradation mechanism of difructose dianhydride III in Blautia species. *Applied Microbiology and Biotechnology*, 108(1):502, November 2024.

[YLC+14] Peng Yang, Xiaoli Li, Hon-Nian Chua, Chee-Keong Kwoh, and See-Kiong Ng. Ensemble Positive Unlabeled Learning for Disease Gene Identification. *PLOS ONE*, 9(5):e97079, May 2014.

[YTZ+22] Qun-Jian Yin, Hong-Zhi Tang, Fang-Chao Zhu, Xu-Yang Chen, De-Wei Cheng, Li-Chang Tang, Xiao-Qing Qi, and Xue-Gong Li. Complete genome sequence and Carbohydrate Active Enzymes (CAZymes) repertoire of *Gilvimarinus* sp. DA14 isolated from the South China Sea. *Marine Genomics*, 65:100982, October 2022.

[ZGY+23] Jinfang Zheng, Qiwei Ge, Yuchen Yan, Xinpeng Zhang, Le Huang, and Yanbin Yin. dbCAN3: automated carbohydrate-active enzyme and substrate annotation. *Nucleic Acids Research*, 51(W1):W115–W121, July 2023.

[ZJL+24] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana Davuluri, and Han Liu. DNABERT-2: Efficient Foundation Model and Benchmark For Multi-Species Genome, March 2024. arXiv:2306.15006 [q-bio].

[ZTV+14] Georg Zeller, Julien Tap, Anita Y. Voigt, Shinichi Sunagawa, Jens Roat Kultima, Paul I. Costea, Aurélien Amiot, Jürgen Böhm, Francesco Brunetti, Nina Habermann, Rajna Hercog, Moritz Koch, Alain Luciani, Daniel R. Mende, Martin A. Schneider, Petra Schrotz-King, Christophe Tournigand, Jeanne Tran Van Nhieu, Takuji Yamada, Jürgen Zimmermann, Vladimir Benes, Matthias Kloor, Cornelia M. Ulrich, Magnus von

Knebel Doeberitz, Iradj Sobhani, and Peer Bork. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Molecular Systems Biology*, 10(11):MSB145645, November 2014.

[ZYH+18] Han Zhang, Tanner Yohe, Le Huang, Sarah Entwistle, Peizhi Wu, Zhenglu Yang, Peter K Busk, Ying Xu, and Yanbin Yin. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, 46(W1):W95–W101, July 2018.

A Supplementary Tables

Supplementary Table 1: Manual curation steps applied to resolve inconsistencies in PUL annotations and genome identifiers across databases.

Action	Database	Description	Accession / Identifier
Added manually	dbCAN-PUL	One PUL-containing genome was downloaded separately from JGI due to missing an entry in NCBI nucleotide.	*Ga0139390_150*
Split into separate PULs	dbCAN-PUL	Two PULs spanning multiple contigs were mapped to a complete genome using BLAST. Although initially appearing connected, they corresponded to distinct loci in the full genome and were split into two PULs.	*ADWO01000021.1*, *ADWO01000020.1* → *CP091800*
Removed	PULDB	Entries were removed due to missing sequence identifiers and lack of availability in public repositories or author-provided data.	SEQ15336-1, SEQ15336-2_ori, Contig5_1_7083079
Identifier recovery	PULDB	Contigs with missing identifiers were recovered by cross-referencing the original publication ([CRU+18]) and replaced with correct genome accessions.	FG27DRAFT_unitig_0_quiver_dupTrim_7536 → NZ_JQNQ01000001 P164DRAFT_scf7180000000008_quiver → JPOL01000002 P164DRAFT_scf7180000000009_quiver → JPOL01000003

Supplementary Table 2: Overview of raw data counts, showing the number of sequences and PULs extracted from each database, before any additional curation and preprocessing steps.

