Assessment of Enzyme Stability in Subsurface Sediments by Computational Methods

A Thesis Presented for the
Master of Science
Degree
The University of Tennessee, Knoxville

Kambiz Kalhor
August 2024

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor, Dr Andrew Steen, whose expertise, understanding, and patience, added considerably to my graduate experience.

I owe my heartfelt gratitude to my committee members, Dr. Erik Zinser and Dr. Frank Löffler whose insights and feedback were invaluable throughout this process. Your perspectives and rigorous reviews significantly enhanced the quality of my research.

ABSTRACT

The microorganisms found in marine subseafloor sediment play a vital role in global carbon and nitrogen cycles, with an estimated 2.9×10^{29} cells, accounting for 0.6% of Earth's total living biomass. These microbes grow at a very slow rate, with carbon turnover occurring over the course of years to thousands of years, about six orders of magnitude slower than sulfate reducing bacteria in pure culture. These slow metabolic rates suggest that the enzymes they produce must also have extended lifespans in order to be effective over such long periods of time. As a result, these enzymes are likely to be highly stable. The central question of my thesis is: "**Do subsurface microbes have more stable proteins compared to other organisms?**". I tested this hypothesis by examining various Carbohydrate-Active enzyme (CAZy) families within sediment samples obtained from the Baltic Sea at two distinct depths (25 cm and 15 meters below the seafloor). previous research suggested a distinction in the stabilities of proteins belonging to two distinct families, GH29 and GH109.

I investigated the stability of proteins by using Gibbs free energy of folding ΔG_{fold}. Some prokaryotic proteins intended for export have an N-terminal signal peptide that directs them to the secretion pathway. Having this signal peptide signifies a substantial environmental change, therefore I conducted a comparative analysis of protein pairs, ensuring both pairs either contained signal peptides or lacked them, revealing a significant stability difference, with higher stability noted at greater depth.

Also, the stability is notably higher in protein pairs lacking signal peptides in deep samples compared to surface samples, predominantly among pairs from the glycoside hydrolase (GH) family. Contrary to our expectations, our investigation revealed no significant difference in the stability of extracellular proteins between deep and surface samples. This outcome does not align with our initial hypothesis, indicating that the environmental depth may impact the stability of these proteins. This suggests that the presence of a signal peptide may play a role in the environmental adaptability and stability of proteins, providing a specific area of interest for further research to understand the protein stability across different environments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION AND BACKGROUND

1.1 Stability of enzymes in subsurface sediments

Marine microorganisms play a crucial role in determining the fate of organic carbon in the oceans. These tiny organisms are able to persist in the low-energy conditions found in deep marine sediments, where they play a vital role in global carbon and nitrogen cycles (LaRowe et al., 2020). It is estimated that there are 2.9×10^{29} microbial cells in global subseafloor sedimentary environments, representing approximately 4.1 petagrams of carbon and approximately 0.6% of Earth's total living biomass (Kallmeyer et al., 2012). Despite the difficult conditions, these microorganisms are an integral part of the marine ecosystem (Kallmeyer et al., 2012).

Despite significant progress in understanding the diversity and importance of marine microorganisms, the vast majority of these organisms remain uncultured in the laboratory. There are several possible reasons for this difficulty in cultivating marine bacteria, including the lack of essential nutrients or osmotic support, slow growth rates, the formation of micro-colonies, and the presence of senescent or viable but nonculturable (VBNC) cells (Zhang et al., 2021). In the marine environment, many bacteria have been found to enter a dormant state, characterized by the VBNC state, in which they are still metabolically active but cannot be grown on standard laboratory media (Zhang et al., 2021).

The microbes that inhabit marine sediments grow at a very slow rate, with biomass carbon turnover occurring over the course of years to thousands of years, which is about six orders of magnitude slower than sulfate-reducing bacteria in pure culture (Braun et al., 2017). The slow rate of metabolism and the long lifespans of these subsurface microbes suggests that the enzymes they produce must also have extended lifespans in order to be effective over such long periods of time (Schmidt et al., 2021). As a result, these enzymes are likely to be highly stable and resistant to environmental conditions. The slow growth rate of these marine microbes may be related to the stability of their proteins (Hoehler & Jørgensen, 2013).

A complex interplay between enzyme activity and stability has been studied, wherein an increase of one is often associated with a decrease in the other. The catalytic efficiency of enzymes is intricately dependent on their operational environment, necessitating adaptative mechanisms in response to the sometimes-extreme conditions characterizing their habitats,

such as variations in pH, temperature, and ion concentrations (Hou et al., 2023; Jaenicke & Böhm, 1998). For example, evolutionary adaptations have enabled certain enzymes to optimize their catalytic efficiency under suboptimal temperatures, primarily through adjustments in the flexibility of their catalytic domains (Arcus et al., 2020; Xie et al., 2014). This adaptation suggests a potential for similar evolutionary strategies among subseafloor enzymes, allowing them to be efficient under suboptimal conditions in trade with their activity in the face of the unique challenges posed by their environment like low biomass, growth-limiting conditions and high pressure.

1.2 Free energy of folding as a proxy for enzyme stability

Proteins are important biological molecules that perform various functions, such as catalyzing chemical reactions and interacting with other molecules. The stability of a protein refers to its ability to maintain its structure and function under different conditions, such as changes in temperature, pH, or the presence of other molecules (Hamborg et al. 2020). In this study, we aim to investigate the stability of proteins by using Gibbs free energy of folding (ΔG_{fold}) as a metric of protein thermodynamic stability (Pucci and Rooman 2016). Several tools can be utilized to compute ΔG_{fold}. The initial step of our computational approach involves identifying the protein structures in our samples. Our goal is to use ΔG_{fold} to understand the stability of proteins in our samples.

1.3 Central question

The central question of my thesis was: "Do these subsurface microbes have more stable proteins compared to other organisms?" My hypothesis was to assess enzyme stability in subsurface sediments by computational methods and compare them to enzyme stability on the surface. I tested my hypothesis by examining various extracellular and intracellular protein families to gain a general understanding of the concept. I studied the stability of the major functional category of CAZy enzymes in bacterial communities living in two different environments, the 15m below surface (15mbsf) bacterial communities and 25cm below surface (25cmbsf) bacterial communities. To test protein stability, I computationally estimated the free energy of folding (thermodynamic stability) of proteins, which is a proxy representing the thermodynamic stability of proteins in our samples. The metagenomes I used in our preliminary analysis were collected from surface and deep sediments from the Little Belt site in the Baltic Sea as part of IODP Expedition 347, and in our future investigation, I will use other datasets in various geographical regions to get comprehensive insight on this concept. I compared the

stability of surface versus subsurface extracellular proteins (four different families including glycoside hydrolases, glycosyltransferases, polysaccharide lyases and carbohydrate esterases) by calculating the free energy of folding of homologous protein pairs in surface and subsurface. I calculated ΔG_{fold} by predicting protein structure using AlphaFold2 (Jumper, et al. 2021) and then using those structures to calculate ΔG_{fold} using FoldX (Schymkowitz et al., 2005). The computational results then were compared by paired t-tests.

Initial findings reveal a difference between the stabilities of surface and deep proteins belonging to two families, GH29 and GH109, as demonstrated in Figure 1 (Steen et al., 2022). The stability of homologous protein pairs from these families was calculated in samples taken from both deep and surface sediment layers of the Baltic Sea. Glycoside Hydrolase Family 29, which includes enzymes known as fucosidases, have a role in carbohydrate metabolism through the hydrolysis of glycosidic bonds in complex carbohydrates, specifically targeting fucose-containing molecules, and Glycoside Hydrolase Family 109, recognized for its enzymatic activity catalyzing the hydrolysis of diverse glycosidic bonds, were examined. The investigation brought to light that homologous protein pairs found in the deeper sediment layers possessed markedly greater stability, suggesting a potential adaptation to conditions prevalent in such environments.

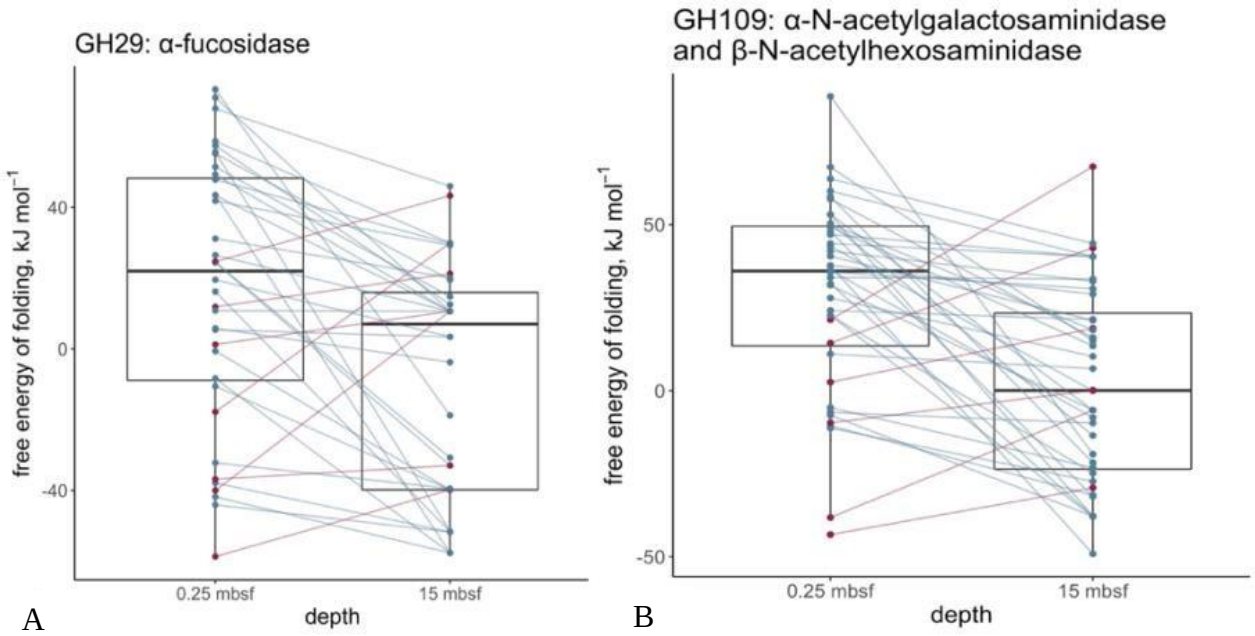

Figure 1 These plots compare the stability of pairs of homologous proteins in two different protein families in our deep and surface samples. The vertical axis represents the free energy of folding, with lower values indicating greater stability. Each dot represents a protein, and each line connects a pair of homologous proteins.

CHAPTER 2 MATERIAL AND METHODS

2.1 Sample Collection and Extraction

Samples were collected during IODP Expedition 347 to assess active microbial communities in the Baltic Sea. The sediment sample, identified as BSB59E15mbsf (Bio Sample: SAMN08473126), was retrieved from the Little Belt, a narrow strait in Denmark, located at 55.079 N 10.238 E (Andrén et al., 2015). The sediment collection was achieved using an advanced piston corer ton October 1, 2013 (*Expedition 347 Summary*, 2015). Cores were divided into 1.5-meter sections onboard to screen for perfluorinated chemicals (PFCs) contamination, a marker of drilling interference (*Expedition 347 Summary*, 2015). These sections underwent analysis with a rapid multiple-scanning core logger and were then segregated into whole round cores within a microbiology container kept at 12°C on the ship. For nucleic acid analysis, sediment cores were immediately frozen at -80°C onboard and transported to land-based laboratories on dry ice. After collection, a 10g sediment sample was processed with the MoBio PowerSoil kit, aimed at extracting high-quality genomic DNA from environmental samples (Bird et al., 2019). To capture sediment from the initial few meters not recovered by drilling, two missions led by the Center for Geomicrobiology at Aarhus University aboard the R/V Aurora took place. In September 2014, site M59 was revisited. Sample M59E-0.25cm was obtained through gravity coring, followed by subsampling via windows cut into the sediment core liner and the use of sterilized 20-ml syringes with the ends removed. Care was exercised to prevent potential seawater contamination by visually examining the cores for signs of seawater intrusion and by extracting DNA from sediments taken from the core's interior. All samples were immediately frozen at -80°C on the ship and dispatched to the United States on dry ice (Andrén et al., 2015).

The depth affected by bioturbation can vary significantly depending on the ecosystem and the specific bioturbating organisms involved. In deep-sea environments like the Baltic Sea, intense bioturbation can enhance the mixing of sediments to depths of 20-100 cm over decadal timescales (Aller & Cochran, 2019). We selected the deep sample from 15 m below the seafloor as well as below the zone of bioturbation and, therefore, in the “deep” biosphere.

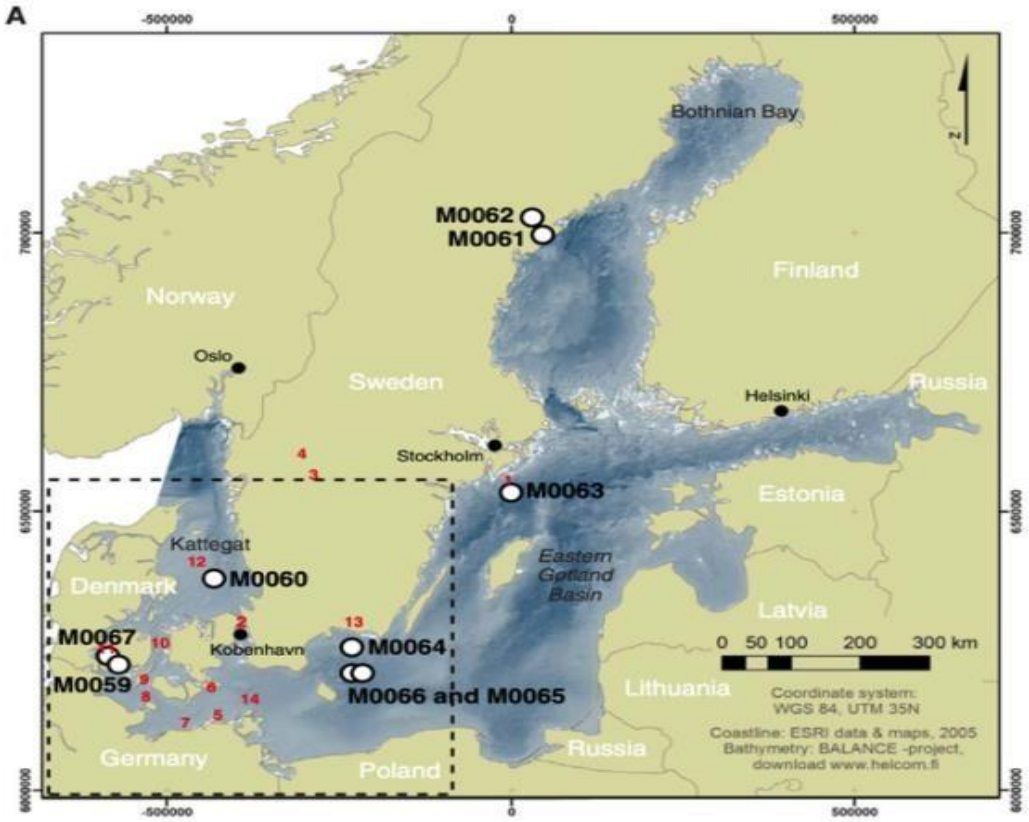

Figure 2 During the Integrated Ocean Drilling Program Expedition 347, sediment samples were obtained from the Baltic Sea at two distinct depths (25 cm and 15 meters below the seafloor) from a specific location designated as M0064. These samples were collected in order to investigate and compare microbial communities at varying depths within the exact location (Andrén et al. 2015).

