Shahjalal University of Science and Technology
Department of Computer Science and Engineering

Deciphering the Biophysical Grammar of Extremophiles: A Hierarchical and Explainable Protein Language Modeling Approach

MD Nasiat Hasan Fahim

Reg. No.: 2020331013

4^{th} year, 1^{st} Semester

Md. Abid Ullah Muhib

Reg. No.: 2020331089

4^{th} year, 1^{st} Semester

Department of Computer Science and Engineering

Supervisor

Dr. M. Shahidur Rahman

Professor

Department of Computer Science and Engineering

28th November, 2025

Abstract

Protein language models (pLMs) have revolutionized protein science yet suffer from severe taxonomic bias, systematically penalizing extremophiles whose molecular adaptations defy mesophilic "grammar." This bias reflects a fundamental limitation: universal models average over fundamentally incompatible biophysical strategies—halophilic surface acidification versus thermophilic core compaction—compromising performance on underrepresented organisms. We address this through a rigorous, evidence-based framework for de-averaging evolutionary signals. Using ESM-2(650M) as foundation, we conducted systematic ablation studies on 535,000 sequences (Gammaproteobacteria, Thermoproteota, Halobacteria) partitioned via GraphPart homology clustering to empirically identify the optimal training strategy. Our hierarchical LoRA architecture (Base→Kingdom→Lifestyle) emerged as superior, reducing trainable parameters by 100-fold while maintaining general protein knowledge (Forgetting Index < 10%). Lifestyle-adapted models exhibited biophysical inversion: Halobacteria-tuned models exclusively favored surface acidification mutations (attribution concentration $Delta$ACI > 0.15, p < 0.001), while Thermoproteota-tuned models prioritized core hydrophobicity, with strong anticorrelation (ρ < -0.30, p < 0.001) confirming incompatible adaptation grammars. Critically, cross-clade validation demonstrated that lifestyle-matched models generalized 0.10 Spearman units better (p < 0.01) to phylogenetically distant extremophiles than phylogenetically closer but lifestyle-agnostic models, establishing biophysical lifestyle as the dominant constraint over ancestry. Integrated gradients analysis automatically recovered established physical adaptation rules without supervision, providing mechanistic interpretability. This work transforms taxonomic bias from a liability into an exploitable feature, delivering parameter-efficient, interpretable models for extremophile protein engineering with immediate applications in industrial biocatalysis and climate adaptation research.

Keywords: Protein language models, extremophile adaptation, taxonomic bias, hierarchical fine-tuning, low-rank adaptation (LoRA), biophysical grammar, GraphPart homology partitioning, variant effect prediction, ProteinGym, explainable AI, integrated gradients, attribution concentration index, catastrophic forgetting, parameter-efficient training, halophilic proteins, thermophilic proteins, cross-clade generalization, synthetic biophysical traps, mechanistic interpretability, protein engineering, biophysical paradox.

Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION AND BACKGROUND

1.1 The Genomic Revolution and the Functional Annotation Crisis

The contemporary landscape of biological sciences is characterized by an unprecedented expansion in our capacity to generate genomic sequence data. High-throughput sequencing technologies have fundamentally transformed biological inquiry from data acquisition to data interpretation challenges [1, 2]. As of 2024, UniProtKB includes annotations to 12,501 Rhea reactions, which are linked to 28,259 857 UniProtKB protein sequence records, including 231,709 reviewed protein sequence records in UniProtKB/Swiss-Prot [3]. This expanding chasm between sequence availability and functional understanding represents one of the most pressing challenges in computational biology, creating what has been termed the "dark matter" of the protein universe.

This functional annotation deficit is not uniformly distributed across the tree of life. Protein databases exhibit significant taxonomic bias, with species representation growing randomly over time without systematic roadmaps for optimal sampling across evolutionary diversity [4–6]. The consequences of this unequal sampling extend far beyond mere data incompleteness. Recent rigorous analyses have demonstrated that protein language models assign systematically higher likelihoods to sequences from well-represented species such as *Escherichia coli* and *Homo sapiens*, while penalizing functionally comparable proteins from underrepresented taxa [6]. This taxonomic bias has profound implications for protein engineering, drug discovery, and our fundamental

understanding of evolutionary adaptation.

Extremophilic microorganisms, which thrive in environments characterized by extreme temperature, salinity, pH, or pressure, are particularly underrepresented in current protein databases. These organisms have evolved through billions of years of relentless physicochemical selection, resulting in proteomes adapted to conditions that would instantly denature proteins from mesophilic organisms [7–10]. Their proteins embody unique structural solutions to maintaining stability and catalytic activity under extreme conditions, yet our computational tools remain largely blind to these specialized adaptations [11].

1.2 The Emergence and Limitations of Protein Language Models

The transformer revolution, sparked by the seminal 'Attention is All You Need' paper [12], has driven a paradigm shift in computational protein analysis through the development of Protein Language Models (pLMs). Drawing inspiration from natural language processing, models like ESM-2 treat amino acid sequences as sentences in a biological language, learning statistical regularities through masked language modeling on billions of unannotated sequences [1, 13]. ESM-2, scaling to 15 billion parameters, represents the largest protein language model evaluated to date, demonstrating emergent capabilities in structure prediction at atomic resolution [13].

The architecture underlying these models has evolved considerably. Transformer protein language models [1, 12], introduced with the principle that biological structure and function emerge from scaling unsupervised learning, employ self-attention mechanisms enabling the model to weigh the importance of every residue relative to every other residue regardless of linear sequence distance. Recent investigations have revealed that ESM-2 learns evolutionary statistics of interacting sequence motifs, storing coevolutionary information analogously to simpler modeling approaches like Markov Random Fields [14]. This finding suggests that while pLMs exhibit impressive performance, they may learn statistical patterns rather than fundamental biophysical principles.

The performance of pLMs on variant effect prediction tasks has been extensively documented. A workflow using ESM1b to predict all approximately 450 million possible missense variant effects in the human genome [15] demonstrated superiority over existing methods in classifying ClinVar and HGMD variants and predicting deep mutational scanning measurements [16]. However, these successes mask a critical limitation that has only recently been systematically characterized.

1.3 The Taxonomic Blind Spot: Species Bias as a Fundamental Challenge

Ding and Steinhardt demonstrated in 2024 that protein language model likelihoods unintentionally encode species bias, with sequences from certain taxa assigned systematically higher probabilities independent of the protein in question [6]. This bias manifests through two distinct mechanisms. First, phylogenetic memorization occurs when models learn to recognize taxonomic markers such as ribosomal binding sites or codon usage patterns rather than functional constraints. Second, biophysical drift emerges when models pull designed sequences toward the average stability profile of the training set, which for extremophiles represents a destabilizing mesophilic baseline.

Variance partitioning analyses have revealed that sequence-only models like ESM2 retain substantial taxonomic dependence, with 19–25% of unexplained variance attributable to species identity even after controlling for protein family and biophysical covariates [6]. This residual bias means that a low model score could indicate either poor sequence quality or origin from an underrepresented species, fundamentally confounding the interpretability of predictions for protein design decisions.

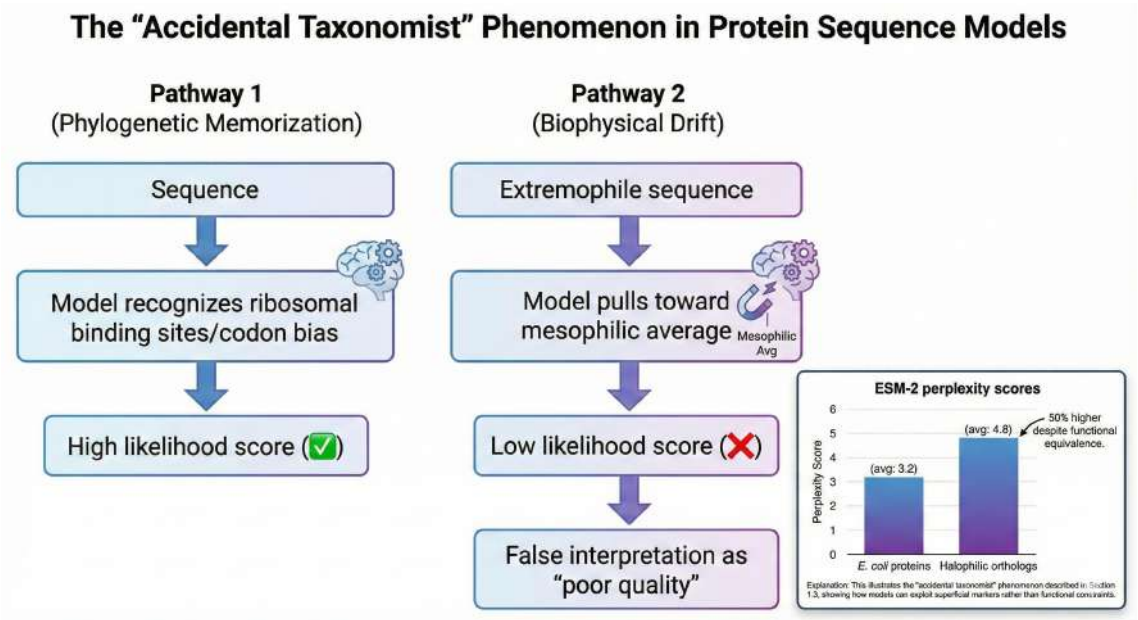

Figure 1.1: Taxonomic Bias Mechanism in pLMs

The implications extend to underrepresented domains such as viral proteins, where fine-tuning