	Total	dbCAN-PUL	PULDB
Genomes/contigs	414	370	44
PULs	991	633	358

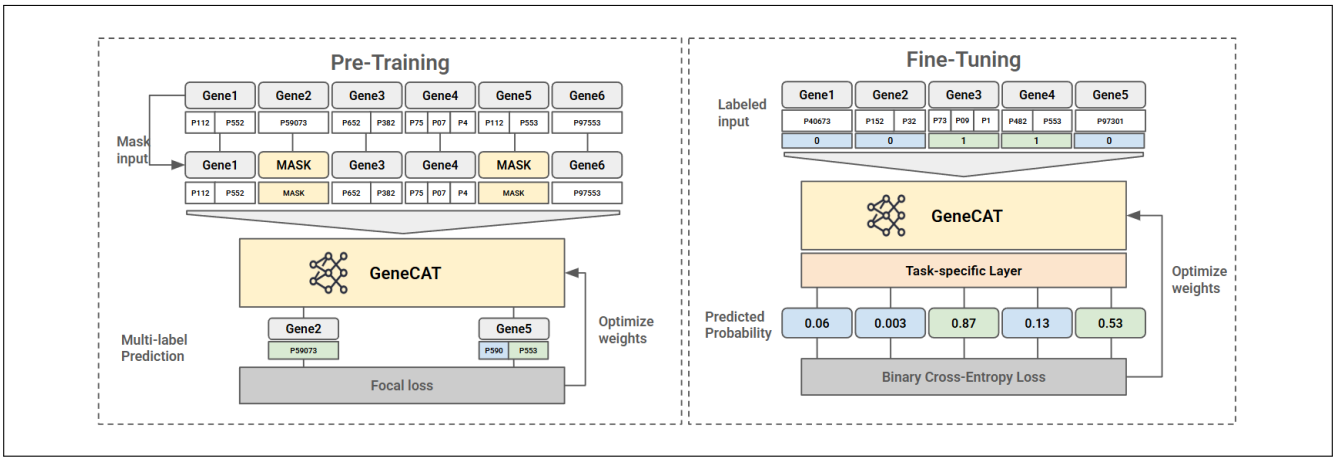
Supplementary Table 3: Domain identifier and names of Pfam domains associated with susC- and susD-like proteins.

Protein	**Pfam domain**	**Domain name**
susC	PF00593	TonB dependent receptor-like, beta-barrel
susC	PF07715	TonB-dependent Receptor Plug Domain
susD	PF07980	susD family
susD	PF12741	Susd and RagB outer membrane lipoprotein
susD	PF12771	Starch-binding associating with outer membrane
susD	PF14322	Starch-binding associating with outer membrane

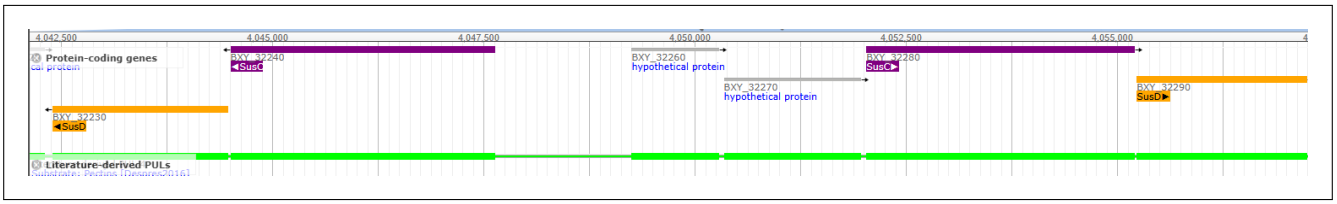
Supplementary Table 4: Performance and training duration of all tested model and feature combinations. Performance is evaluated using AUPRC and training duration is estimated per fold. For zero-shot models, both embeddings generation and logistic regression training are taken into account, but not pretraining.

Model (feature set, masking)	**AUPRC**	**Training duration**
GECCO (Pfam)	0.37	**5–10 mins**
GECCO (Pfam+CAZy)	0.36	**5–10 mins**
GeneCAT zero-shot (Pfam)	0.35	1.2–1.5h
GeneCAT zero-shot (Pfam+CAZy)	0.38	1.2–1.5h
GeneCAT zero-shot (Pfam, masked)	0.41	1.2–1.5h
GeneCAT zero-shot (Pfam+CAZy, masked)	0.44	1.2–1.5h
GeneCAT pretrained-finetuned (Pfam)	0.41	4.5–7h
GeneCAT pretrained-finetuned (Pfam+CAZy)	0.42	4.5–7h
GeneCAT pretrained-finetuned (Pfam, masked)	0.52	4.5–7h
GeneCAT pretrained-finetuned (Pfam+CAZy, masked)	**0.55**	4.5–7h
GeneCAT untrained-finetuned (Pfam+CAZy, masked)	0.48	4.5–7h
ESM Cambrian	0.15	2.5–3h
ESM Cambrian (masked)	0.18	2.5–3h
Bacformer	0.17	3–3.5h
Bacformer (masked)	0.19	3–3.5h