Table 1 Sample descriptions

Sample identifiers	**isolation source**	**collection date**	**geographic location**	**latitude and longitude**	**sample oxygen content**	**sample collection device or method**	**sample material processing**	**sample size**
15mbsf SRR7066492	sediment	2013-10-01	Denmark: Little Belt	55.079 N 10.238 E	anoxic	Advance piston coring	MoBio PowerSoil kit	10g
25cmbsf; SRR7066493	sediment	2016-06-06	Denmark: Little Belt	55.079 N 10.238 E	anoxic	Rumohr corer	MoBio PowerSoil kit	5 g

2.2 DNA extraction and sequencing

DNA was isolated from the sediment samples M59E-0.25m using the DNeasy PowerMax Soil Kit from Mo Bio Laboratories, located in Carlsbad, CA. Sediment for the extraction was taken from frozen whole round cores (M59E-15m) or the ends of cut-off syringes (M59E-0.25m) in a specially equipped clean room at Texas A&M University Corpus Christi (Andrén et al., 2015; Bird et al., 2019; Marshall et al., 2018). All tools were sanitized with ethanol and sterilized by flame, with particular care taken to avoid the outer edges of the sediment cores during sampling. To prevent any contamination of the samples, the research team employed face masks and hair nets. For each extraction, between 5 and 10 grams of sediment were used. The procedure outlined by the kit's manufacturer was adhered to, including the DNA concentration and resuspension steps, where DNA was dissolved in 100 µl of water suitable for molecular biology (Andrén et al., 2015; Bird et al., 2019; Marshall et al., 2018).

Along with the samples, blank controls without any sample material were also processed and sequenced to ensure no contamination; these controls showed no detectable DNA (below 0.5 ng) when assessed with the Qubit DS high-sensitivity kit and failed to produce any results in PCR amplification of the 16S rRNA gene, as evidenced in a 1% agarose gel. The sequencing of the metagenomes from samples M59E-0.25m and M59E-15m was carried out at the Marine Biological Laboratories in Woods Hole, MA, USA. The preparation of the metagenomic libraries and their sequencing were based on the Census of Deep Life guidelines, with specific adjustments made by Vineis et al. Notably, the sequencing was performed on the NextSeq platform by Illumina, based in San Diego, CA, USA, generating 150-bp-long paired-end reads, and the process did not include a step for microbiome enrichment (Andrén et al., 2015; Bird et al., 2019; Marshall et al., 2018). The libraries had insert sizes of 170 bp. The sequence data obtained from this sample has been archived in the Sequence Read Archive (SRA) under accession number SRS3221385, enabling further study and comparison within the scientific community (Andrén et al., 2015; Bird et al., 2019; Marshall et al., 2018).

2.3 Quality control

Quality control was performed using FASTQC on both samples. *(Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data, n.d.).* We assessed various aspects of the data, such as per base sequence quality, per sequence quality scores, per base sequence content, and per sequence GC content, by performing quality control using fastQC. The evaluation done by fastQC shows the satisfactory quality of the sequences

obtained, affirming the effectiveness of the sequencing effort and the reliability of the data. The trimming procedure was performed on both data sets using Trimmomatic, which performs a variety of useful trimming tasks for Illumina paired end and single ended data. The selection of trimming steps and their associated parameters are supplied on the command line (*KneadData – The Huttenhower Lab, n.d.*).

The current trimming steps performed are:

1. ILLUMINACLIP: Cut adapter and other Illumina-specific sequences from the read.
2. SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
3. LEADING: Cut bases off the start of a read, if below a threshold quality
4. TRAILING: Cut bases off the end of a read, if below a threshold quality
5. CROP: Cut the reads to a specified length
6. HEADCROP: Cut the specified number of bases from the start of the read
7. MINLEN: Drop the read if it is below a specified length
8. TOPHRED64: Convert quality scores to Phred-64

2.4 Identification and removal of contaminant sequences

The decontamination of sequencing data was conducted using Bowtie 2 (v2.4.2), a fast and memory-efficient tool for aligning sequencing reads to long reference sequences. The reference sequences used for decontamination included the human genome (GRCh38), obtained from the National Center for Biotechnology Information (NCBI) database, to identify and remove contaminant reads. As a result of this step, approximately 57.17 percent of reads were removed as contaminated. Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters and particularly good at aligning to relatively long (e.g. mammalian) genomes. We used the human genome as a database to decontaminate reads (Langmead et al., 2019).

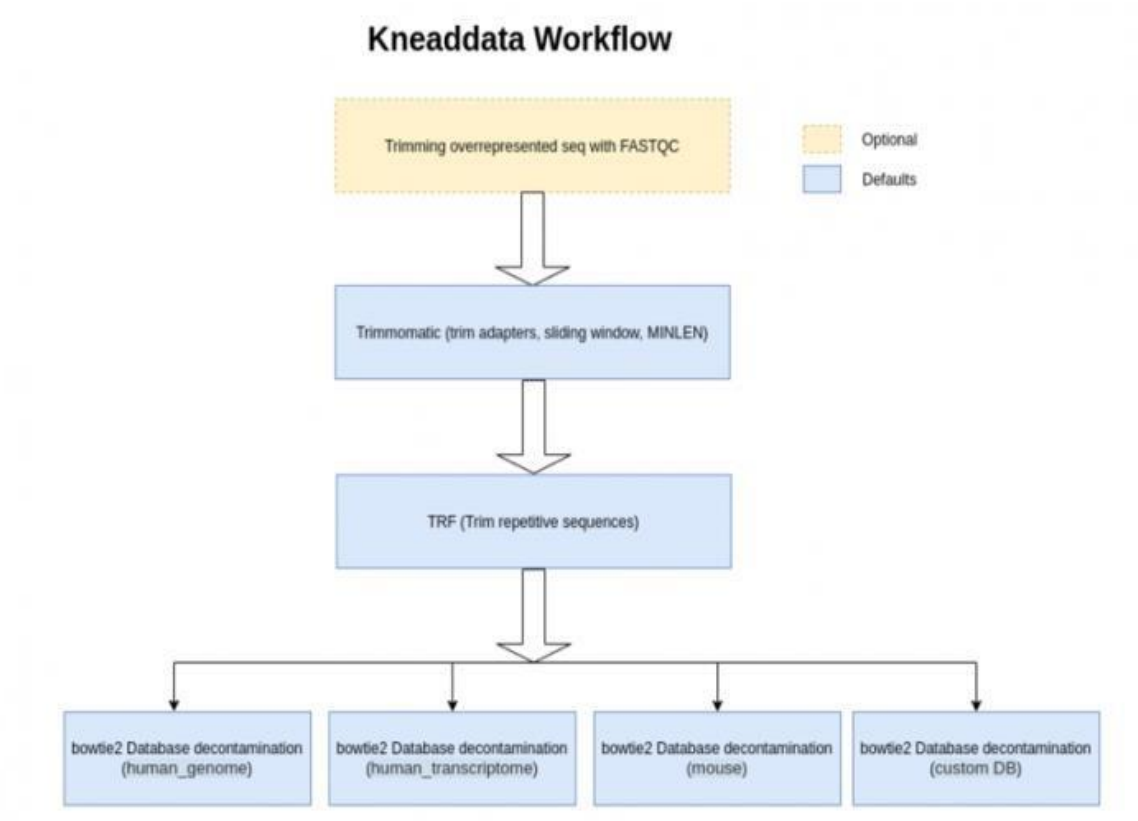

Figure 3 Knead Data is a computational pipeline that carries out a series of quality control and data preprocessing operations on raw DNA sequencing data. The primary steps of the pipeline involve utilizing Trimmomatic, Tandem Repeat Finder (TRF) and Bowtie2 to conduct quality control, adapter trimming, host genome removal, and contaminant removal.

2.5 Assembly

In the early stages of research, several Linux-based software programs were used, including de novo and reference-based assemblers. For the first try, Meta Velvet was selected as the preferred software (Liang & Sakakibara, 2021). Meta velvet is an extension of the Velvet assembler designed explicitly for de novo metagenome assembly from short sequence reads (Afiahayati et al., 2015). Later, during the literature review, MetaSpades was substituted with Meta velvet (Forouzan et al. 2018). MetaSPAdes is a metagenomic assembler that addresses various challenges of metagenomic assembly by capitalizing on computational ideas that have proven to be helpful in assemblies of single cells and highly polymorphic diploid genomes. It is designed to handle the difficulties associated with metagenomic data, which often contain complex bacterial populations composed of multiple related strains. MetaSPAdes implements several novel features, such as efficient assembly graph processing to address micro-diversity challenges, a new repeat resolution approach that utilizes rare strain variants to improve consensus assembly of strain mixtures, and fast algorithms for constructing assembly graphs and error-correcting reads. It has been benchmarked against other state-of-the-art metagenome assemblers and has demonstrated high-quality assemblies across diverse data sets.

2.6 Identifying secreted proteins

SignalP was used to identify proteins with signal peptides, which are targeted to the secretory pathway. This tool can predict the presence of signal peptides and the location of their cleavage sites in proteins from Archaea, Gram-positive bacteria, Gram-negative bacteria, and Eukarya. It is also able to distinguish between three types of signal peptides in bacteria and Archaea: Sec/SPI (standard signal peptide), Sec/SPII (lipoprotein signal peptide), and Tat/SPI (tat signal peptide). SignalP 6.0 is based on a neural network architecture that combines deep convolutional and recurrent layers, as well as a conditional random field (Teufel et al., 2022).

2.7 Annotation

We used DRAM (Distilled and Refined Annotation of Metabolism) to annotate the genes. DRAM is a framework to translate the deluge of microbiome-based genomic information into a catalogue of microbial traits (Shaffer et al., 2020). Microbial genome annotation involves recognizing the structural and operational components within DNA sequences and associating biological details with these components. DRAM, a software tool, is designed specifically for annotating genomes of bacteria, archaea, and viruses obtained from either pure cultures or metagenomes. Unlike conventional annotation methods, DRAM stands out for its ability to condense numerous gene annotations into comprehensive summaries of functional capabilities

at the genome level (Shaffer et al., 2023). We performed DRAM on all assembled genes and made a database of genes and their annotated functions.

2.8 Amino acid sequence alignments between deep and surface metagenomes

DNA sequences, translated *in silico* to amino acid sequences, from deep samples, were aligned with their closest homologs in surface samples, using the diamond algorithm, resulting in a total of 2,638 matches. We considered homologous pairs to be those with sequence identity above than 70 percent and matching with an e-value of smaller than 10^{-5}.

2.9 Protein Structure Prediction

We used AlphaFold to predict the protein structures of both our protein sets. AlphaFold is a machine learning system developed by DeepMind that can predict the 3D structure of proteins from their amino acid sequences (Jumper et al. 2021). It uses a combination of neural networks and physical simulations to make these predictions and has achieved state-of-the-art performance on a number of benchmark datasets. The accuracy of AlphaFold's predictions has significant implications for protein structure predictions, which are prerequisites for calculating protein stabilities (Jumper et al. 2021). In computational studies, researchers often start with a crystal structure or a model of a protein in a specific conformation. This initial structure is usually energetically minimized or equilibrated to obtain a more "relaxed" structure that represents a stable state or a local minimum on the energy landscape. This is done to study the dynamics and behaviors of the protein under physiological conditions.

Among Alphafold predictions, we used the primary relaxed protein structure, which reflects a prediction of a stable conformation that the protein might adopt under physiological conditions. This is crucial for realistic simulations and predictions of stability and molecular dynamics. AMBER was applied to all structures using the Colab service. AMBER is a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules (Case et al., 2005).

I have selected 135 sequences for initiating structure prediction. The previous computational procedure was resource-intensive and executed on a Local computer (Marie), so all the computations were carried out on Google Colab (Mirdita et al., 2022). The associated code is now accessible on GitHub (https://github.com/KambizKalhor/proteostab-Proteins-of-the-deep).

2.10 Protein stability prediction

FoldX is a tool that quickly and quantitatively estimates the free energy of folding of proteins from the energies and entropies of contributions of various interactions between protein residues (Schymkowitz et al., 2005). The resulting free energy of the folding function is based on empirical data obtained from protein engineering experiments, and it is designed to use minimal computational resources while still being accurate and fast. This makes it useful in protein design algorithms and in predicting protein structure and folding pathways (Schymkowitz et al., 2005). FoldX, an algorithm for predicting protein stability and mutation sites, has shown varying degrees of accuracy across different studies. When applied to predict and rationalize protein mutant stabilities, FoldX's extended protocol demonstrates good agreement with experimental stabilities (Buß et al., 2018).

I used FoldX to calculate the free energy of folding each pair of enzymes, for which the structure was predicted via AlphaFold. FoldX requires specific environmental parameters in order to estimate the free energy of folding, including temperature, pH, and ionic strength, which were supplied from IODP Expedition 347 data (Andrén, T., Jørgensen, B.B., Cotterill, 2015).

2.11 Statistical Assumption Checks

To ensure the robustness of our inferential statistics, we performed hecks of statistical assumptions before using t-tests, which compared differences between two independent protein pairs. We evaluated the normality assumption, crucial for the t-test, using the Shapiro-Wilk test for both surface and deep sample data regarding total stability. Additionally, we applied the Anderson-Darling normality test to complement it.

2.12 Homogeneity of Variances and Outlier Detection

Levene's Test for Homogeneity of Variance, with the median as the central tendency measure, was performed to confirm the homogeneity of variances across samples. For outlier detection, Grubbs's test was performed, followed by the Z-score method to flag outliers, prompting considerations for data transformation or robust statistical techniques.