pre-trained models on domain-specific datasets using parameter-efficient strategies significantly enhances representation quality and downstream task performance [17]. The "accidental taxonomist" hypothesis proposes that neural networks can exploit phylogenetic distances across labels in protein datasets rather than genuine interaction features, particularly problematic when negative sampling in training datasets creates systematic taxonomic separation between positive and negative examples [18].

Investigations into genomic language models have revealed that superior performance on certain tasks, such as melting point prediction, largely reflects the models' ability to identify and condition on species based on codon bias fingerprints present in coding sequences rather than capturing intrinsic protein properties [19]. This finding underscores that apparent predictive success may mask reliance on superficial taxonomic markers rather than genuine understanding of molecular mechanisms.

1.4 The Biophysical Paradox: Conflicting Grammars of Protein Stability

The taxonomic bias problem is particularly acute for extremophiles because these organisms employ fundamentally different strategies for protein stability that are mutually incompatible with mesophilic adaptations. Current universal models attempt to learn a single grammar of protein folding, effectively averaging contradictory biophysical constraints.

1.4.1 Halophilic Adaptation: The Acidic Surface Paradigm

Halophilic organisms cope with high intracellular salinity by modifying protein chemical composition to enrich surfaces with acidic and short polar side chains while depleting lysines and bulky hydrophobic residues [20–23]. Structural analyses demonstrate that acidic residues cluster on protein surfaces, facilitating excess hydration that maintains surface hydrophilicity and flexibility while promoting nonspecific electrostatic interactions with salts in solution [24–26].

Neutron diffraction studies of aspartic acid solutions at physiological pH in varying potassium chloride concentrations have provided structural evidence for solvent-stabilization by aspartic acid as a mechanism for halophilic protein stability [27], with surface aspartic acid residues maintaining

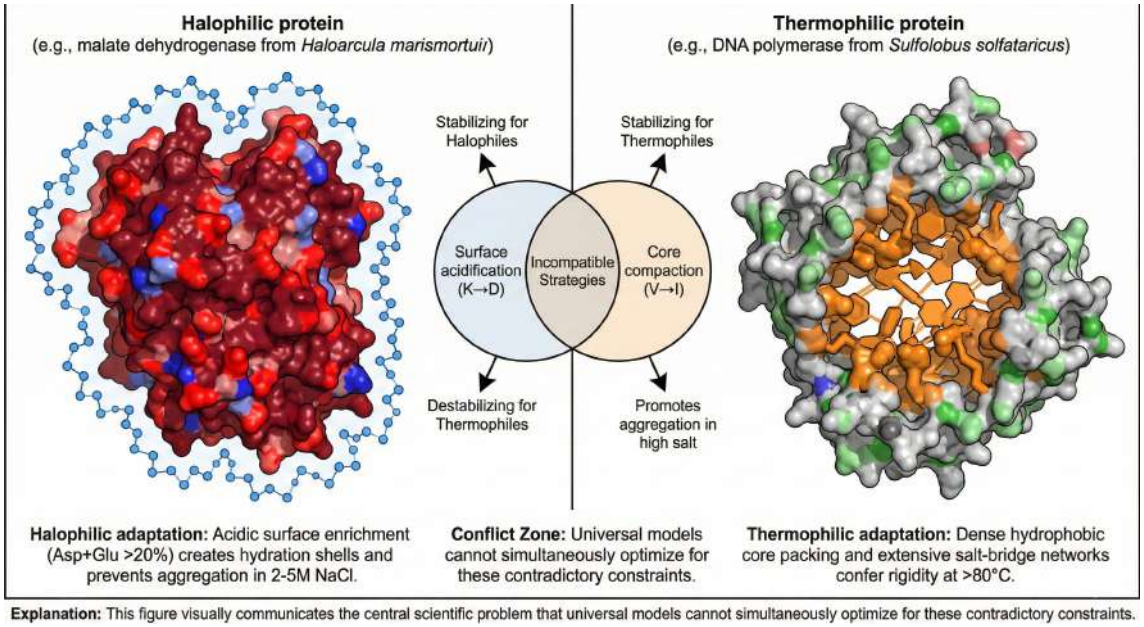

Figure 1.2: The Biophysical Paradox—Conflicting Adaptation Strategies

hydration layers at molar salt concentrations [27]. At the proteomic level, halophilic species are characterized by low hydrophobicity, over-representation of acidic residues (especially aspartate), under-representation of cysteine, lower propensities for helix formation, and higher propensities for coil structure [28].

Comparative structural analysis of "salt-in" versus "osmolyte" halophilic strategies reveals that haloadaptation in the presence of molar salt concentrations necessitates weakening of conserved hydrophobic contact surfaces, with the most frequent amino acid exchange at contact sites being isoleucine to valine [29]. Statistical analyses demonstrate that halophilic proteins prefer higher average flexibility and polarity while avoiding higher positive charge, bulkiness, and hydrophobicity, with significant enrichment of aspartate, glutamate, histidine, proline, threonine, and valine [22,30].

1.4.2 Thermophilic Adaptation: The Rigid Core Paradigm

Thermophilic adaptation mechanisms include high hydrophobic core density, compact loops, increased hydrogen bonding, elevated isoelectric points, and extensive salt bridge networks that confer thermostability [31–34]. High-throughput comparative analysis of hyperthermophilic archaea and bacteria reveals that organisms originating in extreme environments, such as *Pyrococcus furiosus*, possess proteins significantly more compact and hydrophobic than mesophilic counter-

parts [35–37].

Electrostatic interactions represent one of the most critical factors in protein thermal stability [32, 33], with increased numbers of surface charges and salt bridges observed in thermophilic proteins contributing to conformational specificity, thermal stability, and oligomerization [38]. Molecular dynamics simulations reveal that salt bridge networks, such as Glu3-Lys62-Glu36 triads and Asp79-Lys83 ion pairs, maintain hydrophobic core protection and structural packing at elevated temperatures [33, 39].

Salt bridges between charged residue pairs produce enhanced protein stability at elevated temperatures due to reduced dielectric constant of water, with uniformity of internal motion within thermophilic protein domains reducing onset of local thermal unfolding [40]. Organisms that recolonized hot environments, such as *Thermotoga maritima*, relied on sequence-based mechanisms involving specific interactions like additional salt bridges rather than global structural compaction [37].

1.4.3 The Incompatibility of Adaptation Strategies

These adaptation strategies are fundamentally incompatible. A mutation increasing surface acidity (e.g., lysine to aspartate) would be stabilizing for a halophile by enhancing hydration and preventing aggregation, but potentially destabilizing for a mesophile by creating electrostatic repulsion. Conversely, mutations increasing core hydrophobicity essential for thermophile stability could promote aggregation in high-salt environments. A universal protein language model trained on aggregated data cannot simultaneously maximize both probabilities without environmental context.

1.5 Fine-Tuning Strategies for Protein Language Models

The adaptation of large pre-trained protein language models to specialized domains requires careful consideration of training strategies that balance task performance, computational efficiency, and preservation of general protein knowledge. Multiple approaches have been explored in the literature, ranging from parameter-efficient methods to full model fine-tuning, each with distinct advantages and limitations.

Parameter-efficient fine-tuning methods, particularly Low-Rank Adaptation (LoRA), have been

introduced to proteomics to address the computational and memory barriers of full model fine-tuning [41, 42]. Task-specific supervised fine-tuning almost always improves downstream predictions, with parameter-efficient approaches achieving similar improvements while consuming substantially fewer resources, enabling up to 4.5-fold acceleration of training over full model fine-tuning [43].