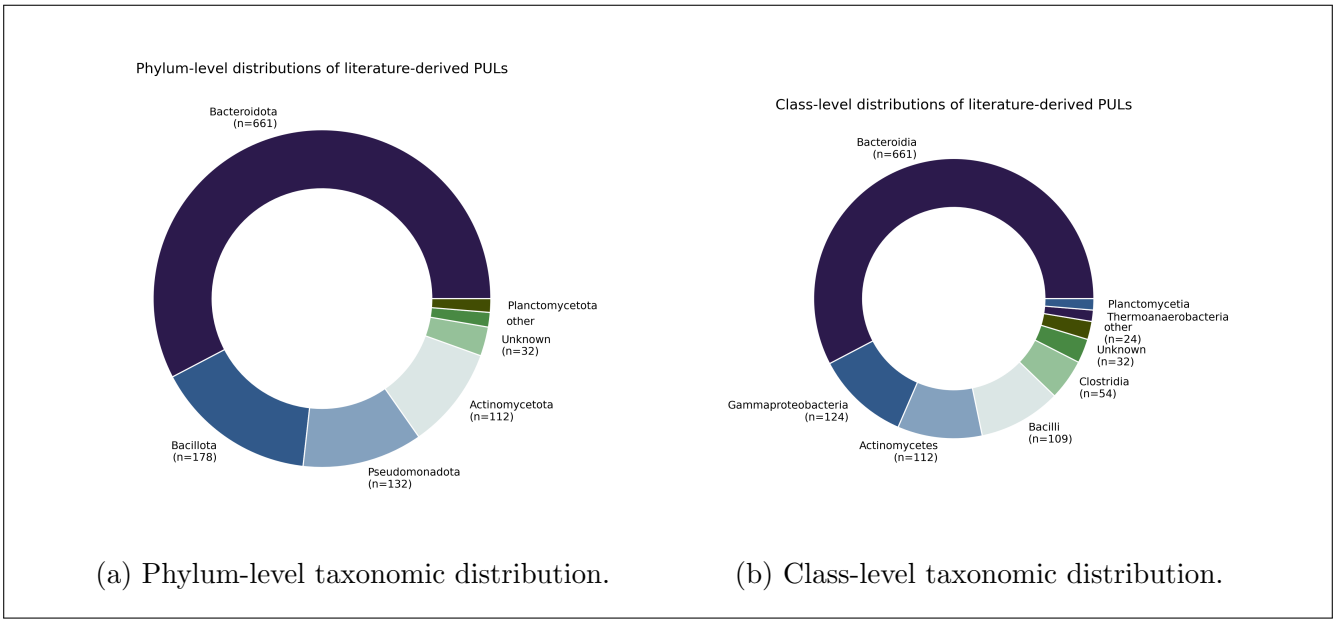
B Supplementary Figures



Supplementary Figure 1: Graphical overview of transfer learning with GeneCAT. Left: pre-training with self-supervised learning. Right: Fine-tuning on a binary classification task.



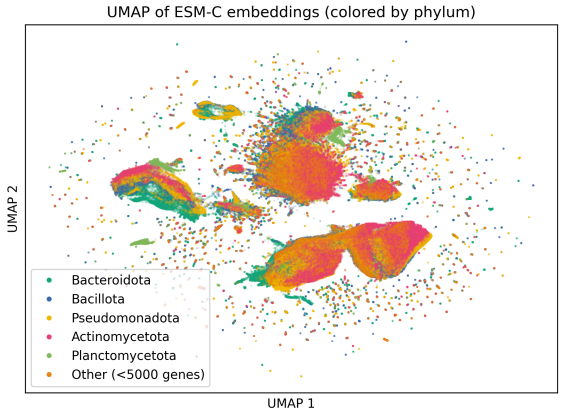
Supplementary Figure 2: Screenshot of Literature-Derived PUL 4 from PULDB.



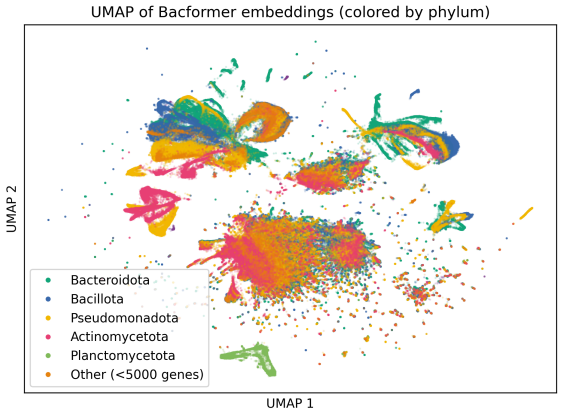
(a) Phylum-level taxonomic distribution.

(b) Class-level taxonomic distribution.