2.13 Statistical Analysis Approach

The Paired Samples t-test was employed to compare means, given the paired nature of our data. Robustness checks were implemented to mitigate the influence of outliers. Additionally, the non-parametric Wilcoxon signed-rank test was performed as a counterpart to the t-test. The goal was to compare the mean stabilities of protein families between two surface and deep samples. Independent samples paired t-tests were conducted for each CAZy Family to detect

significant location-based differences. The same test was also applied based on the existence of signal peptides. All analyses and figure generation were performed using R statistical software, with significance set at $p < 0.05$.

CHAPTER 3 RESULTS

3.1 Quality of Sequencing Data

Quality control analysis was conducted on the sequencing data to ensure the integrity and reliability of the reads for subsequent analysis. The "per base sequence quality" plot (Figure 4) indicated a high-quality threshold across the dataset, with initial base calls exceeding a quality score of 29 and an expected decrease in quality towards the end of reads. The interquartile range of scores and the absence of significant anomalies attest to the overall high quality of the sequencing efforts.

Moreover, the "Per sequence quality scores" plot (Figure 5) exhibited a desirable distribution, with the majority of reads showing an average quality score above 30 and no notable frequency of lower-quality scores. This distribution indicates that the sequencing data are predominantly of high quality, with minimal noise introduced by poor-quality reads.

Lastly, the "Per sequence GC content" plot (Figure 6) displayed a normal distribution of GC content among the sequences, consistent with the expected genomic content of oceanic bacterial communities. The lack of sharp peaks in the GC content distribution suggests an absence of biased representation of sequences or contamination from extraneous sources.

The sequencing data passed the quality control metrics with satisfactory scores, demonstrating that the data are robust and suitable for further analysis. The consistency and reliability of the sequence quality endorse the effectiveness of the sequencing process and allow us to proceed with confidence to the next phase of our research.

3.2 Sequence decontamination

Subsequent to the implementation of quality control measures and decontamination using Kneaddata, a notable reduction in read counts was observed. Specifically, the read count for the deep sample was reduced from an initial 58,587,498 to 33,394,877. Similarly, the surface sample experienced a decrease from 98,478,162 to 51,162,989 reads. This reduction is indicative of the removal of low-quality reads and potential contaminants, thereby enriching the dataset for high-quality sequences suitable for our analysis.

Following the assembly of reads with metaSPAdes, gene prediction was carried out using MetaGeneMark. The analysis identified 7,915 genes within the deep sample and 28,154 genes in the surface sample. This differential gene count shows the variability and abundance of the microbial communities present in the two distinct marine environments.

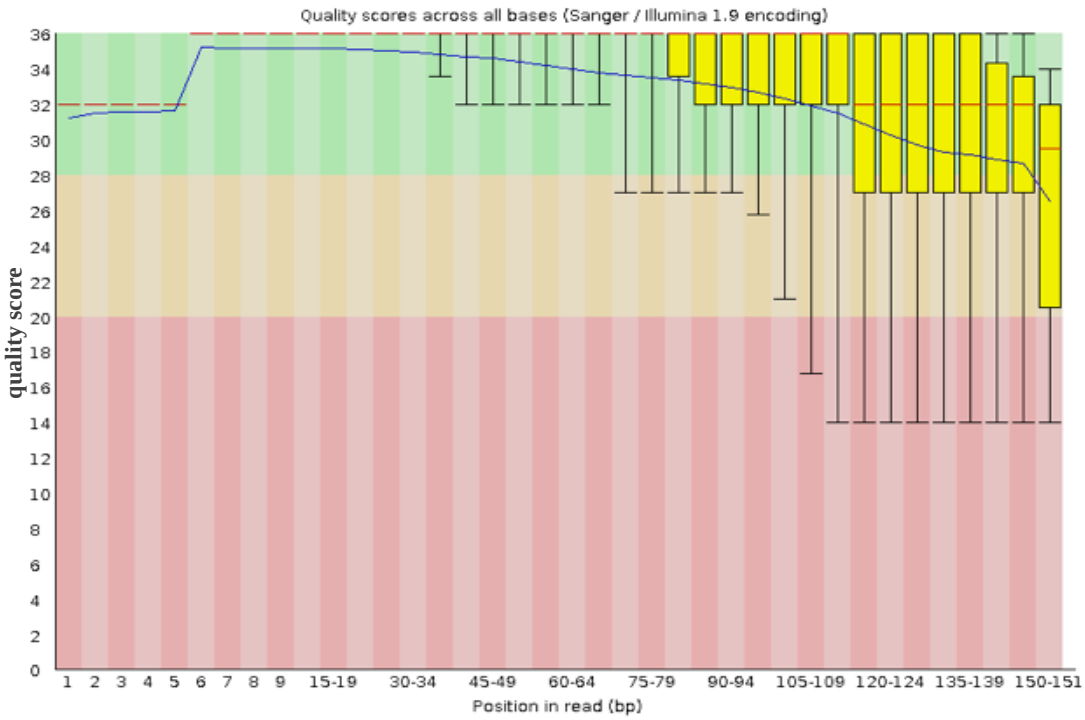

Figure 4 Distribution of quality score at each read position. The y-axis quantifies the quality scores, while the x-axis indexes the position within each read. Scores within the green band are considered high quality; yellow and red represent medium and low quality, respectively. The yellow box represents the interquartile range, with a red line representing the median value. The whiskers span the 10th to 90th percentiles, with a blue line marking the average score for the position.

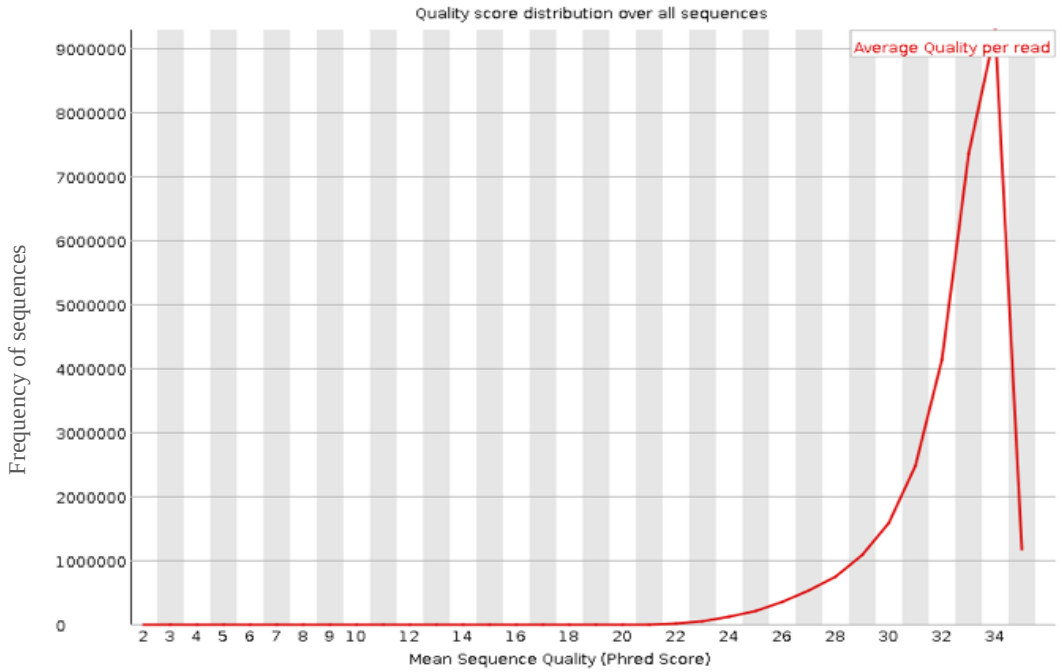

Figure 5 The "Per sequence quality scores" plot displays the average quality score along the x-axis against the frequency of sequences possessing that average score on the y-axis. It is desirable for the bulk of the reads to exhibit a high average quality score, ideally without any significant spikes in the lower quality range. Our data do not contain any small bumps at lower quality scores, and most quality scores are above 30, indicating that the sequence quality is satisfactory.

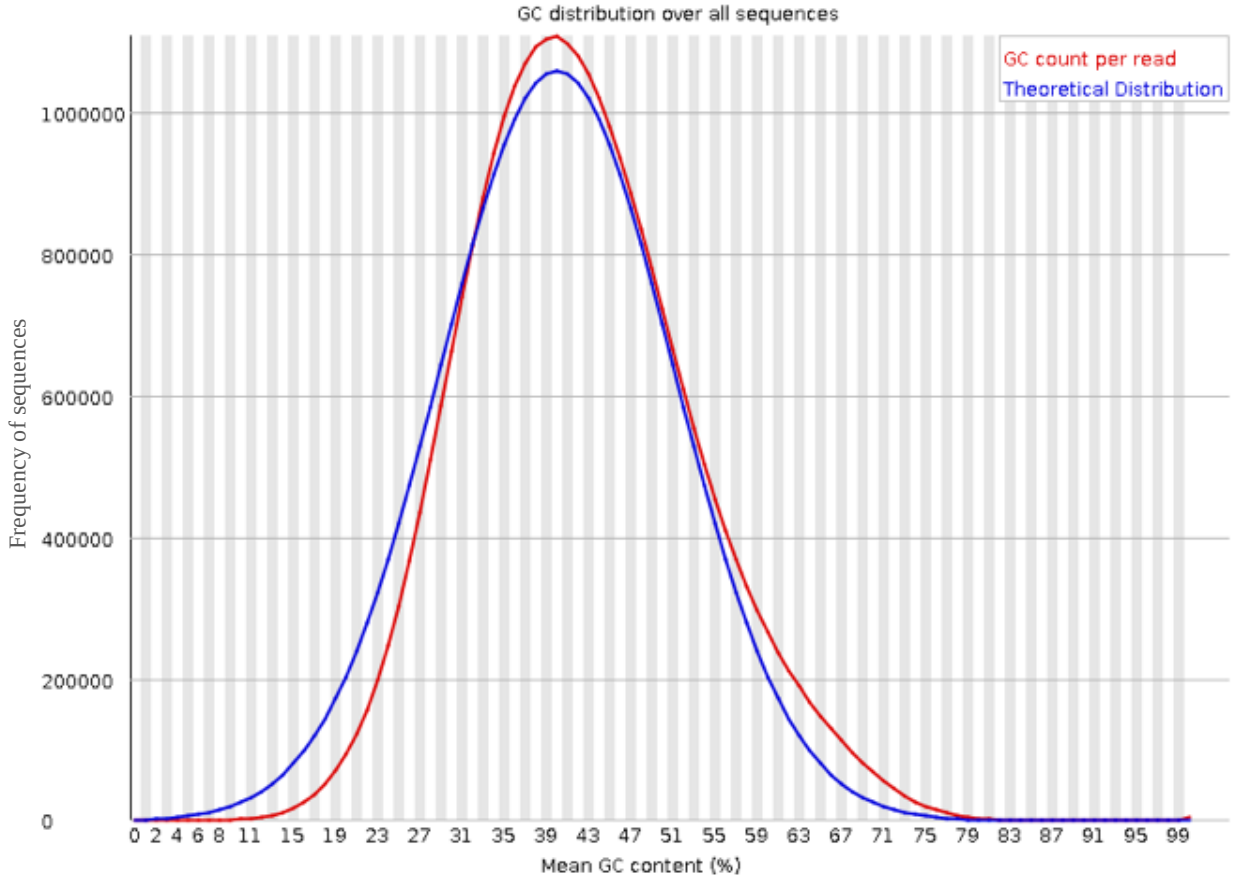

Figure 6 The "Per sequence GC content" plot illustrates the distribution of GC content across all sequences. It's beneficial to compare the GC content at the central peak with the anticipated GC percentage for the bacterial community of the ocean. Ideally, the distribution should appear normal, with deviations such as sharp peaks indicating over-represented sequences or a broad peak suggesting contamination by another organism. In the case of ocean bacteria, whose genomic G+C content ranges from under 30% to above 60%, this figure reflects similar variability. Additionally, the absence of sharp peaks and the presence of a normal distribution are observed.

3.3 Protein annotation

Our analysis of the annotated genes reveals insightful distributions pertinent to their potential functions and classification confidence. Signal peptide analysis via SignalP version 6 (Figure 7) shows that only a minor fraction (15.7%) of the annotated genes possesses signal peptides. These data suggest that a relatively small subset of the annotated genes is equipped with signal peptides, which are short peptides present at the N-terminus of newly synthesized proteins that are destined towards the secretory pathway, including five types of bacterial signal peptides: Sec/SPI, Tat/SPI, Lipoprotein, Sec/SPII, and Tat/SPII. The DRAM annotation (Figure 8) provides a confidence level for gene annotations, ranking them based on reciprocal best hits across multiple databases. In DRAM annotation of genes, each gene is assigned a summary rank (from A to E) based on reciprocal best hits (RBH) from multiple databases like KEGG, Uniref90, Pfam, CAZy, MEROPS, and VOGDB. The ranks indicate the confidence level of the annotation, with A being the highest confidence and E being the lowest. These ranks help assess the reliability of the annotations provided by DRAM for different genes. The results show that 34.6 percent of genes are hit to KEGG genes with a bit score >60 (C rank) and UniRef90 with a bit score >60 (C rank). 35.3 percent of genes had Pfam, dbCAN or MEROPS matches (D rank), but hits to KEGG or UniRef90 were <60-bit score. 30.1 percent are ranked E, which represents proteins that had no significant hits to any DRAM database, including KEGG, UniRef90, dbCAN, Pfam, MEROPS, or only had significant hits to VOGDB.

The analysis of the annotated genes revealed a diverse composition of carbohydrate-active enzyme families. A total of 135 genes encoding CAZymes were identified, which were classified into five distinct families based on their functional domains.

The Glycoside Hydrolase (GH) family was the most prevalent, with 88 genes representing 65% of the total CAZyme-encoding genes identified. This family is known for its role in the hydrolysis of glycosidic bonds, suggesting a significant capacity for Glycoside degradation within the microbial community studied. The glycosyl transferase (GT) family, which catalyzes the transfer of sugar moieties from activated donors to specific acceptor molecules to form glycosidic bonds (Breton et al., 2006), comprised 21 genes, accounting for 15.5% of the CAZyme gene pool. Polysaccharide Lyases (PL) that act as an enzyme which claves certain activated glycosidic linkages present in acidic polysaccharides (Linhardt et al., 1986), were represented by 15 genes, making up 11% of the CAZyme genes. This family's enzymes are typically involved in the degradation of complex and highly structured polysaccharides such as pectin and chitin. The Carbohydrate-Binding Module (CBM) family was represented by ten

genes, corresponding to 7.5% of the identified CAZyme genes, which are protein domains found in carbohydrate-active enzymes that bind specifically to polysaccharides (Boraston et al., 2004). Lastly, the Carbohydrate Esterase (CE) family, which is a group of enzymes that catalyze the de-O or de-N-acylation of carbohydrates by removing ester decorations, was the least represented with a single gene, accounting for 1% of the CAZyme genes (Nakamura et al., 2017).