Integration of LoRA into ESM-2 models for signal peptide prediction achieved maximum Matthews correlation coefficient gains of 87.3% for categories with small training samples and overall gains of 6.1%, with LoRA requiring fewer computing resources and less memory than adapter tuning [44]. End-to-end LoRA fine-tuning of ESM-2 for protein property prediction tasks utilizing only sequence information, combined with multi-head attention mechanisms for integrating contact map information, demonstrates strong performance and faster convergence across regression and classification tasks [45].

For the 1.2 billion parameter ProtT5 model, LoRA reduces trainable parameters to approximately 3 million, enabling fine-tuning on commercial GPUs with around 10 GB memory by freezing original weight matrices while updating only low-rank matrices [46]. Extensions combining LoRA with techniques like Fully Sharded Data Parallel (FSDP) and quantization (QLoRA) enable efficient multi-GPU training with reduced memory footprint.

Alternatively, full fine-tuning approaches that unfreeze and train the final transformer layers have demonstrated higher capacity for domain adaptation in natural language processing, though at the cost of increased memory requirements and potential catastrophic forgetting. Hybrid strategies combining parameter-efficient adapters with selective layer unfreezing represent a middle ground, potentially capturing adaptations that low-rank decomposition cannot represent while maintaining computational tractability. Hierarchical or stacked adapter architectures, where domain-specific adapters build upon broader phylogenetic adapters, offer the promise of compositional knowledge disentanglement.

However, a critical gap exists in the protein modeling literature: systematic empirical comparison of these training strategies specifically for extremophile protein adaptation remains absent. Prior work has typically assumed a single approach without rigorous ablation studies evaluating trade-offs between task performance, catastrophic forgetting, computational cost, and generalization across the unique constraints of limited extremophile training data. This research addresses this

gap through comprehensive comparative evaluation of four distinct training strategies—single adapters, hierarchical stacking, full fine-tuning, and hybrid approaches—on carefully controlled pilot datasets, ensuring that our final methodology is empirically validated rather than based on untested assumptions.

1.6 Explainable AI: Toward Mechanistic Understanding

Explainable artificial intelligence methods, particularly integrated gradients extended to latent representations inside transformer models fine-tuned for Gene Ontology and Enzyme Commission number prediction, enable identification of amino acids that transformers pay particular attention to, reflecting expectations from biology and chemistry [47, 48]. Integrated gradients satisfy axiomatic properties including sensitivity and completeness, with extensions addressing noise sources through improved integral path selection, handling of high-curvature decision surfaces, and refined Riemann approximations [49].

In protein modeling contexts, gradient-based interpretability techniques including GraphGrad-CAM [50], GNN-LRP, DeepLIFT [51], and integrated gradients [47] trace gradient flow through trained models to attribute prediction relevance to specific nodes or features, though attention weights can exhibit unreliable consistency and biological relevance across training runs [52]. Application of integrated gradients and LIME [53] to protein secondary structure prediction models reveals key features influencing protein shape, with ROUGE-L scores from natural language processing proving effective for protein sequence evaluation [54].

1.7 Rigorous Evaluation: Homology Partitioning and Benchmark Standards

The GraphPart algorithm addresses the critical challenge of homology partitioning by dividing datasets such that closely related sequences always end up in the same partition while retaining maximum sequence numbers, constructing graphs where sequences are nodes and edges represent identity above chosen thresholds [55]. GraphPart employs restricted agglomerative single-linkage clustering followed by iterative sequence removal or movement between partitions to eliminate cross-partition edges, using global alignment percentage identity as the similarity measure [55].

Updated benchmarking of variant effect predictors using deep mutational scanning demonstrates that assessment against clinical observations introduces biases, necessitating use of independently generated protein function measurements from multiplexed assays of variant effects [56–58]. ProteinGym encompasses over 250 standardized deep mutational scanning assays spanning millions of mutated sequences alongside curated clinical datasets, providing comprehensive benchmarks specifically designed for protein fitness prediction and design [59, 60].

Fine-tuning protein language models with deep mutational scanning data using normalized log-odds ratio heads produces consistent improvements on held-out protein test sets and independent benchmarks from ProteinGym and ClinVar [61]. Zero-shot prediction methods utilizing multimodal deep representation learning from approximately 160 million proteins achieve state-of-the-art performance in mutational effect prediction [16], with attention mechanisms successfully identifying functional sites such as metal-binding residues [62].

Chapter 2

RESEARCH OBJECTIVES AND HYPOTHESES

2.1 Central Research Questions

This research addresses three fundamental questions at the intersection of machine learning, evolutionary biology, and protein biophysics:

RQ1 (Lifestyle-Specific Probability Divergence): Does hierarchical fine-tuning of ESM-2 on lifestyle-specific extremophile datasets produce statistically significant divergences in zero-shot likelihood scores for identical mutations compared to generic kingdom-level or universal models, and do these divergences align with known biophysical adaptation strategies?

RQ2 (Generalization to Phylogenetically Distant Taxa): Does a lifestyle-adapted model trained on a specific extremophile class generalize more effectively to phylogenetically distant organisms sharing the same environmental niche than models trained on broader taxonomic groups, thereby demonstrating that biophysical lifestyle constitutes a more potent constraint for protein modeling than deep phylogenetic ancestry?

RQ3 (Mechanistic Reconstruction via Explainable AI): Can integrated gradients and complementary explainability techniques automatically recover known physicochemical adaptation rules without explicit physical supervision, as evidenced by concentration of attribution scores on functionally relevant structural regions that differ systematically between lifestyle-adapted models?

2.2 Research Hypotheses

H1 (Biophysical Inversion Hypothesis): Mutations increasing surface acidity (e.g., lysine to aspartate substitutions) will be assigned high probability by the Halobacteria-tuned model but low probability by Thermoproteota and control models. Conversely, mutations increasing core packing density (e.g., valine to isoleucine substitutions in buried positions) will be favored exclusively by the Thermoproteota model. This biophysical inversion will manifest as negative correlation ($\rho < -0.3$, $p < 0.001$) between lifestyle-specific model scores for the same mutation applied to orthologous positions.

H2 (Lifestyle Constraint Dominance Hypothesis): A model trained specifically on Class Halobacteria will predict variant effects in phylogenetically distant halophilic species (e.g., Nanohaloarchaea from different phyla) with correlation to experimental deep mutational scanning data at least 0.10 Spearman units higher ($\Delta\rho \geq 0.10$, $p < 0.01$) than a model trained on the entire Phylum Euryarchaeota, demonstrating that shared biophysical constraints transcend phylogenetic distance.

H3 (Mechanistic Attribution Hypothesis): When analyzing identical orthologous proteins, the Halobacteria-tuned model will concentrate attribution scores (as measured by Attribution Concentration Index, ACI) on surface acidic residues ($ACI_{surface\_acidic} > 0.35$, $p < 0.001$ vs. random permutation baseline), while the Thermoproteota-tuned model will shift attribution to hydrophobic core residues and salt-bridge networks ($ACI_{core\_hydrophobic} > 0.40$, $p < 0.001$), with these differences validated against experimental structural data and mutagenesis studies.

2.3 Expected Contributions to the Field

This research will advance computational biology through five primary contributions:

2.3.1 De-averaging Evolutionary Signals

Current approaches aggregate evolutionary information into universal statistical potentials, treating lineage-specific adaptations as noise. This work introduces the first systematic framework for de-averaging these signals, moving from universal translator models that speak broken average protein to multilingual models fluent in specific biophysical dialects. This directly transforms the taxonomic bias identified by Ding & Steinhardt from a limitation into an exploitable feature.

2.3.2 Empirically Validated Training Strategy Framework

Rather than assuming a priori that any single training approach (LoRA, full fine-tuning, hybrid methods) is optimal, this research implements rigorous ablation experiments on pilot datasets to empirically identify the most effective strategy. This evidence-based methodology, including comprehensive evaluation of catastrophic forgetting and multi-dimensional performance metrics, provides a replicable blueprint for future protein language model adaptation studies. The comparative framework we establish—evaluating single adapters, stacked pipelines, full fine-tuning, and hybrid approaches—will guide researchers working with other specialized protein domains (e.g., membrane proteins, disordered proteins, antimicrobial peptides) in selecting appropriate training strategies for their specific applications.