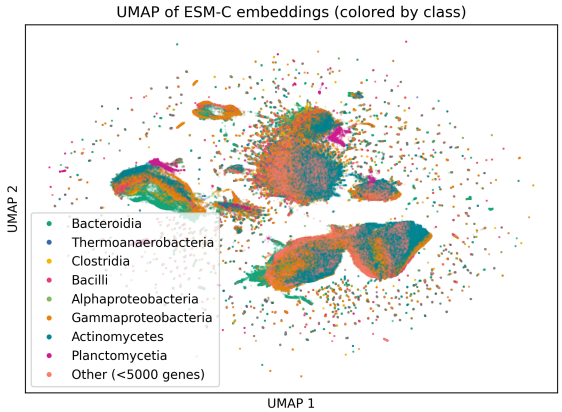
Supplementary Figure 3: Taxonomic distributions of literature-derived PULs in the dataset, based on GTDB annotations.



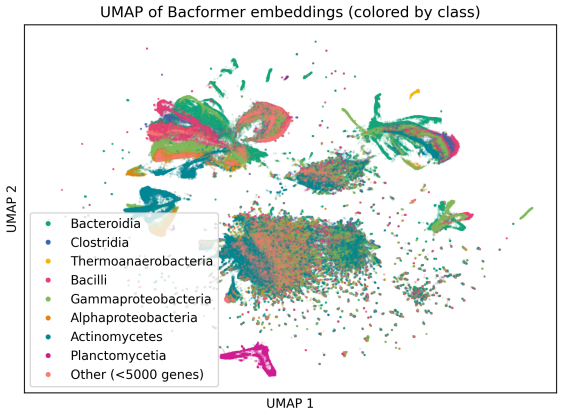
(a) UMAP projection of embeddings from ESMC colored by phylum



(b) UMAP projection of embeddings from Bacformer colored by phylum

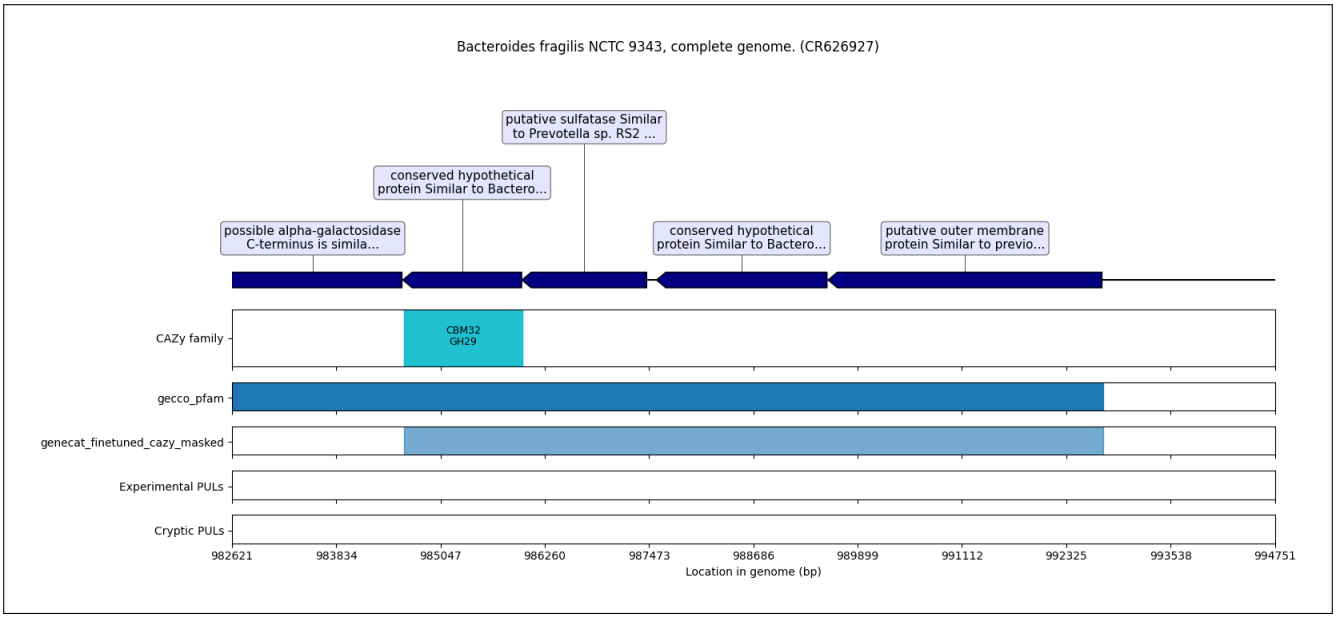


(c) UMAP projection of embeddings from ESMC colored by class



(d) UMAP projection of embeddings from Bacformer colored by class

Supplementary Figure 4: UMAP projections of ESM Cambrian and Bacformer Large gene embeddings, colored by taxonomic rank. Taxa containing less than 5000 gene embeddings were grouped together as "Other".



Supplementary Figure 5: Putative PUL predicted by GECCO and GeneCAT in the *Bacteroides fragilis* NCTC 9343 genome, without known experimental annotations and not predicted by PULpy.