Within the glycoside hydrolases, the most abundant class was GH2(13%), which plays crucial roles in the breakdown and synthesis of complex carbohydrates (Schröder et al., 2015), including the metabolism of xylan, oligo and disaccharides, resistant starch, and glycogen.

Another abundant subfamily was glycoside hydrolase family 74 (GH74, 7.5%), which are endo-β-glucanases that play a crucial role in breaking down xyloglucans (XyG), which are complex carbohydrates found in plant cell walls.

In the Glycosyl transferase Family, the most abundant subfamily was Glycosyl transferase family 35 (GT35, 2.2%) enzymes, which play a crucial role in glycosylation processes. These enzymes are responsible for catalyzing the formation of glycosidic bonds between sugars, contributing to the biosynthesis of complex carbohydrates, glycoproteins, and glycolipids.

The most abundant Polysaccharide lyase was the PL22 subfamily, with 3% abundance, and it is involved in the degradation of disaccharides, specifically oligogalacturonate, through a β-elimination mechanism.

3.4 Similarity of deep subsurface proteins to surface homologs

In our study, to identify homologous protein pairs from both samples, we conducted BLAST searches, which resulted in the identification of 2,638 pairs characterized by sequence identity, bit-score, and E-values, as detailed in Figure 9. Subsequently, we refined our selection to those BLAST hits where both proteins were annotated as Carbohydrate-Active enzymes (CAZy).

3.5 Stability conditions

To effectively analyze protein stability, it is important to understand in which conditions the stability should be evaluated. Extracellular proteins in surface versus deep sediments experience somewhat different geochemical conditions. One could reasonably compare the stability of proteins in their native environments, which would provide a measure of the actual stability of the enzymes, or in surface conditions only since species in deep sediments derive

Table 2 DRAM databases

Carbohydrate-active enzymes family	number of genes	relative abundance	Subcategory	Relative abundance	Number of genes
Glycoside Hydrolase Family	88	65%	GH2	13%	18
			GH74	7.5%	10
			GH38	5%	7
			GH78	4%	6
			GH127	3.7%	5
			GH94	3%	4
			GH106	3%	4
			GH32	2.2%	3
			Other	23%	31
Glycosyltransferase Family	21	15.5%	GT35	2.2%	3
			GT66	1.5%	2
			GT27	1.5%	2
			GT2	1.5%	2
			Other	9%	12
Polysaccharide Lyase Family	15	11%	PL22	3%	4
			PL9; PL31	3%	4
			PL9	2.2%	3
			PL31	2.2%	3
			Other	0.7%	1
Carbohydrate-Binding Module Family	10	7.5%	CBM40	2.2%	3
			CBM57	2.2%	3
			Other	3%	4
Carbohydrate Esterase Family	1	1%	CE1	0.7%	1
Total	*137*	*100%*			

Table 3 Distribution of Carbohydrate-Active enzyme (CAZy) families within the collection of annotated genes.

Database	KEGG	Uniref90	Pfam	CAZy	MEROPS	VOGDB
Description	Kyoto Encyclopedia of Genes and Genomes	clustered sets of sequences from UniProtKB and selected UniParc records that have at least 90% sequence identity and 80% overlap.	a comprehensive database of protein families and domains	The Carbohydrate-Active Enzymes (CAZy) database	proteases, proteinases and proteolytic enzymes	Virus Orthologous Groups Database
Reference	Kanehisa & Goto, 2000	Suzek et al., 2007	Mistry et al., 2021	Lombard et al., 2014	Rawlings et al., 2018	Ritsch et al., 2023

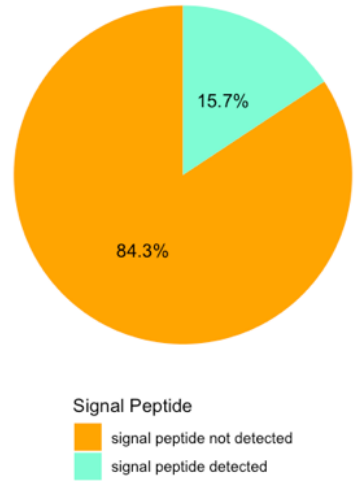

Figure 7 Distribution of signal peptides in annotated genes. The pie chart illustrates the proportion of annotated genes possessing signal peptides prior to any filtration process, as predicted by SignalP version 6. A majority of the genes, 84.3%, are detected as sequences with no signal peptide, while in 15.7% of annotated genes, a signal peptide is predicted.

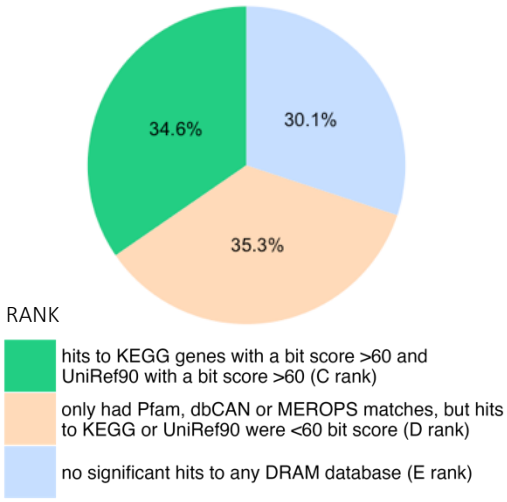

Figure 8. Confidence level of the annotation in DRAM. For each gene annotated, DRAM provides a single, summary rank (A–E), which represents the confidence of the annotation. The highest rank includes reciprocal best hits (RBH) with a bit score >350, against KEGG genes (A rank), followed by reciprocal best hits to UniRef90 with a bit score >350 (B rank), hits to KEGG genes with a bit score >60 (C rank) and UniRef90 with a bit score >60 (C rank). The next rank represents proteins that only had Pfam, dbCAN or MEROPS matches (D rank), but hits to KEGG or UniRef90 were <60-bit score. The lowest rank (E) represents proteins that had no significant hits to any DRAM database including KEGG, UniRef90, dbCAN, Pfam, MEROPS, or only had significant hits to VOGDB.

from surface sediments (Chen et al., 2017) or both in deep conditions (in order to get an apples-to-apples comparison of the stability of deep enzymes, the focus of this work, with proteins in surface conditions).In order to assess the importance of the decision, I assessed how ionic strength, pH and temperature affected the stability of some of these proteins.

To do this, I evaluate the stability of a protein pair under identical conditions and then replicate this analysis across hundreds of different conditions, with PHs ranging from 6 to 10 and ionic strength ranging from 0 to 1. If the disparity in stability remains constant across all conditions, it suggests that the stability difference is purely based on structure and unaffected by varying conditions. However, there are other possibilities to consider. One scenario might involve fluctuating stability differences, where in some conditions, the protein located in the deep sample is more stable, and in other conditions, the surface proteins exhibit greater stability. Another scenario could involve both proteins experiencing a shift in stability in the same direction but with one protein consistently maintaining higher stability. This would result in either an increase or decrease in the gap between their stability levels. All three scenarios mentioned have been seen in our analysis.

3.6 Protein stability

The The stability of all 135 CAZyme proteins was assessed using FoldX. However, there were insufficient data pairs in the Carbohydrate Esterase Family (CE), Auxiliary Activities (AAs), and Cohesins. Furthermore, two protein sequences with atypical lengths were identified and subsequently excluded from the analysis. Figure 10 presents a boxplot illustrating the distribution of stability among Carbohydrate-Active Enzyme (CAZy) families, categorized by location and the presence of signal peptides. It was observed that the stability levels of Polysaccharide Lyases (PLs) surpassed those of all other families. Additionally, boxplots for certain CAZy families are not displayed due to the absence of adequate protein pair data.

The following plots provide an overview of how the presence of a signal peptide influences the free energy of folding within distinct enzymatic families, suggesting an interplay between peptide signalling, protein family and sample location with enzyme stability. In the analysis depicted in Figure 11, the free energy of folding in two enzyme families is interrogated under varying experiments (comparisons). The upper left box plot elucidates the distribution of free energy changes within the Glycoside Hydrolases family that lack a signal peptide. The upper right counterpart details the same for Glycoside Hydrolases with the presence of a signal peptide.

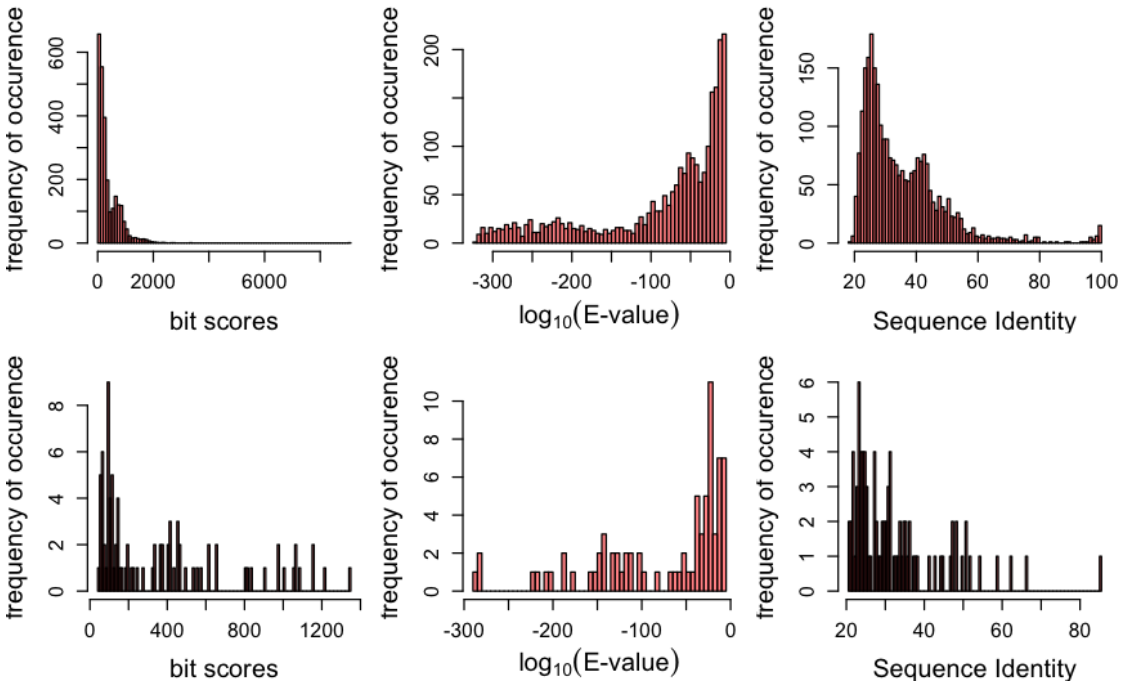

Figure 9 Histograms depicting the distribution of Bit Scores, E-value, and sequence identity of BLAST Hits. The upper panel shows the distribution of the entire dataset, while the lower panel displays the distributions after the exclusion of specific pairs that are not of interest to the study. A. The first histogram displays the distribution of bit scores from a BLAST search, with the frequency of occurrence on the y-axis and bit scores on the x-axis. B. The second histogram illustrates the distribution of E-values on a logarithmic scale (Log10[E-values]) from the same BLAST search, with the frequency of hits on the y-axis. This graph shows a right-skewed distribution, with a higher frequency of higher E-values, but still, most of the E-values are significant. C. The third histogram shows the distribution of sequence identity percentages for the BLAST hits, with sequence identity on the x-axis and frequency on the y-axis. The distribution peaks at lower sequence identity values, suggesting that most hits have a lower percentage of sequence identity.

Similarly, the lower plots offer a comparative visualization for the Glycosyl Transferases family. The lower left plot presents data for those enzymes without signal peptides, while the lower right plot contrasts those proteins with the signal peptide (no protein set was found under this criterion).

Each sub-plot compares the stability of pairs of homologous proteins in our deep and surface samples. The vertical axis represents the free energy of folding, with lower values indicating greater stability. Each dot represents a protein, and each line connects a pair of homologous proteins.

Figures 11, 12, and 13 illustrate the results of the same experiments conducted under different free energy folding calculations. Figure 11 replicates the environmental conditions of Sample 1, sourced from deep-sea sediment, characterized by a pH of 7.58, an ionic strength of 0.66, and a temperature of 281.15 Kelvin. Figure 12 presents outcomes from free energy of folding calculations performed under surface environmental conditions, which differ notably with a pH of 7.88, an ionic strength of 0.49, and a maintained temperature of 281.15 Kelvin. The final figure, Figure 13, showcases free energy calculations executed under conditions identical to those of their respective sample origins (all conditions are mentioned in the figures).

A series of independent paired t-tests were conducted to evaluate the mean ΔG_{fold} values of homologous carbohydrate-active enzymes sourced from both surface and deep samples. This comparative analysis was executed across three distinct enzyme categories: the Glycoside Hydrolase (GH) family with a signal peptide, the Glycoside Hydrolase (GH) family without a signal peptide, and the Glycosyl Transferase (GT) family, also lacking a signal peptide. Among these, a statistically significant mean difference in ΔG_{fold} was identified solely within the GH family pairs lacking a signal peptide. For this subgroup, the surface samples exhibited a ΔG_{fold} mean that was significantly higher than that of the deep samples, with a t-value of 3.4181 (df=12) and a p-value of 0.005096. The difference in means was 81.666, with a 95% confidence interval extending from 29.61 to 133.72, denoting the higher stability of the surface enzymes relative to those from the deep samples.