2.3.3 Parameter-Efficient Adaptation Architecture

Whether the final selected approach involves hierarchical stacking of Low-Rank Adapters (Base → Kingdom → Lifestyle), single lifestyle-specific adapters, or hybrid methods, the architecture will create a modular system where biophysical rules can be compositionally combined or selectively applied. Parameter-efficient approaches reduce trainable parameters by 100–1,000-fold compared to full fine-tuning (depending on the selected strategy), democratizing access to high-performance biological prediction tools for laboratories with modest computational resources. If full fine-tuning of last N layers proves optimal despite higher costs, we will provide detailed guidance on memory optimization techniques (gradient checkpointing, mixed precision, ZeRO optimization) to make this approach accessible.

2.3.4 Comparative Explainable AI Framework

Rather than static interpretation asking "what does the model see," this research introduces comparative XAI methodology. By computing differential attribution maps between lifestyle-specific models analyzing identical proteins, we perform *in silico* differential scanning that highlights precisely which residues are evolutionarily re-weighted by environmental selection, transforming deep learning models into hypothesis-generation engines for structural biology.

2.3.5 Rigorous Evaluation Methodology

Implementation of GraphPart homology partitioning, multi-baseline integrated gradients computation, adversarial deconfounding, and phylogenetic signal quantification establishes new standards for rigorous assessment of protein language models, addressing the data leakage and evaluation circularity that have plagued previous studies.

Chapter 3

DETAILED METHODOLOGY

3.1 Experimental Design Overview

The experimental protocol comprises five sequential phases executed over an estimated 20-month timeline: (1) data curation with strict homology partitioning, (2) preliminary ablation study comparing training strategies on pilot datasets, (3) full-scale training using the empirically validated optimal strategy, (4) comprehensive benchmarking across multiple evaluation frameworks, and (5) mechanistic interpretation via explainable AI techniques. Each phase incorporates multiple validation checkpoints to ensure robustness, with the ablation study (Phase 2) serving as a critical decision point that determines the methodology for all subsequent phases.

3.2 Phase 1: Data Curation and Homology Partitioning

3.2.1 Dataset Selection and Taxonomic Stratification

Three distinct datasets will be curated from UniProtKB [3] (accessed December 2024, with accessions archived for reproducibility), representing orthogonal points in the space of biophysical adaptation strategies:

3.2.1.1 Dataset A (Control – Mesophilic Baseline):

Class Gammaproteobacteria [63] will serve as the mesophilic control, representing standard protein grammar under moderate temperature, neutral pH, and physiological salt concentrations. Represen-

tative genera include *Escherichia*, *Pseudomonas*, *Vibrio*, *Salmonella*, and *Shewanella*. The dataset will be down-sampled to approximately 500,000 sequences to match the scale of extremophile groups while maintaining taxonomic diversity through stratified sampling ensuring representation across all orders within the class.

Selection Query: `(taxonomy_id:1236) AND (fragment:false) AND (length:[100 TO 800])`

Quality filters will include exclusion of fragments and partial sequences, length restriction to 100–800 residues to focus on single-domain proteins or well-characterized multi-domain architectures, prioritization of Swiss-Prot reviewed entries supplemented with high-confidence TrEMBL entries for taxa with limited reviewed sequences, and removal of sequences with greater than 20% low-complexity regions as determined by SEG algorithm to prevent attribution artifacts.

3.2.1.2 Dataset B (Heat – Hyperthermophilic Adaptation):

Phylum Thermoproteota [64] (formerly TACK superphylum) encompasses the most extensively characterized hyperthermophiles, with optimal growth temperatures exceeding 80°C. Target orders include Sulfolobales (e.g., *Sulfolobus*, *Acidianus*), Thermoproteales (e.g., *Thermoproteus*, *Pyrobaculum*), and Desulfurococcales (e.g., *Aeropyrum*, *Hyperthermus*).

Estimated volume: ~100,000 sequences from UniProtKB plus ~25,000 sequences from high-quality Metagenome-Assembled Genomes (MAGs) sourced from hot spring metagenomic studies.

Selection Query: `(taxonomy_id:28889) AND (fragment:false) AND (length:[100 TO 800])`

MAG Quality Filters (following MIMAG standards [65]): All MAGs must demonstrate completeness exceeding 90% as assessed via CheckM [66], contamination below 5%, GTDB-tk [67] taxonomic assignment confidence exceeding 95%, presence of essential single-copy genes, and assembly N50 exceeding 10 kilobases.

3.2.1.3 Dataset C (Salt – Extreme Halophilic Adaptation):

Class Halobacteria [64] represents obligate extreme halophiles requiring 2–5M NaCl for growth [21], employing the "salt-in" strategy with proteome-wide acidic enrichment. Target orders include Halobacteriales (e.g., *Halobacterium*, *Haloarcula*), Haloferacales (e.g., *Haloferax*, *Halogeomet-*

ricum), and Natrialbales (e.g., *Natronobacterium*, *Natrialba*). Methanogens will be strictly excluded despite phylogenetic proximity to isolate the halophilic adaptation signal from methanogenic metabolic pathways.

Estimated volume: ~150,000 sequences.

Selection Query: `(taxonomy_id:183963) AND (fragment:false) AND (length:[100 TO 800])`

Table 3.1: Dataset Summary

Dataset	Taxonomic Query	Total Sequences	Reviewed (Swiss-Prot)	Post-GraphPart Estimate
A: Gammaproteobacteria (Mesophilic Control)	`taxonomy_id:1236`	18M	109k	350k
B: Thermoproteota (Hyperthermophilic)	`taxonomy_id:28889`	222k	4k	75k
C: Halobacteria (Extreme Halophilic)	`taxonomy_id:183963`	1M	2k	110k

3.2.2 Physicochemical Validation of Dataset Biophysical Signatures

Prior to model training, comprehensive physicochemical characterization will confirm that curated datasets exhibit expected biophysical signatures:

3.2.2.1 Isoelectric Point Distribution:

Halophilic proteomes should exhibit median pI ~4.2–4.8 [21,22] (significantly lower than mesophiles at ~6.0–6.5), reflecting surface acidic enrichment. Distribution comparisons via Kolmogorov–Smirnov test ($\alpha = 0.01$) will verify significant shifts.

3.2.2.2 Amino Acid Composition Profiles:

Per-proteome frequency analysis for the 20 standard amino acids will be conducted with particular focus on characteristic signatures. Halobacteria are expected to show Asp+Glu exceeding 20%, Lys below 2%, and (Asp+Glu)/Lys ratio exceeding 10. Thermoproteota should demonstrate Ile+Val enrichment in predicted buried positions and Gly+Ala enrichment overall. Control datasets should exhibit balanced composition approximating universal amino acid frequencies.

3.2.2.3 Surface vs. Core Composition:

Using AlphaFold2 [68,69] or ESMFold [13] structure predictions for a random 1000-protein subset per dataset, Relative Solvent Accessibility (RSA) will be computed via DSSP [70]. Amino acid frequency will be stratified by RSA, with surface residues defined as RSA exceeding 20% and core residues as RSA below 5%. This analysis will validate halophile surface acidic enrichment and thermophile core hydrophobic enrichment.

3.2.2.4 Low-Complexity Region Quantification:

Sequences with >20% low-complexity content (as determined by SEG [71] with window=12, trigger=2.2, extension=2.5) will be flagged for potential exclusion or special handling during attribution analysis to prevent confounding from intrinsically disordered regions.
All QC metrics will be compiled into Supplementary Table S1, establishing that datasets genuinely capture distinct biophysical adaptation strategies rather than arbitrary taxonomic divisions.

3.2.3 GraphPart Homology Partitioning Protocol

GraphPart divides datasets such that closely related sequences always reside in the same partition while retaining maximum sequence numbers [72]. The algorithm constructs a sequence graph where nodes represent proteins and edges connect sequences exceeding a specified identity threshold, then applies restricted agglomerative clustering followed by iterative partition refinement.

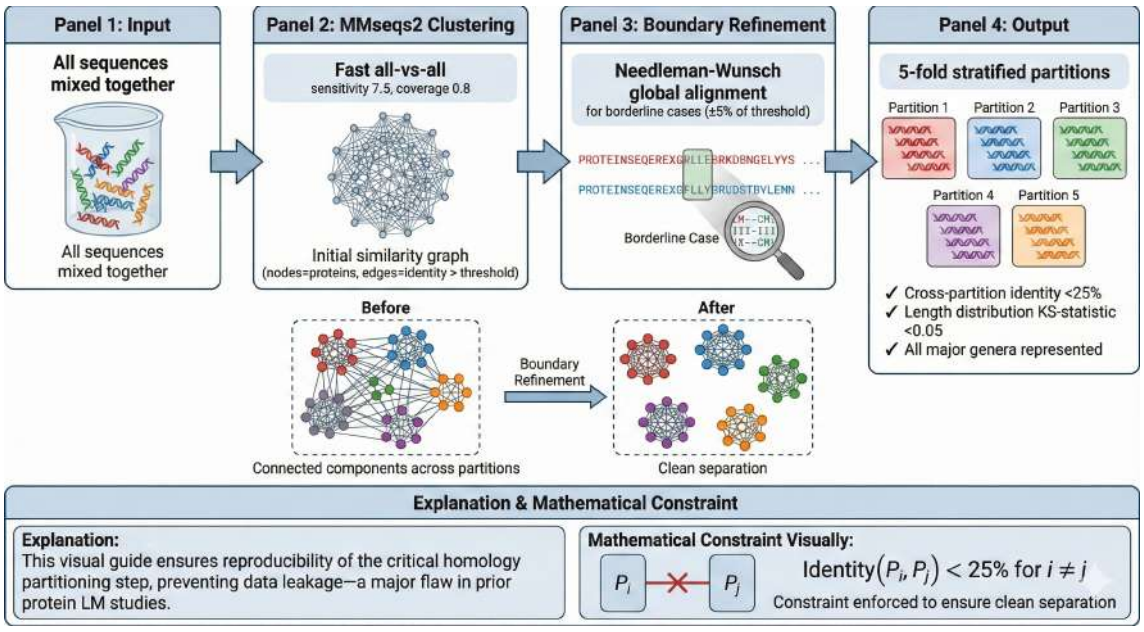

Figure 3.1: GraphPart Homology Partitioning Workflow

3.2.3.1 Implementation Parameters:

Identity Threshold Selection: Following established protein modeling conventions and preliminary sensitivity analyses, identity thresholds will be set as follows. Gammaproteobacteria will use 30% global identity, which is standard for generalizable protein models. Thermoproteota will employ a more stringent 25% global identity threshold due to higher sequence conservation within hyperthermophile families. Halobacteria will similarly use 25% global identity given within-class conservation patterns.

Alignment Strategy: A two-stage hybrid approach will balance computational efficiency with accuracy. Initial Clustering will employ MMseqs2 [73] with sensitivity 7.5, coverage 0.8, and e-value 0.001 for rapid all-versus-all comparison, generating an approximate identity graph. Boundary Refinement will apply Needleman–Wunsch global alignment [74] via EMBOSS needleall [75] for sequence pairs with MMseqs2 identity within ±5% of the threshold, ensuring accurate boundary decisions.

Partition Configuration: Five-fold stratified partitioning will enable robust cross-validation. Training will utilize 60% of sequences spanning 3 folds, validation will comprise 20% spanning 1 fold for hyperparameter optimization and early stopping, and testing will employ the remaining 20%

in 1 fold strictly held-out for final evaluation. Label stratification will be applied across sequence length bins of 100–200, 201–400, and 401–800 residues to ensure balanced representation.

3.2.3.2 Expected Data Reduction:

Preliminary GraphPart analyses on pilot datasets comprising 10,000 sequences per group indicate the following reductions. Gammaproteobacteria are expected to undergo approximately 30% reduction, retaining approximately 350,000 sequences, due to moderate within-genus similarity. Thermoproteota should experience approximately 40% reduction, retaining approximately 75,000 sequences, reflecting high conservation within thermophile lineages. Halobacteria are similarly expected to undergo approximately 40% reduction, retaining approximately 110,000 sequences, reflecting within-class conservation patterns.
The more aggressive reduction for extremophiles, while reducing absolute training data volume, is essential for obtaining unbiased performance estimates and preventing memorization of homologous test sequences.

3.2.3.3 Validation of Partition Quality:

Post-partitioning validation will verify three critical properties. Homology Separation will ensure maximum cross-partition identity remains below the threshold for all sequence pairs, verified via random sampling of 10,000 cross-partition pairs per dataset. Label Balance will confirm length distribution deviation below 5% across partitions, measured via Kolmogorov–Smirnov statistic. Taxonomic Representation will ensure all major genera are represented in each partition with a minimum of 10 sequences per genus per partition where taxonomically feasible.

3.2.4 Ortholog Identification for Comparative Analysis

To enable direct comparison of model predictions across lifestyles, orthologous protein groups spanning all three datasets will be identified:

3.2.4.1 Ortholog Clustering Pipeline:

The pipeline proceeds in three stages. Initial Candidates will be identified via BLAST reciprocal best hits with e-value below 10^{-20}, identity exceeding 35%, and coverage exceeding 70% between

all dataset pairs. Structural Validation will employ AlphaFold2-Multimer or Foldseek structure alignment with TM-score exceeding 0.7 to confirm shared fold. Functional Annotation will verify UniProt/InterPro functional domain architecture consistency, ensuring the same PFAM domains appear in the same order.

3.2.4.2 Target Ortholog Groups

(preliminary identification): Target ortholog groups will include Malate Dehydrogenase, which has been extensively studied with available structures for all lifestyles; Elongation Factor Tu, representing universal translation machinery with multiple DMS studies; Superoxide Dismutase, involved in oxidative stress response with available structural data; and DNA Polymerase subunits, representing replication machinery with conserved active sites. A minimum of 50 ortholog groups with at least 5 sequences per lifestyle will be identified.

These ortholog groups will serve as controlled test cases for RQ1 (biophysical inversion) and RQ3 (mechanistic attribution), enabling comparison of identical mutational effects predicted by different lifestyle-adapted models.

3.3 Phase 2: Training Strategy Selection and Hierarchical Fine-Tuning Architecture

3.3.1 Foundation Model Selection and Rationale

The ESM-2 model with 650 million parameters (esm2_t33_650M_UR50D) will serve as the foundation for all fine-tuning experiments. ESM-2 represents the largest protein language model systematically evaluated to date, achieving state-of-the-art results on structure prediction at atomic resolution [13]. The 650M parameter variant provides an optimal balance between representational capacity and computational tractability, requiring approximately 2.5GB GPU memory for inference and enabling training on widely available A100 or H100 GPUs.

Alternative foundation models were considered but rejected for specific technical reasons. The 15B parameter ESM-2 variant, while more powerful, imposes prohibitive memory requirements (60GB+ for training even with gradient checkpointing) that would severely constrain experimental throughput. Smaller variants (8M, 35M, 150M parameters) were deemed insufficient to capture

the complex long-range dependencies characteristic of extremophile protein adaptations. ESMFold, built on the ESM-2 architecture, demonstrated capability to predict accurate structures for orphan proteins with limited sequence homologs while operating an order of magnitude faster than AlphaFold2 [6], confirming that the ESM-2 embedding space captures evolutionarily meaningful structural information independent of multiple sequence alignments.

3.3.2 Preliminary Ablation Study: Comparative Training Strategy Evaluation

Rather than assuming that LoRA represents the optimal training strategy, we will conduct systematic ablation experiments on small pilot datasets to empirically identify the most effective approach. This data-driven strategy selection ensures that our final methodology is optimized for the specific characteristics of extremophile protein adaptation.

3.3.2.1 Pilot Dataset Construction:

For rapid experimentation, three balanced pilot datasets will be created, each containing 10,000 sequences randomly sampled from the full curated datasets. Pilot_Gamma will comprise 10,000 Gammaproteobacteria sequences, Pilot_Thermo will contain 10,000 Thermoproteota sequences augmented with MAGs if necessary, and Pilot_Halo will consist of 10,000 Halobacteria sequences. These pilot datasets will undergo the same GraphPart homology partitioning (80%/10%/10% train/val/test split) as the full datasets, ensuring that ablation study findings transfer to full-scale training.