In Figure 14, we evaluated the stability of 135 protein pairs from both deep and surface samples, calculating their stability under conditions retrieved from their respective environments. We used a comparative boxplot to show the distribution of the free energy of folding (ΔGfold, expressed in kJ·mol-1) for these proteins, categorizing them by their location

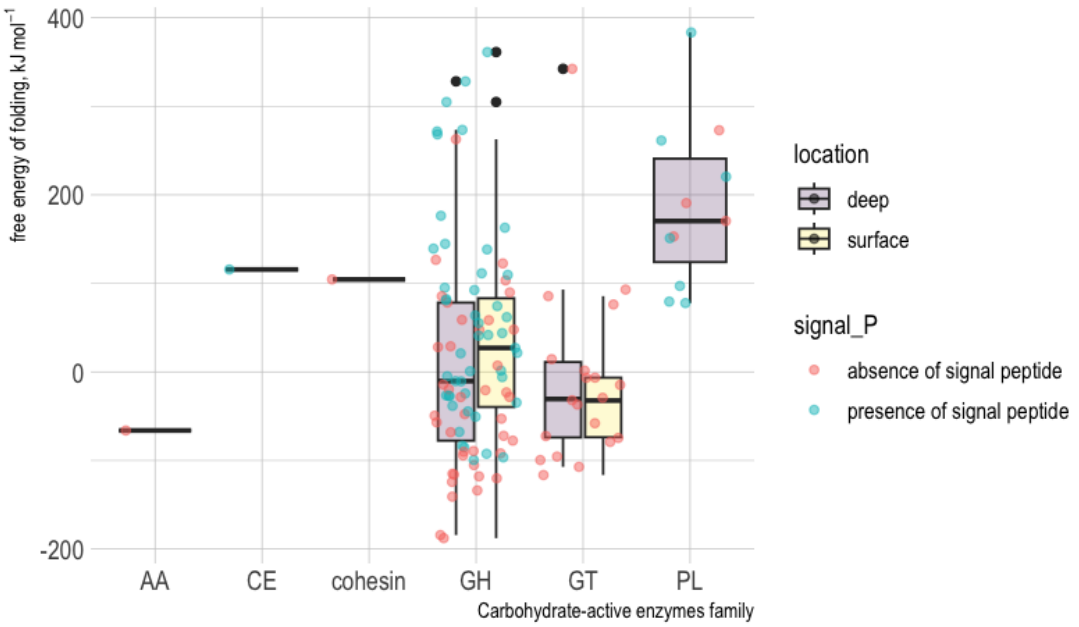

Figure 10 Boxplot of Carbohydrate-Active Enzyme (CAZy) family distribution by location and signal peptide presence. The figure presents a comparative analysis of the total stability for different CAZy families, specifically AA (Auxiliary Activities), CE (Carbohydrate Esterases), cohesin modules, GH (Glycoside Hydrolases), and GT (Glycosyl Transferases), along with PL (Polysaccharide Lyases), categorized by the location of the sample (deep or surface) and the presence of signal peptides (signal peptide exist or no signal peptide). Boxplots represent the interquartile range of stability values with medians, and individual points indicate a protein stability value. The points are colored according to the presence of a signal peptide, with red indicating absence of signal peptide and blue indicating s presence of signal peptide.

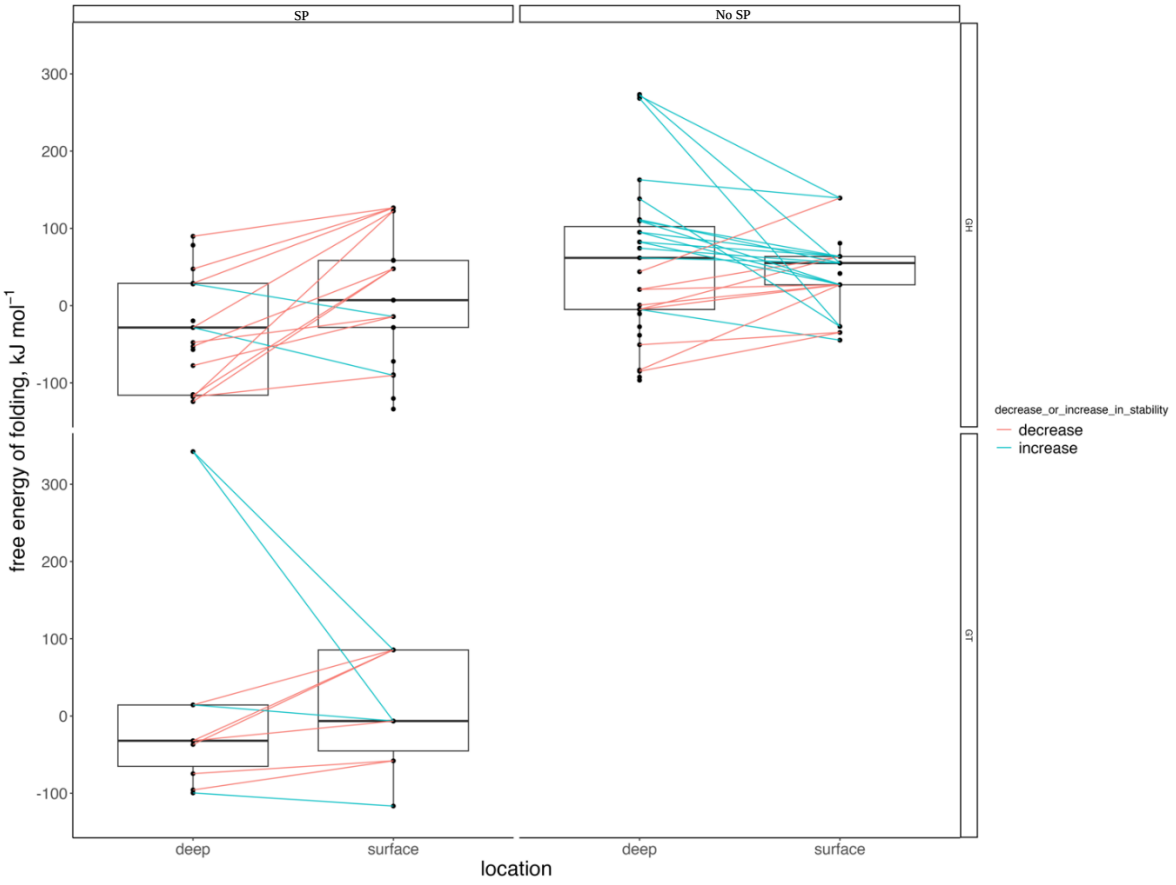

Figure 11 Comparative boxplot of free energy of folding by location, signal peptide prediction, and the enzyme family. This graph displays four sets of Boxplots that compare the free energy of folding (ΔG_{fold}, kJ.mol^{-1}) for proteins based on their location (deep vs. surface) in four different subsets. The top left is Glycoside Hydrolases without signal peptide. On the top right are Glycoside Hydrolases with signal peptides. The bottom left is Glycosyl Transferases without signal peptide.
The bottom right is Glycosyl Transferases with signal peptide. (No protein set was found with these criteria.
The stability of 135 pairs of proteins in our deep and surface sample was measured using surface conditions (ionic strength = 0.66, pH = 7.58, Temperature = 281.15 Kelvin). The vertical axis is the free energy of folding. More negative free energy means more stability. Each dot presents a protein, and each line connects a pair of homologue proteins.

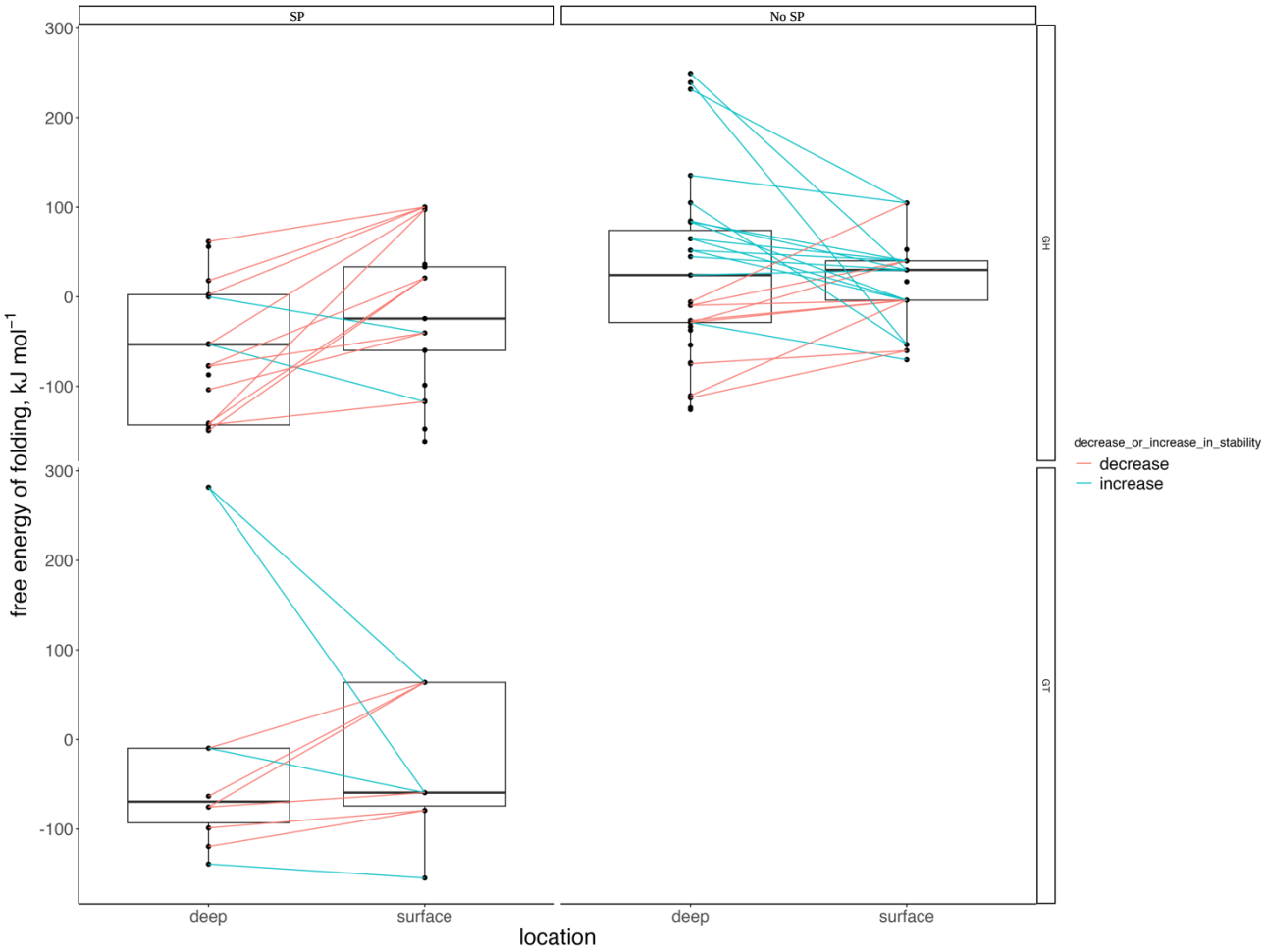

Figure 12 Comparative boxplot of free energy of folding by location, signal peptide prediction, and the enzyme family. This graph displays four sets of boxplots that compare the free energy of folding (ΔG_{fold}, kJ.mol^{-1}) for proteins based on their location (deep vs. surface) in four different subsets. The left column indicates enzymes with predicted signal peptide (SP) while the right column indicates enzymes lacking a predicted signal peptide (No SP). The top row indicates annotate glycosyl hydrolases (GH) while the bottom row indicates annotated glycosyl transferases (GT). Stability of 135 pairs of proteins in our deep and surface sample was measured using deep conditions(ionic strength = 0.49, PH = 7.88, Temperature = 281.15)
The vertical axis is the free energy of folding. More negative free energy means more stability. Each dot presents a protein, and each line connects a pair of homologue proteins.

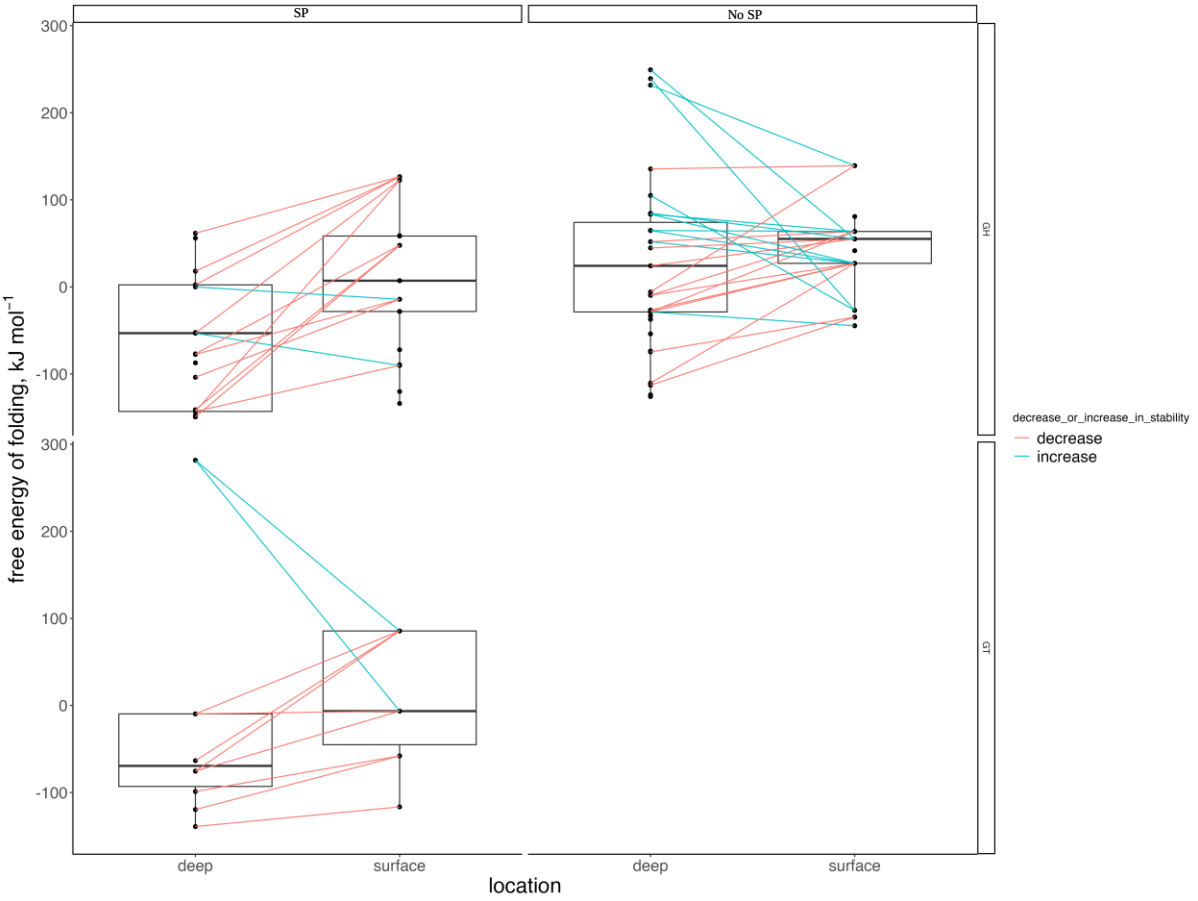

Figure 13 Comparative boxplot of free energy of folding by location, signal peptide prediction, and the enzyme family. This graph displays four sets of boxplots that compare the free energy of folding (ΔG_{fold}, kJ.mol^{-1}) for proteins based on their location (deep vs. surface) in four different subsets. The left column indicates enzymes with predicted signal peptide (SP) while the right column indicates enzymes lacking a predicted signal peptide (No SP). The top row indicates annotate glycosyl hydrolases (GH) while the bottom row indicates annotated glycosyl transferases (GT). Stability of 135 pairs of proteins in our deep and surface sample was measured using the conditions sample retrieved from. The vertical axis is the free energy of folding. More negative free energy means more stability. Each dot presents a protein, and each line connects a pair of homologue proteins.