3.3.2.2 Candidate Training Strategies:

Four distinct training strategies will be systematically compared:

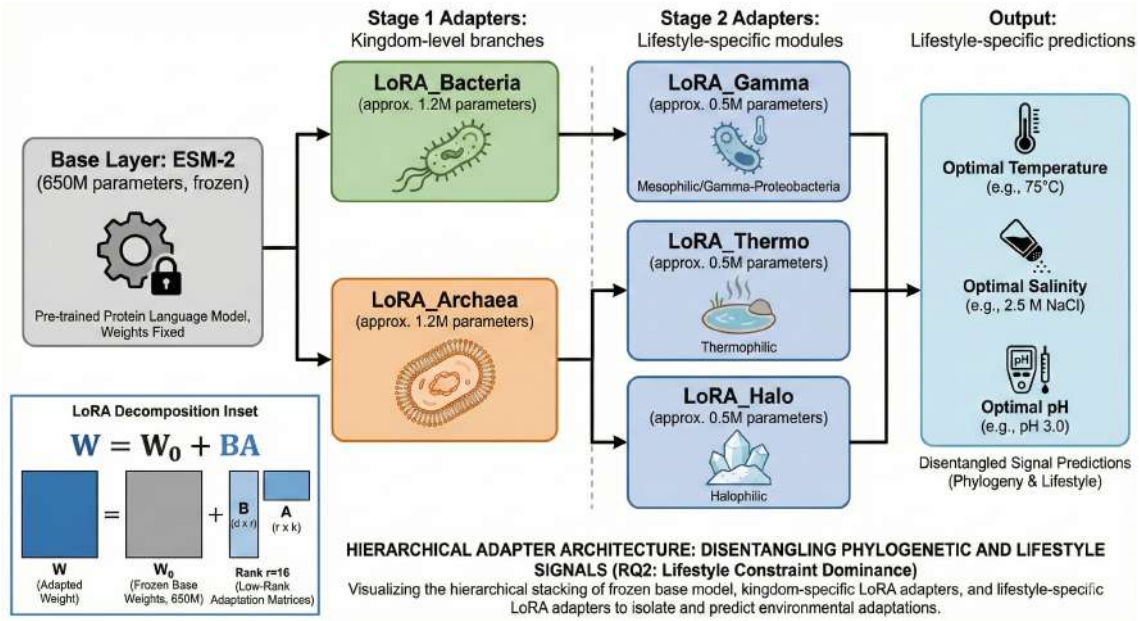

Figure 3.2: Hierarchical Adapter Architecture

Strategy 1: Single Adapter per Lifestyle (Baseline)

This approach trains a single, independent adapter module for each lifestyle category directly on the base ESM-2 model, without hierarchical structuring.

Architecture: One LoRA adapter applied to query/key/value projection matrices in all transformer layers, trained from scratch for each lifestyle.

Advantages: Simplest architecture, minimal interference between lifestyle-specific and general protein knowledge, fastest training time per model.

Potential Limitations: Does not explicitly capture kingdom-level phylogenetic structure, may need to relearn basic archaeal/bacterial features independently for each lifestyle.

Trainable Parameters: ~3.2 million (rank $r = 16$, 33 transformer layers).

Strategy 2: Stacked Adapter Pipeline (Hierarchical)

This is the hierarchical approach described previously, where lifestyle-specific adapters are stacked on top of kingdom-level adapters.

Architecture: A two-stage training curriculum will be employed. Stage 1 trains kingdom-level LoRA adapters including LoRA_Bacteria and LoRA_Archaea on broad phylogenetic datasets. Stage 2 trains lifestyle-specific adapters including LoRA_Gamma, LoRA_Thermo, and LoRA_Halo stacked on frozen Stage 1 adapters.

Advantages: Explicitly disentangles phylogenetic signal from biophysical adaptation, modular architecture enables compositional reuse of kingdom-level knowledge.

Potential Limitations: More complex training pipeline, risk of suboptimal Stage 1 adapters constraining Stage 2 performance, potential for error accumulation across stages.

Trainable Parameters: ~6.4 million total (3.2M per stage), but stages trained sequentially.

Strategy 3: Full Fine-Tuning of Last N Layers

Rather than using parameter-efficient adapters, this approach unfreezes and fully fine-tunes the final N transformer layers while keeping earlier layers frozen.

Architecture: For the 33-layer ESM-2 model, we will test $N \in \{2, 4, 6\}$ layers. All parameters in the last N layers (self-attention, layer norm, feed-forward networks) become trainable.

Advantages: Higher capacity for adaptation than low-rank adapters, no architectural modifications required, allows model to fundamentally restructure high-level representations.

Potential Limitations: Substantially higher memory requirements (gradient computation for full layers), increased risk of overfitting on small extremophile datasets, potential for catastrophic forgetting of general protein knowledge.

Trainable Parameters: ~40 million per layer (for 650M model), so 80–240 million depending on N.

Strategy 4: Hybrid Adapter + Selective Layer Unfreezing

This strategy combines the parameter efficiency of LoRA with targeted full fine-tuning, training adapters throughout the model while also unfreezing the last 2 transformer layers.

Architecture: LoRA adapters in all layers (as in Strategy 1) + full fine-tuning of final 2 transformer layers, trained simultaneously.

Advantages: Balances parameter efficiency with high-capacity adaptation where most beneficial (later layers encode task-specific features), potentially captures adaptations that low-rank decomposition cannot represent.

Potential Limitations: Most complex training configuration, highest memory requirements among parameter-efficient approaches, requires careful regularization to prevent overfitting.

Trainable Parameters: ~83 million (3.2M from LoRA + 80M from 2 layers).

3.3.2.3 Comparative Evaluation Metrics:

All four strategies will be trained on the pilot datasets using identical computational budgets (3 epochs, early stopping with patience=5) and evaluated across multiple dimensions:

Primary Task Performance: Evaluation will include masked language modeling perplexity on held-out test sets where lower values indicate better performance, per-token prediction accuracy stratified by amino acid class including charged, hydrophobic, and polar residues, and Spearman correlation with available DMS data for pilot proteins.

Catastrophic Forgetting Assessment:

To quantify the extent to which lifestyle-specific training erases general protein knowledge, we define the **Forgetting Index (FI)**:

$$\text{FI} = \frac{\text{Perplexity}_{\text{FineTuned}}(\mathcal{D}_{\text{general}}) - \text{Perplexity}_{\text{Base}}(\mathcal{D}_{\text{general}})}{\text{Perplexity}_{\text{Base}}(\mathcal{D}_{\text{general}})} \times 100\% \tag{3.1}$$

where $\mathcal{D}_{\text{general}}$ is a held-out set of 5,000 proteins spanning diverse non-extremophile organisms (metazoans, plants, fungi, mesophilic bacteria). FI quantifies the percentage increase in perplexity on general proteins after fine-tuning. A low FI (<10%) indicates successful knowledge retention, while high FI (>25%) signals catastrophic forgetting.

Additionally, we will compute the **Selective Forgetting Ratio (SFR)**:

$$\text{SFR} = \frac{\text{FI}_{\text{target\_lifestyle}}}{\text{FI}_{\text{other\_lifestyles}}} \tag{3.2}$$

Ideally, SFR should approach 0, indicating that performance on the target lifestyle improves while other lifestyles are minimally affected. High SFR (>0.5) suggests appropriate specialization, while SFR $\approx$ 1.0 indicates non-specific degradation.

Generalization and Transfer Learning: Evaluation will include cross-lifestyle transfer by evaluating each lifestyle-specific model on held-out test sets from other lifestyles to quantify specialization versus brittleness, and embedding space structure analysis using t-SNE visualization of [CLS] token embeddings to assess cluster separation.

Computational Efficiency: Metrics will include training time per epoch measured in wall-clock hours on standardized A100 GPUs, peak GPU memory consumption during training, and inference latency for single proteins measured in milliseconds.

Robustness and Stability: Assessment will include gradient norm variance across training batches where lower values indicate stable optimization, convergence speed measured as epochs required to reach 95% of final performance, and sensitivity to hyperparameter perturbations including learning rate variations of ±50% and rank variations of ±50% for LoRA strategies.

3.3.2.4 Decision Criteria for Final Strategy Selection:

The optimal training strategy will be selected via multi-objective decision analysis incorporating the following weighted criteria. Task Performance, weighted at 40%, prioritizes lowest perplexity on lifestyle-matched test sets with bonus for exceeding baseline by more than 5%. Knowledge Retention, weighted at 25%, requires Forgetting Index below 15% and penalizes strategies with FI exceeding 20%. Computational Efficiency, weighted at 20%, prefers parameter counts below 50M and training times below 24 hours per model. Generalization, weighted at 15%, requires graceful degradation on mismatched lifestyles with perplexity increase below 30% relative to matched conditions.

Statistical significance will be assessed via paired bootstrap tests (10,000 resamples, $\alpha = 0.05$ with Bonferroni correction for six pairwise comparisons). A strategy must significantly outperform alternatives on at least two of the four criteria to be selected.

3.3.2.5 Expected Outcomes and Contingency Planning:

Scenario 1 (Hypothesis: Stacked Pipeline Optimal): If Strategy 2 (stacked adapters) demonstrates superior performance, the full-scale experiments will proceed with the hierarchical architecture as originally planned, scaling to full datasets.

Scenario 2 (Hypothesis: Single Adapter Sufficient): If Strategy 1 (single adapter) performs comparably to Strategy 2 with substantially lower complexity, we will adopt the simpler single-adapter approach, eliminating the kingdom-level training stage.

Scenario 3 (Hypothesis: Full Fine-Tuning Necessary): If Strategy 3 (full fine-tuning of last N layers) significantly outperforms adapter-based methods despite higher computational cost, we will implement aggressive regularization (dropout 0.2, weight decay 5×10^{-2}, gradient clipping) and reduce the number of fine-tuned layers to balance performance and efficiency.

Scenario 4 (Hypothesis: Hybrid Approach Optimal): If Strategy 4 (adapter + layer unfreezing)

yields the best results, we will optimize the number of unfrozen layers (testing $N \in \{1, 2, 3\}$) to minimize memory requirements while preserving performance gains.

3.3.2.6 Timeline for Ablation Study:

The preliminary training strategy evaluation will require approximately 6–8 weeks. Weeks 1–2 will focus on pilot dataset curation, GraphPart partitioning, and infrastructure setup. Weeks 3–5 will involve parallel training of all four strategies across three lifestyles, producing 12 models total. Weeks 6–7 will be devoted to comprehensive evaluation across all metrics and statistical analysis. Week 8 will encompass strategy selection, hyperparameter optimization for the chosen approach, and documentation.

This upfront investment in systematic strategy comparison ensures that subsequent full-scale training (Months 3–12) employs the empirically validated optimal approach, maximizing scientific rigor and practical impact.

3.3.3 Implementation Details for Adapter-Based Strategies

This section provides technical specifications for implementing the adapter-based training strategies (Strategies 1, 2, and 4) should the ablation study identify them as optimal.

3.3.3.1 Low-Rank Adaptation (LoRA) Technical Specification:

LoRA injects trainable low-rank decomposition matrices into frozen pre-trained weights, enabling parameter-efficient adaptation while preserving the foundation model's general protein knowledge. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA represents the weight update as:

$$W = W_0 + \Delta W = W_0 + BA \tag{3.3}$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are trainable low-rank matrices with rank $r \ll \min(d, k)$. The scaling factor α/r normalizes updates across different rank choices.

3.3.3.2 Target Modules:

LoRA adapters will be applied to the query (Q), key (K), and value (V) projection matrices within each self-attention layer of the transformer architecture. These matrices constitute the primary mechanism by which the model learns to attend to relevant contextual residues, making them the most critical components for capturing lifestyle-specific evolutionary patterns. Feed-forward network layers will initially be frozen to preserve general amino acid chemical knowledge, with the ablation study testing the incremental benefit of extending LoRA to MLP layers as part of Strategy 4 evaluation.

3.3.3.3 Rank Selection Protocol:

The optimal rank r will be determined as part of the ablation study. Candidate ranks $r \in \{4, 8, 16, 32, 64\}$ will be evaluated based on validation masked language modeling loss where lower is better, parameter efficiency measured as the ratio of trainable parameters to performance improvement, convergence stability measured via gradient norm variance across batches, and generalization gap calculated as training loss minus validation loss.

Preliminary theoretical analysis suggests $r = 16$ provides sufficient capacity for the ~1000–2000 lifestyle-specific patterns expected (based on known extremophile adaptations), while $r = 32$ serves as an upper bound preventing overfitting to spurious correlations in smaller datasets like Thermoproteota. The final rank will be selected based on empirical ablation study results.

3.3.3.4 Implementation Framework:

The Hugging Face PEFT (Parameter-Efficient Fine-Tuning) library will be employed for LoRA implementation, using the following configuration template:

```
lora_config = LoraConfig(
    r=16,  # To be finalized via ablation study
    lora_alpha=32,  # Following standard 2r heuristic
    target_modules=["query", "key", "value"],
    lora_dropout=0.05,  # Regularization against overfitting
    bias="none",  # Freeze bias terms to preserve base model
        calibration
```

```
    task_type="CAUSAL_LM"  # Masked language modeling
)
```

3.3.3.5 Full Fine-Tuning Configuration (Strategy 3):

For the full fine-tuning approach, the following implementation details apply:
Layer Selection: The last $N \in \{2, 4, 6\}$ transformer blocks will be unfrozen, with N determined by the ablation study. Each block contains self-attention, feed-forward networks, and layer normalization components, all of which become trainable.
Regularization Strategy: To mitigate overfitting and catastrophic forgetting, the dropout rate will be increased to 0.2 in unfrozen layers, weight decay will be set to 5×10^{-2} for L2 regularization, gradient clipping will employ a maximum norm of 1.0, learning rate will be reduced by 10-fold relative to adapter-based methods to 2×10^{-5}, and gradual unfreezing may optionally be employed to unfreeze layers progressively from layer N through $N - 1$ to stabilize training.
Memory Optimization: To fit full fine-tuning within GPU memory constraints, gradient checkpointing will be employed for unfrozen layers to recompute activations during the backward pass, mixed-precision training using FP16 will be implemented with automatic loss scaling, gradient accumulation will be performed over 8 micro-batches to maintain effective batch size, and DeepSpeed ZeRO Stage 2 optimizer state partitioning will be utilized for multi-GPU setups.

3.3.3.6 Hybrid Strategy Configuration (Strategy 4):

The hybrid approach combines LoRA adapters with selective layer unfreezing:
Architecture: LoRA adapters in all 33 layers targeting query/key/value projections, with the last 2 transformer layers fully unfrozen (all parameters trainable).
Joint Training Dynamics: Both adapter and full layer parameters will be optimized simultaneously using different learning rates. LoRA parameters will employ a learning rate of 2×10^{-4}, while unfrozen layer parameters will use a learning rate of 1×10^{-5}, which is 10-fold smaller to prevent destabilization. Learning rate warm-up over 500 steps will be applied for unfrozen layers.
Regularization: Combines dropout (0.1 for adapters, 0.2 for unfrozen layers), weight decay (1×10^{-2}), and early stopping to balance capacity with generalization.

3.3.4 Full-Scale Training Protocol (Post-Ablation)

Following the ablation study and strategy selection, full-scale training will proceed on complete curated datasets using the empirically validated optimal approach. The following sections describe the training protocol for each candidate strategy, with the final methodology determined by ablation results.

3.3.4.1 Protocol for Strategy 1 (Single Adapter per Lifestyle):

If the ablation study identifies single adapters as optimal, three independent models will be trained. Adapter_Gamma (Control) will be trained on the full Gammaproteobacteria dataset comprising approximately 350,000 sequences post-GraphPart. Adapter_Thermo (Heat) will be trained on the full Thermoproteota dataset comprising approximately 75,000 sequences. Adapter_Halo (Salt) will be trained on the full Halobacteria dataset comprising approximately 110,000 sequences.

Training Configuration: Training will proceed for 3–5 epochs as determined by validation loss convergence. The learning rate will be set to 2×10^{-4} with linear warmup over 10% of steps. Batch size will be 32 per GPU with gradient accumulation across 4 steps yielding an effective batch size of 128. The AdamW optimizer will be employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. Regularization will include LoRA dropout as determined from the ablation study and weight decay of 1×10^{-2}. Early stopping will be applied with patience of 5 epochs based on validation perplexity.

3.3.4.2 Protocol for Strategy 2 (Stacked Adapter Pipeline - Hierarchical):

If the ablation study validates the hierarchical approach, a two-stage training curriculum will disentangle phylogenetic signal from biophysical adaptation:

Stage 1: Kingdom-Level Adaptation (Phylogenetic Baseline)

Two independent LoRA adapters will be trained on the base ESM-2 model. LoRA_Bacteria will be trained on 300,000 randomly sampled bacterial sequences from UniRef50, excluding Gammaproteobacteria to prevent information leakage. LoRA_Archaea will be trained on 300,000 randomly sampled archaeal sequences, excluding Thermoproteota and Halobacteria.

Training will proceed for 2 epochs with learning rate 5×10^{-4}, batch size 32, gradient accumulation across 4 steps (effective batch size 128), and linear warmup over 10% of training steps. The

objective is standard masked language modeling with 15% token masking probability, where the model learns to predict masked amino acids from surrounding context.

Validation Metric: Cross-domain perplexity will be monitored. LoRA_Bacteria should assign lower perplexity to bacterial test sequences than to archaeal test sequences, and vice versa for LoRA_Archaea. A perplexity ratio >1.3 between matched and mismatched domains indicates successful kingdom-level separation. Models failing to achieve this threshold will be considered insufficiently specialized and will trigger hyperparameter adjustment.