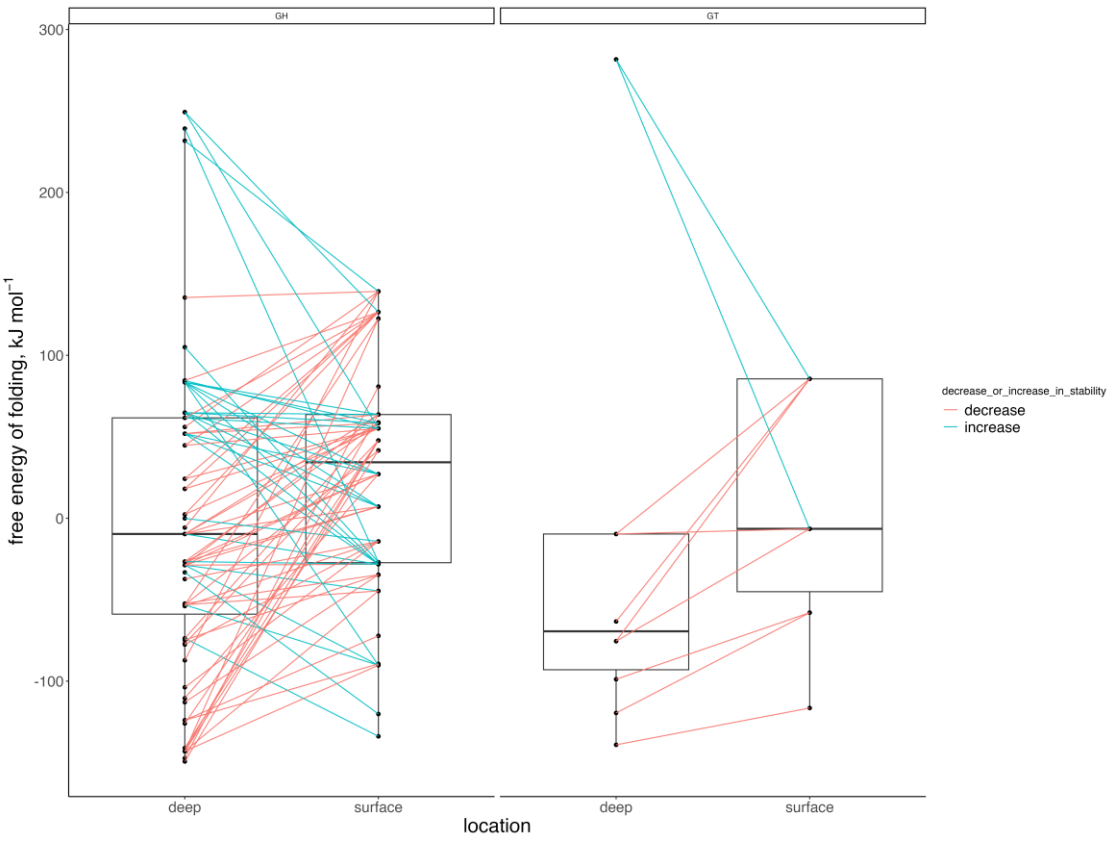

Figure 14 Comparative boxplot of free energy of folding by location and the enzyme family. This graph displays two sets of boxplots that compare the free energy of folding (ΔG_{fold}, kJ.mol^{-1}) for proteins based on their location (deep vs. surface) in two different protein families (GT and GH).
Stability of 135 pairs of proteins in our deep and surface sample was measured using the conditions sample retrieved from. The vertical axis is the free energy of folding. More negative free energy means more stability. Each dot presents a protein, and each line connects a pair of homologue proteins.

(either deep or surface) and protein family to which they belong, regardless of the presence or absence of signal peptide. The analysis revealed distinct patterns in the free energy of folding between proteins from different locations and based on the protein family. Specifically, among the Glycoside Hydrolase Family, 63.75% exhibited higher free energy of folding when located at the surface compared to those in the deep. independent samples paired t-test was performed to compare the means of ΔGfold and no significant difference been observed ($t(79) = 0.203$, $p = 0.839$). The mean ΔGfold of the surface sample was 2.248 (95% confidence interval ranging from -19.77 to 24.267), higher than the mean ΔGfold of the deep sample.

Similarly, for the Glycosyltransferase Family, 80% of those located at the surface demonstrated a higher free energy of folding than their deep counterparts. independent samples paired t-test was performed and there was no significant difference between ΔGfold in pairs of surface and deep proteins ($t(9) = -0.5127$, $p = 0.62$). The mean ΔGfold of surface sample was -25.24 (95% confidence interval ranging from -136.6 to 86.12) lower than the mean ΔGfold of deep sample.

For Glycosyl transferase (GT) family pairs, there was a no significant difference between ΔG_{fold} in pairs of surface and deep proteins ($t(9) = -0.5127$, $p = 0.62$). The mean ΔG_{fold} of surface sample was -25.24 (95% confidence interval ranging from -136.6 to 86.12) lower than the mean ΔG_{fold} of deep sample.

We evaluated the stability of 135 protein pairs from both deep and surface samples, calculating their stability under conditions retrieved from their respective environments (Figure 15). We used a comparative boxplot to show the distribution of the free energy of folding (ΔG_{fold}, expressed in $kJ \cdot mol^{-1}$) for these proteins, categorizing them by their location (either deep or surface) and the presence or absence of a signal peptide, regardless of the protein family to which they belong. The analysis revealed distinct patterns in the free energy of folding between proteins from different locations based on the presence of signal peptides. Specifically, among the protein pairs lacking a signal peptide, 73% exhibited a higher free energy of folding when located at the surface compared to those in the deep. independent samples paired t-test was performed and there was a significant difference between the means of ΔGfold on paired homologues carbohydrate-active enzymes of surface and deep sample ($t(20) = 4.03$, $p = 0.0006$). The mean ΔGfold of the surface sample was 67.36 (95% confidence interval ranging from 32.49 to 102.24), higher than the mean ΔGfold of the deep sample.

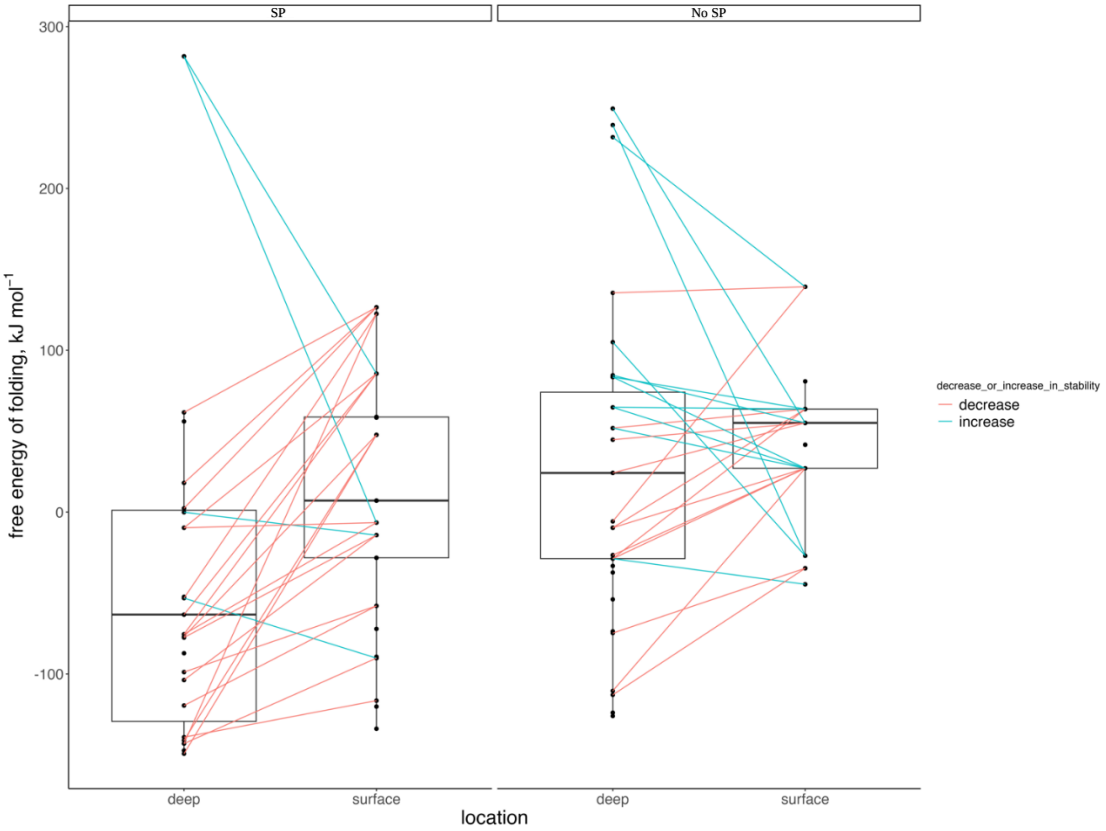

Figure 15 Comparative boxplot of free energy of folding by location and the presence of signal peptide. This graph displays two sets of boxplots that compare the free energy of folding (ΔG_{fold}, kj.mol^{-1}) for proteins based on their location (deep vs. surface) in proteins with and without signal peptide. Stability of 135 pairs of proteins in our deep and surface sample was measured using the conditions sample retrieved from. The vertical axis is the free energy of folding. More negative free energy means more stability. Each dot presents a protein, and each line connects a pair of homologous proteins.

Conversely, for proteins equipped with a signal peptide, 58% of those located at the surface demonstrated a higher free energy of folding than their deep counterparts and there was no significant difference between ΔGfold on paired homologues carbohydrate-active enzymes of surface and deep sample ($t(20) = 0.56$, $p = 0.58$). The mean ΔGfold of surface sample was 7.82 (95% confidence interval ranging from -36.98 to 21.32) lower than the mean ΔGfold of the deep sample. This differential folding energy underscores the potential influence of signal peptide presence (activity location of protein) on protein stability.

3.7 Protein size and Total Stability

We analyzed the protein folding energy relative to their amino acid sequence length to understand the influence of sequence length on protein folding energy. The scatter plot (Figure 16A) reveals differentiation in stability trends based on the protein sample's location, categorized as surface or deep. For the surface sample proteins, indicated by the blue line, there was a lack of a clear linear relationship. This shows that as the sequence length varies, the folding energy of these proteins remains relatively unchanged.

Conversely, proteins derived from deep samples exhibit a different trend. The red line in Figure 16A shows a more pronounced positive correlation, suggesting that as the protein sequence length increases, there is a corresponding increase in folding energy for these deep sample proteins. This trend was significant and indicates that longer sequence lengths may contribute positively to protein folding energy in the deep sample, with the effect being more substantial as compared to surface samples.

In the overall comparison of sequence length versus folding energy without differentiation by sample location (Figure 16B), the aggregated data suggest a general trend where longer sequence lengths could be associated with higher protein folding energy across the combined data set. This indicates that while local environment specifics play a significant role in stability, the sequence length is a contributing factor to the folding energy of proteins, with its effect being more pronounced in deep subseafloor environments. In the analysis of the relationship between total stability and sequence length within our dataset, Pearson’s product-moment correlation was employed. The results revealed a weak yet statistically significant, positive linear correlation with a coefficient of 0.27 ($p = 0.0029$). This indicates that as sequence length increases, there is a tendency for total stability to also increase, albeit modestly. The 95% confidence interval for this correlation ranges from 0.095 to 0.43, further substantiating the presence of a statistically significant relationship.

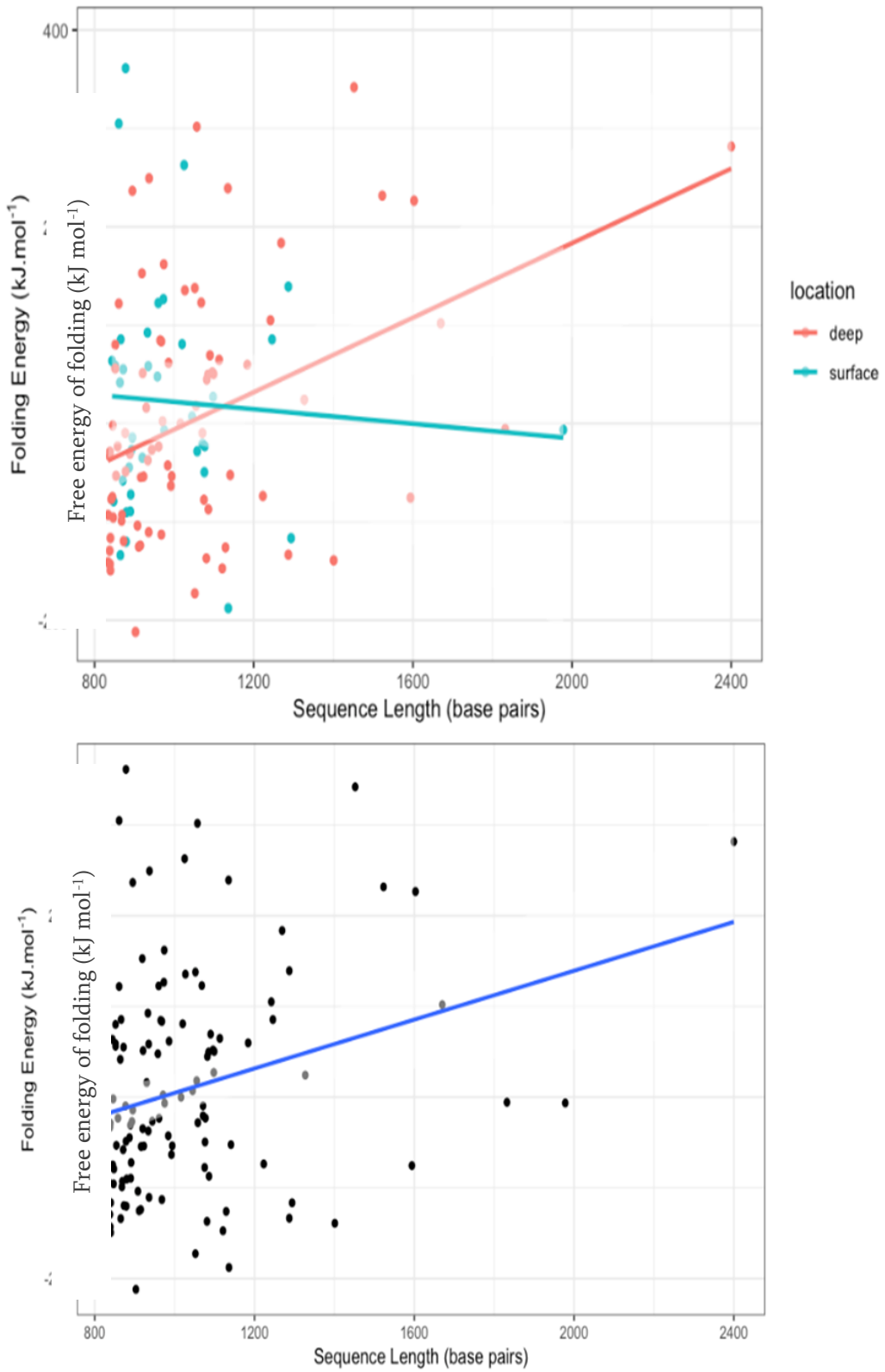

Figure 16 Relationship between protein size and total stability. The scatter plot displays the total stability of proteins as a function of their amino acid sequence length. Data points are differentiated by color in panel A, indicating the surface and deep sample. The blue line represents surface samples, and the red trend line represents proteins from deep sample.
In panel B, the plot combines both surface and deep sample data points without color differentiation, displaying an overall trend of sequence length versus folding energy.

3.8 Statistical Analysis

To ensure the validity of the inferential analyses, rigorous statistical assumption checks were performed prior to the execution of t-tests. These checks were integral to confirm the appropriateness of the t-test for our dataset, which compared means between two independent groups.

We addressed the assumption of normality, which is crucial for the t-test, especially in small sample sizes. Prior to the application of parametric statistical analyses, the normality of the data distribution was evaluated using the Shapiro-Wilk test. This test was conducted separately for the surface and deep sample data for the total stability. The test for the surface data yielded a W statistic of 0.93 with an associated p-value of 0.03, suggesting a departure from normality at the 5% significance level. On the other hand, the test for the deep data produced a W statistic of 0.91 with a p-value of 2.8×10^{-5}, further indicating a more pronounced deviation from normality.

To complement the Shapiro-Wilk test results and further assess the normality of our data distributions, we also applied the Anderson-Darling normality test to the total stability for both surface and deep data sets. For the surface data, the Anderson-Darling test yielded a statistic of A = 0.708 with a corresponding p-value of 0.059. This result indicates a marginal non-significance, suggesting that the distribution does not significantly deviate from normality at the conventional alpha level of 0.05.

In contrast, the deep data exhibited a higher Anderson-Darling statistic of A = 1.9, with a highly significant p-value of 4.6×10^{-5}, firmly suggesting that the assumption of normality is not met for the total stability within this dataset.

Taken together, the results from both the Shapiro-Wilk and Anderson-Darling tests indicate that while the surface sample data may approximate a normal distribution, the deep sample data demonstrate a significant deviation from normality. The significance of the Shapiro-Wilk test denotes that the data for Total stability do not follow a Gaussian distribution, and hence, caution should be exercised when considering parametric tests for these data. The non-normality of the data necessitated the consideration of alternative analytical approaches. Subsequently, non-parametric methods were employed to assess the differences in protein stability between the two environments.

In assessing the assumption of homogeneity of variances, which is critical for certain parametric tests, Levene's Test for Homogeneity of Variance was utilized. With the median serving as the measure of central tendency to mitigate the impact of any potential outliers, Levene's Test on the stability data of deep and surface samples returned an F-value of 1.169.

The corresponding p-value was 0.28, which exceeds the conventional alpha level of 0.05. This result suggests that there is no statistically significant difference in variances between the groups, thereby satisfying the assumption of homoscedasticity for the given dataset. With 1 degree of freedom for the groups and 122 degrees of freedom for the error term, the test confirms that variance across the groups is consistent, permitting further parametric analysis based on this assumption.

In our analysis, we applied Grubbs's test to identify and quantify the presence of outliers within our datasets. For the deep sample stabilities, Grubbs's test yielded a G statistic of 3.86 with an associated U statistic of 0.83 and a p-value of 0.002. This significant result led us to reject the null hypothesis, indicating that the highest value in the dataset, 681.715, is an outlier.

Similarly, for the surface sample, Grubbs's test reported a G statistic of 5.18, a U statistic of 0.36, and an extremely low p-value of 1.7×10^{-9}. This strongly suggests that the highest observed value of 2611.83 is an outlier with respect to the rest of the data distribution.

Upon application of the Z-score method, there were 4 data points exhibiting Z-scores greater than 3 or less than -3 that were flagged as outliers. This identification is rooted in the empirical rule that for a normally distributed dataset, approximately 99.7% of the data should fall within three standard deviations from the mean. Therefore, data points falling outside of this range are considered to possess a low probability of occurrence and, thus, are treated as potential outliers.

Given the identification of these outliers, we carefully considered their impact on the subsequent statistical analysis. The presence of such extreme values can distort the results and interpretations of parametric tests, which assume that data do not include outliers. Consequently, we evaluated the need for outlier mitigation strategies, such as data transformation or the adoption of robust statistical techniques, to ensure the integrity of our findings.

Given the paired nature of our data, alongside its homoscedasticity and normal distribution, a Paired Samples t-test was deemed appropriate for comparing the means of our two conditions. This choice is predicated on the t-test's sensitivity to differences in paired observations, making it well-suited for our data structure. However, to account for the potential influence of outliers, robustness checks were implemented, including outlier impact assessment and consideration of data transformation techniques. Also, the Wilcoxon signed-rank test is the non-parametric counterpart to the paired t-test that was performed. The primary objective of the statistical analysis was to compare the mean stability of different protein families between the two soil samples (deep vs surface). For each CAZy Family, an independent sample paired t-test was

conducted to determine if there were statistically significant differences between the two locations. Data were analyzed using statistical software ®, setting the level of significance at $p < 0.05$. All statistical analyses and the generation of figures were performed using R, a comprehensive statistical software package.

3.9 Statistical Assumption Checks

To ensure the robustness of our inferential statistics, we performed rigorous checks of statistical assumptions before using t-tests, which compared means between two independent protein pairs. We evaluated the normality assumption, which is crucial for the t-test, using the Shapiro-Wilk test for both surface and deep sample data regarding total stability. The Shapiro-Wilk test indicated a departure from normality for the surface data ($W = 0.94$, $p = 0.03$) and a more pronounced deviation for the deep data ($W = 0.91$, $p = 2.8\times10^{-5}$). Additionally, we applied the Anderson-Darling normality test, which showed a marginal non-significance for the surface data ($A = 0.7$, $p = 0.056$) but a significant deviation for the deep data ($A = 1.97$, $p = 4.6\times10^{-5}$), suggesting that the deep sample distribution notably deviates from normality.

3.10 Homogeneity of Variances and Outlier Detection

Levene's test for Homogeneity of Variance, with the median as the central tendency measure, confirmed the homogeneity of variances across groups ($F = 1.16$, $p = 0.2817$), satisfying homoscedasticity assumptions. For outlier detection, Grubbs's test identified significant outliers in deep ($G = 3.87$, $p = 0.002$) and surface ($G = 5.18$, $p = 1.8\times10^{-9}$) sample stabilities, necessitating further outlier impact analysis. The Z-score method flagged four additional data points as outliers, prompting considerations for data transformation or robust statistical techniques.

CHAPTER 4 DISCUSSION AND CONCLUSION

To prevent artifacts from the fact that a signal protein itself is likely to add to the free energy of folding of a protein (Bhirde et al., 2018, show that, on average, addition of amino acids to proteins increases their thermodynamic stability), we compared only proteins that both had or both lacked signal peptides.

Our analysis encompassed proteins sampled from both surface and deep environments, revealing a significant stability difference between these two samples. A paired t-test, assessing stability differences at surface versus deeper levels, demonstrated a statistically significant mean difference of 30.5 kJ/mol (95% CI: 8.53 to 52.6, p-value = 0.007), with higher stability noted at greater depth. To further mitigate any bias introduced by size, we carefully removed protein pairs with significant size difference. Previous literature supports the relationship between protein size and stability through changes in the melting temperature (ΔTM) and standard free energy of folding(ΔG_{fold}) (Watson, Monroe, and Raleigh, 2018). A linear correlation exists between Δ_{TM} and ΔG_{fold}, with the effect of a given change in ΔG_{fold} being larger for small proteins compared to large proteins (Thomas, Joris, and Brasseur 2010). in our study, Pearson's product-moment correlation between total stability and sequence length revealed a weak, yet statistically significant, positive linear correlation with a coefficient of 0.27 ($p = 0.0029$). This indicates that as sequence length increases, there is a tendency for total stability to also increase, albeit modestly. These findings suggest that while sequence length contributes to total stability, the correlation is relatively weak, implying the influence of additional factors on stability that warrant further investigation.

In our investigation, the stability disparity was particularly notable in protein pairs lacking signal peptides. A significant difference was observed in the folding free energy (ΔG_{fold}) between paired homologous carbohydrate-active enzymes derived from surface and deep samples. The analysis revealed a t-value of 4.03with 20 degrees of freedom ($t(20) = 4.029$), indicating a significant difference ($p < 0.001$). Specifically, the mean ΔG_{fold} for the surface sample exceeded that of the deep sample by 67.4 $kJ.mol^{-1}$, with a 95% confidence interval spanning from 32.5 to 102. This pronounced difference can underscore the influence of environmental factors on the stability of carbohydrate-active enzymes, particularly those without signal peptides, which can suggest a modulation of protein folding energetics in adapting to varying depths. Within the analysis, a notable distinction in ΔG_{fold} was observed predominantly among pairs from the glycoside hydrolase (GH) family that were devoid of

signal peptides. This particular subset displayed a marked increase in ΔG_{fold} for surface samples compared to their deep counterparts, evidencing a t-value of 3.4 (degrees of freedom = 12) and a p-value of 0.005. The mean difference between these groups was calculated to be 81.7, with a 95% confidence interval ranging from 29.6 to 133.7. This data underscores the enhanced stability of enzymes located at the surface in contrast to those found in deeper samples.

Conversely, when examining the glycosyltransferase (GT) family under identical conditions, the analysis indicated an anomaly that significantly impacted the results. Despite the anticipated trend, an outlier led to a diminished statistical significance, with the mean difference between surface and deep samples being -25.2 kJ/mol (95% CI: -136.6 to 86.1, p-value = 0.6). Consequently, this outcome suggested no statistically significant difference in ΔG_{fold} means between the surface and deep samples for the GT family.

In our analysis of carbohydrate-active enzymes, it was observed that 66% of the protein pairs in the deep samples exhibited lower Gibbs free energy of folding (ΔG_{fold}) compared to their counterparts. A lower ΔG_{fold} in the context of proteins generally denotes increased stability.

Contrary to our expectations, our investigation revealed no significant difference in the stability of extracellular proteins between deep and surface samples. This outcome does not align with our initial hypothesis, indicating that the environmental depth may impact the stability of these proteins as predicted. This was substantiated by a paired t-test assessing the overall stability difference between proteins with signal peptide from surface and deep environments, which did not show a statistically significant difference (mean difference: -4.17, with a 95% confidence interval from -44.6 to 36.2, and a p-value of 0.8).

Nonetheless, a distinct pattern emerged upon examining proteins without signal peptides; here, a difference in stability between the two environments was discernible. This suggests that the presence of a signal peptide may play a role in the environmental adaptability and stability of proteins($t(20) = 4.03$, $p = 0.0006$), providing a specific area of interest for further research to understand the mechanisms at play and their implications for protein function across different environments.

4.1 Protein structure

The structural differences between protein pairs have not been extensively studied and will be the focus of our future research, but an examination of various protein pairs shows a considerable diversity in their structures.

In many protein pairs, notable differences are observed in the N-terminal region, which functions as the signal peptide (Figure 17). Since these N-terminal signal peptides are cleaved from proteins prior to secretion and are absent during stability analyses, they can be excluded from consideration in structural studies.

In some protein pairs there are some loops which can be related to higher stability. These are regions where the polypeptide chain extends away from the other homologous structure. They are flexible regions that often connect secondary structural elements. They do not have a regular, repeating structure and are often found on the protein surface. Loops can vary significantly in length and can be involved in binding sites or functional regions of the protein.

4.2 Limitations and future directions

A notable limitation of our study was the sequencing depth and coverage achieved during the genomic analysis, which led to a small sample size, which may have constrained the statistical power and, consequently, the conclusiveness of our findings. This limitation is particularly significant in the context of employing t-tests for analyzing differences between deep and surface proteins, where the adequacy of the sample size is essential for ensuring reliable and valid results. A smaller sample can also increase susceptibility to outliers, potentially resulting in less precise estimates of the true effect sizes. The small sample size may increase the likelihood of type II errors, where true associations or effects may fail to reach statistical significance. Recognizing this limitation, it is imperative for future research to employ larger sample sizes that can more accurately reflect the population diversity and enhance the robustness of the findings. Such studies would not only validate the initial observations made in this work but also potentially reveal additional insights that were not detectable in the current study due to its constrained sample size.

The stability difference observed among protein pairs was significant in many cases but not consistent across all protein classes. Several reasons account for this. First of all, the observed difference aligns with our hypothesis, although its lack of statistical significance suggests it could be within the range of random variation if the null hypothesis (no difference between depths) were true. Secondly, with only 135 proteins in the analysis, statistical power is poor, potentially affecting the robustness of the findings.

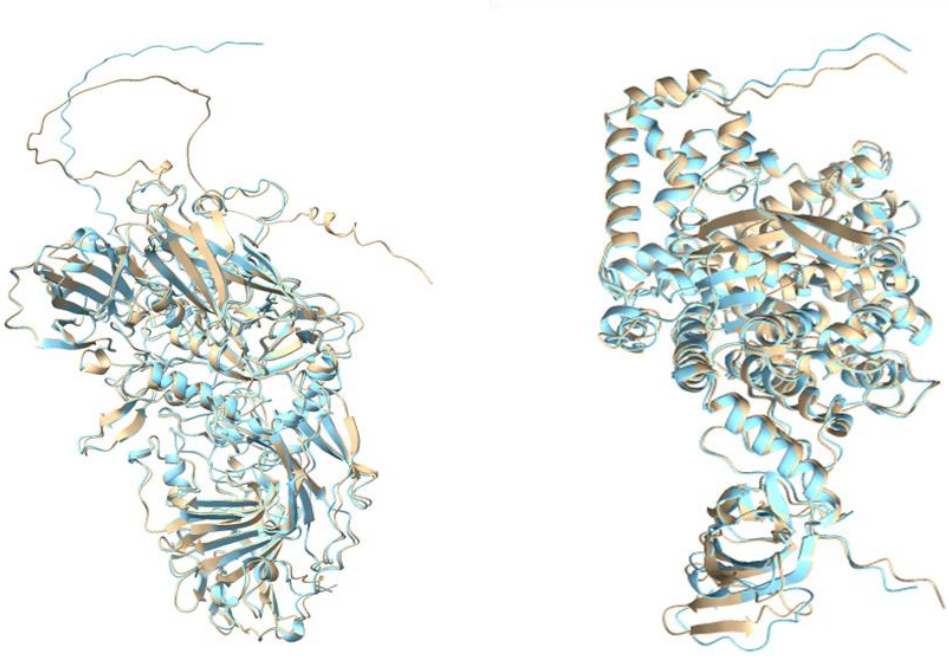

Figure 17 Presents the structural differences between two pairs of proteins. On the left, the pair SRR7066493_gene_5956 (blue) and SRR7066492_gene_39840 (orange), both from the Glycosyl Hydrolases Family 2, exhibit significant variations in the length and position of their signal peptides. In contrast, on the right, the pair SRR7066492_gene_23617 (blue) and SRR7066493_gene_8504 (orange), both classified as Carbohydrate Phosphorylases, display minor differences in the N-terminal regions of the two proteins.

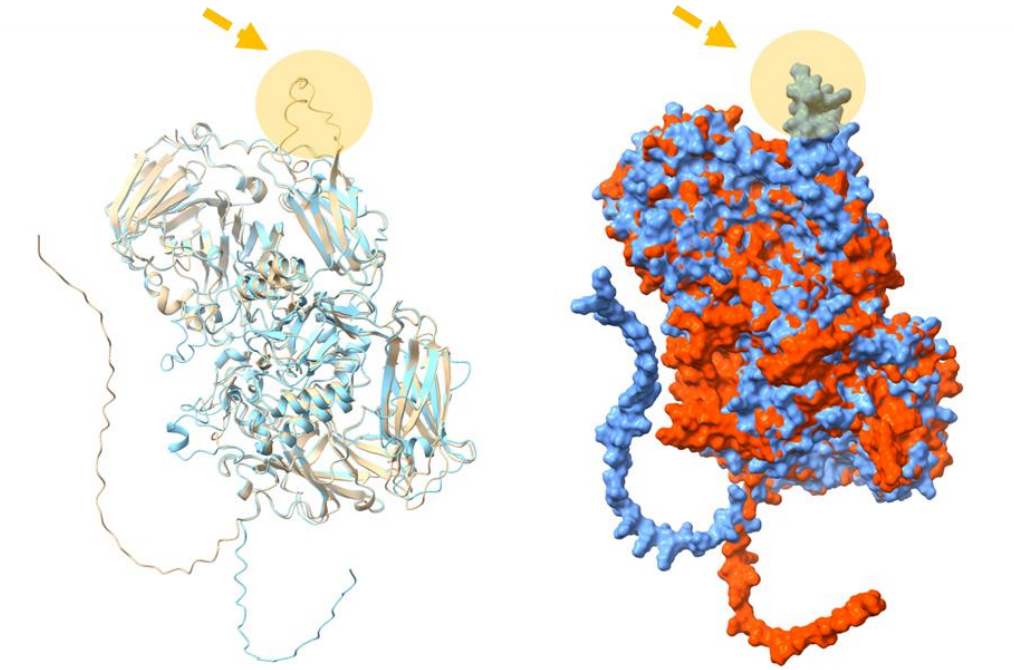

Figure 18 Illustrates the protein pair SRR7066492_gene_14490 and SRR7066493_gene_5956, both members of the Glycosyl Hydrolases Family 2, presented in both cartoon and spheres display. Notably, an extrusion at the top of the protein is observed, a characteristic that has been consistently identified across multiple protein pairs in our samples.

Contrary to our hypothesis, these data fail to support the notion that all classes of proteins in deep microbes are inherently more stable than those in surface microbes. In addition, certain classes of proteins might be more affected by this phenomenon than others. For example, glycosyl transferases could be stabilized by other means, such as chaperone proteins or osmolytes, or their stability might not be as crucial for reasons yet to be understood.

While signal peptides often direct proteins to be secreted outside the cell, not all proteins with signal peptides end up in extracellular environments; some proteins are directed to specific organelles within the cell or to the cell membrane. Therefore, while signal peptides are commonly associated with extracellular proteins, they are not exclusive to them. Investigating peptide localization can be achieved computationally using various tools and methods, as outlined previously (Imai & Nakai, 2020). Future research could focus on elucidating peptide localization and exploring its potential impact on protein stability.

Through the discovery of enzymes with higher stabilities, the richness of marine biodiversity and the potential inherent in further exploring the oceans can be highlighted. It underscores the importance of preserving marine environments, as they could be sources of novel enzymes and compounds beneficial for various applications. Stable enzymes could potentially indicate adaptations of marine organisms to their environment, shedding light on evolutionary processes and the ability of marine life to persist in diverse and challenging conditions.

LIST OF REFERENCES

1. LaRowe, D. E. *et al.* The fate of organic carbon in marine sediments - New insights from recent data and analysis. *Earth-Sci. Rev.* **204**, 103146 (2020).
2. Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C. & D'Hondt, S. Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl. Acad. Sci.* **109**, 16213–16216 (2012).
3. Zhang, X.-H., Ahmad, W., Zhu, X.-Y., Chen, J. & Austin, B. Viable but nonculturable bacteria and their resuscitation: implications for cultivating uncultured marine microorganisms. *Mar. Life Sci. Technol.* **3**, 189–203 (2021).
4. Braun, S. *et al.* Microbial turnover times in the deep seabed studied by amino acid racemization modelling. *Sci. Rep.* **7**, 5680 (2017).
5. Hoehler, T. M. & Jørgensen, B. B. Microbial life under extreme energy limitation. *Nat. Rev. Microbiol.* **11**, 83–94 (2013).
6. Schmidt, J. M., Royalty, T. M., Lloyd, K. G. & Steen, A. D. Potential Activities and Long Lifetimes of Organic Carbon-Degrading Extracellular Enzymes in Deep Subsurface Sediments of the Baltic Sea. *Front. Microbiol.* **12**, 702015 (2021).
7. Fang, Y. Gibbs Free Energy Formula for Protein Folding. in *Thermodynamics - Fundamentals and Its Application in Science* (ed. Morales-Rodriguez, R.) (InTech, 2012). doi:10.5772/52410.
8. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res.* **33**, W382–W388 (2005).
9. Atsavapranee, B., Stark, C. D., Sunden, F., Thompson, S. & Fordyce, P. M. Fundamentals to function: Quantitative and scalable approaches for measuring protein stability. *Cell Syst.* **12**, 547–560 (2021).
10. Jumper, J. Highly accurate protein structure prediction with AlphaFold.
11. Matthieu, G. *et al.* Protein denaturation and protein:drugs interactions from intrinsic protein fluorescence measurements at the nanolitre scale. *Protein Sci.* **19**, 1544–1554 (2010).
12. Cecconi, C., Shank, E. A., Bustamante, C. & Marqusee, S. Direct Observation of the Three-State Folding of a Single Protein Molecule. *Science* **309**, 2057–2060 (2005).
13. Kellermayer, M. S. Z., Smith, S. B., Granzier, H. L. & Bustamante, C. Folding-Unfolding Transitions in Single Titin Molecules Characterized with Laser Tweezers. *Science* **276**, 1112–1116 (1997).
14. Puchner, E. M. & Gaub, H. E. Force and function: probing proteins with AFM-based force spectroscopy. *Curr. Opin. Struct. Biol.* **19**, 605–614 (2009).
15. Andrén, T., Jørgensen, B.B., Cotterill. (2015). Expedition 347 summary DOI: 10.2204/iodp.proc.347.2015
16. Marshall, I. P. G., Karst, S. M., Nielsen, P. H. & Jørgensen, B. B. Metagenomes from deep Baltic Sea sediments reveal how past and present environmental conditions determine microbial community composition. *Mar. Genomics* **37**, 58–68 (2018).
17. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
18. KneadData – The Huttenhower Lab. https://huttenhower.sph.harvard.edu/kneaddata/.
19. Langmead, B., Wilks, C., Antonescu, V. & Charles, R. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics* **35**, 421–432 (2019).
20. Liang, K. & Sakakibara, Y. MetaVelvet-DL: a MetaVelvet deep learning extension for de novo metagenome assembly. *BMC Bioinformatics* **22**, 427 (2021).
21. Afiahayati, Sato, K. & Sakakibara, Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. *DNA Res.* **22**, 69–77 (2015).
22. Teufel, F. *et al.* SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025 (2022).
23. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
24. Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* **46**, D624–D632 (2018).
25. Drula, E. *et al.* The carbohydrate-active enzyme database: functions and literature. *Nucleic*

Acids Res. **50**, D571–D577 (2022).

26. Durowoju, I. B., Bhandal, K. S., Hu, J., Carpick, B. & Kirkitadze, M. Differential Scanning Calorimetry — A Method for Assessing the Thermal Stability and Conformation of Protein Antigen. *J. Vis. Exp. JoVE* 55262 (2017) doi:10.3791/55262.
27. Johnson, C. M. Differential scanning calorimetry as a tool for protein folding and stability. *Arch. Biochem. Biophys.* **531**, 100–109 (2013).
28. Greenfield, N. J. Analysis of the kinetics of folding of proteins and peptides using circular dichroism. *Nat. Protoc.* **1**, 2891–2899 (2006).
29. Corrêa, D. & Ramos, C. The use of circular dichroism spectroscopy to study protein folding, form and function. *Afr. J. Biochem. Res.* **3**, 164–173 (2009).
30. Liu, P.-F., Avramova, L. V. & Park, C. Revisiting absorbance at 230nm as a protein unfolding probe. *Anal. Biochem.* **389**, 165–170 (2009).
31. Hamborg, L., Horsted, E. W., Johansson, K. E., Willemoës, M., Lindorff-Larsen, K., & Teilum, K. (2020). Global analysis of protein stability by temperature and chemical denaturation. *Analytical Biochemistry*, *605*, 113863. https://doi.org/10.1016/j.ab.2020.113863
32. Bhirde, A. A., Chiang, M.-J., Venna, R., Beaucage, S., & Brorson, K. (2018). High-Throughput In-Use and Stress Size Stability Screening of Protein Therapeutics Using Algorithm-Driven Dynamic Light Scattering. *Journal of Pharmaceutical Sciences*, *107*(8), 2055–2062. https://doi.org/10.1016/j.xphs.2018.04.017
33. Watson, M. D., Monroe, J., & Raleigh, D. P. (2018). Size-Dependent Relationships between Protein Stability and Thermal Unfolding Temperature Have Important Implications for Analysis of Protein Energetics and High-Throughput Assays of Protein–Ligand Interactions. *The Journal of Physical Chemistry B*, *122*(21), 5278–5285. https://doi.org/10.1021/acs.jpcb.7b05684
34. Thomas, A., Joris, B., & Brasseur, R. (2010). Standardized evaluation of protein stability. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1804*(6), 1265–1271. https://doi.org/10.1016/j.bbapap.2010.02.008
35. Davis, C. M., & Gruebele, M. (2018). Non-Steric Interactions Predict the Trend and Steric Interactions the Offset of Protein Stability in Cells. *ChemPhysChem*, *19*(18), 2290–2294. https://doi.org/10.1002/cphc.201800534
36. Speer, S. L., Zheng, W., Jiang, X., Chu, I.-T., Guseman, A. J., Liu, M., Pielak, G. J., & Li, C. (2021). The intracellular environment affects protein–protein interactions. *Proceedings of the National Academy of Sciences*, *118*(11), e2019918118. https://doi.org/10.1073/pnas.2019918118
37. Hou, Q., Rooman, M., & Pucci, F. (2023). Enzyme Stability-Activity Trade-Off: New Insights from Protein Stability Weaknesses and Evolutionary Conservation. *Journal of Chemical Theory and Computation*, *19*(12), 3664–3671. https://doi.org/10.1021/acs.jctc.3c00036
38. Jaenicke, R., & Böhm, G. (1998). The stability of proteins in extreme environments. *Current Opinion in Structural Biology*, *8*(6), 738–748. https://doi.org/10.1016/S0959-440X(98)80094-8
39. Arcus, V. L., Van Der Kamp, M. W., Pudney, C. R., & Mulholland, A. J. (2020). Enzyme evolution and the temperature dependence of enzyme catalysis. *Current Opinion in Structural Biology*, *65*, 96–101. https://doi.org/10.1016/j.sbi.2020.06.001
40. Xie, Y., An, J., Yang, G., Wu, G., Zhang, Y., Cui, L., & Feng, Y. (2014). Enhanced Enzyme Kinetic Stability by Increasing Rigidity within the Active Site. *Journal of Biological Chemistry*, *289*(11), 7994–8006. https://doi.org/10.1074/jbc.M113.536045
41. Steen, A. D., Royalty, T. M., Schmidt, J. M., & Lloyd, K. G. (2022). Do deep marine sediments host unusually stable enzymes? *Biophysical Journal*, *121*(3), 296a–297a. https://doi.org/10.1016/j.bpj.2021.11.1271
42. Frauenfelder, H., Sligar, S. G., & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science (New York, N.Y.)*, *254*(5038), 1598–1603. https://doi.org/10.1126/science.1749933

VITA

Kambiz Kalhor was born on January 28, 1993, in Qazvin, Iran. he graduated from Alama Jafari Special High School, Iran. He received his Bachelor of Science in biotechnology in 2020. Before attending the University of Tennessee, he spent two years teaching bioinformatics at a research institute in Tehran.

Currently, Kambiz is a Master's student at the University of Tennessee Knoxville, working at Dr. Steen's lab. He lives in Knoxville, Tennessee, at this time.