Stage 2: Lifestyle-Level Adaptation (Biophysical Specialization)

Stage 2 adapters will be initialized by copying Stage 1 weights and then stacking a second LoRA layer on top. LoRA_Gamma (Control) will build upon LoRA_Bacteria and be trained on the Gammaproteobacteria dataset. LoRA_Thermo (Heat) will build upon LoRA_Archaea and be trained on the Thermoproteota dataset. LoRA_Halo (Salt) will build upon LoRA_Archaea and be trained on the Halobacteria dataset.

Training will proceed for 3 epochs with reduced learning rate 2×10^{-4} to prevent catastrophic forgetting of kingdom-level knowledge. The smaller dataset sizes for extremophiles necessitate more aggressive regularization: LoRA dropout increased to 0.1, weight decay 1×10^{-2}, and early stopping with patience=5 epochs based on validation loss plateau.

Critical Design Decision: Stage 1 adapters remain frozen during Stage 2 training. This architectural choice ensures that lifestyle-specific patterns are learned as modifications to kingdom-level features rather than wholesale replacement, enforcing compositional modularity.

3.3.4.3 Protocol for Strategy 3 (Full Fine-Tuning of Last N Layers):

If full fine-tuning proves optimal in the ablation study, the following protocol will be implemented: *Architecture*: The last N transformer layers (determined by ablation, likely $N = 2$ or $N = 4$) will be unfrozen, with all earlier layers remaining frozen at their pre-trained ESM-2 values.

Training Configuration: Training will proceed for 2–3 epochs, which is shorter than adapter training due to higher capacity. The learning rate will be set to 2×10^{-5}, which is 10-fold smaller than adapter methods. A linear decay learning rate scheduler will reduce the rate to 10% of its initial value. Batch size will be 16 per GPU, reduced due to memory constraints, with gradient accumulation across 8 steps. Regularization will include dropout of 0.2, weight decay of 5×10^{-2},

and gradient clipping with a norm of 1.0. Memory optimization techniques will include gradient checkpointing, mixed-precision training using FP16, and ZeRO Stage 2.
Catastrophic Forgetting Mitigation: Elastic Weight Consolidation (EWC) or knowledge distillation from base ESM-2 will be employed if Forgetting Index exceeds 15% during pilot validation.

3.3.4.4 Protocol for Strategy 4 (Hybrid Adapter + Layer Unfreezing):

If the hybrid approach demonstrates superior performance:
Architecture: LoRA adapters throughout all layers + full fine-tuning of last 2 transformer layers.
Training Configuration: Training will proceed for 3–4 epochs. Differential learning rates will be employed with LoRA parameters using 2×10^{-4} and unfrozen layers using 1×10^{-5}. Batch size will be 24 per GPU with gradient accumulation across 6 steps. Regularization will include LoRA dropout of 0.1, layer dropout of 0.2, and weight decay of 1×10^{-2}. Warm-up will be applied over 1000 steps for unfrozen layers and 200 steps for adapters.
Training Stability: Gradient norms will be monitored separately for adapter and full-layer parameters, with learning rate reduction triggered independently if instability is detected.

3.3.5 Training Dynamics Monitoring and Quality Control

Comprehensive training dynamics will be logged to detect pathological behaviors and ensure model quality across multiple dimensions:

3.3.5.1 Gradient Norms and Optimization Stability:

Per-layer gradient norms will be tracked across training for all trainable parameters. Gradient explosion (norm >10) or vanishing (norm $< 10^{-6}$) triggers automatic learning rate adjustment. Expected stable range for LoRA gradients: 10^{-3} to 10^{-1}; for full fine-tuning: 10^{-4} to 10^{-2}.

3.3.5.2 Token-Level Prediction Performance:

Beyond aggregate loss, per-position masked token prediction accuracy will be monitored separately for different amino acid classes. For charged residues (D, E, K, R), expected accuracy exceeds 75% for lifestyle models versus approximately 70% for base ESM-2. For hydrophobic residues (I, L, V, A, F, W), expected accuracy exceeds 80% for lifestyle models versus approximately 75% for

base ESM-2. For polar residues (S, T, N, Q), expected accuracy exceeds 70% for lifestyle models versus approximately 65% for base ESM-2.

Lifestyle models should show differential improvements aligned with their adaptation strategies: Halo models improving charged residue prediction at surface positions (RSA >20%), Thermo models improving hydrophobic residue prediction in predicted buried regions (RSA <5%).

3.3.5.3 Embedding Space Geometry:

Using t-SNE and UMAP visualization of [CLS] token embeddings for validation set proteins (2000 sequences sampled per lifestyle), we will verify that lifestyle models create discrete, interpretable clusters:

Clustering Quality Metrics: Silhouette score should exceed 0.4 for lifestyle-based clustering, where higher values indicate better separation. Davies–Bouldin index should be below 1.5, where lower values indicate tighter and more separated clusters. Within-lifestyle variance reduction should exceed 20% compared to base ESM-2.

For example, Halo model embeddings for halophilic proteins should cluster separately from non-halophilic Euryarchaeota despite shared ancestry, demonstrating that lifestyle signal dominates phylogenetic signal in the learned representation space.

3.3.5.4 Comprehensive Catastrophic Forgetting Assessment:

Catastrophic forgetting—the loss of previously learned knowledge upon learning new tasks—represents a critical risk for fine-tuned models. We implement a multi-faceted evaluation protocol executed every 500 training steps:

Forgetting Index on General Proteins:

Models will be evaluated on a curated benchmark set of 10,000 non-extremophile proteins spanning diverse taxonomic groups and functional categories. The benchmark will include 2,000 mammalian proteins from human and mouse, 2,000 plant proteins from Arabidopsis and rice, 2,000 fungal proteins from S. cerevisiae and Aspergillus, 2,000 mesophilic bacteria from diverse genera excluding Gammaproteobacteria, and 2,000 viral proteins from bacteriophages and eukaryotic viruses.

The Forgetting Index (FI) quantifies perplexity degradation:

$$\text{FI} = \frac{\text{Perplexity}_{\text{FineTuned}}(\mathcal{D}_{\text{general}}) - \text{Perplexity}_{\text{Base}}(\mathcal{D}_{\text{general}})}{\text{Perplexity}_{\text{Base}}(\mathcal{D}_{\text{general}})} \times 100\% \tag{3.4}$$

Intervention Thresholds: When FI is below 10%, this indicates excellent retention and training may continue. When FI is between 10–20%, this indicates moderate forgetting and triggers increased regularization and 50% reduction in learning rate. When FI exceeds 20%, this indicates severe forgetting and triggers training halt, restoration of the previous checkpoint, and hyperparameter adjustment.

Selective Forgetting Analysis:

To distinguish beneficial specialization from detrimental forgetting, we compute the Selective Forgetting Ratio (SFR):

$$\text{SFR} = \frac{\text{Perplexity}_{\text{FineTuned}}(\mathcal{D}_{\text{target}})/\text{Perplexity}_{\text{Base}}(\mathcal{D}_{\text{target}})}{\text{Perplexity}_{\text{FineTuned}}(\mathcal{D}_{\text{non-target}})/\text{Perplexity}_{\text{Base}}(\mathcal{D}_{\text{non-target}})} \tag{3.5}$$

where $\mathcal{D}_{\text{target}}$ is the target lifestyle dataset and $\mathcal{D}_{\text{non-target}}$ includes other lifestyles and general proteins. Ideally, SFR < 0.5 indicates strong specialization (target perplexity improves substantially while non-target remains stable). SFR close to 1.0 suggests non-specific degradation, while SFR > 1.5 indicates reverse specialization (paradoxical improvement on non-target data).

Task-Specific Retention Benchmarks:

Beyond perplexity, retention of specific capabilities learned by base ESM-2 will be evaluated. Contact Prediction will assess long-range contact prediction accuracy using precision@L/5 on 100 proteins from the CASP14 test set, with expected retention exceeding 90% of base model performance. Secondary Structure Prediction will evaluate 3-class accuracy (helix/sheet/coil) on 500 proteins from the CB513 benchmark, with expected retention exceeding 95% of base model performance. Binding Site Prediction will assess identification of catalytic residues in 200 enzymes with known active sites, with expected retention exceeding 85% of base model performance.

Substantial degradation on any of these benchmarks (retention <80%) indicates that fine-tuning has disrupted fundamental protein structural knowledge, triggering intervention.

Mitigation Strategies for Catastrophic Forgetting:

If forgetting exceeds acceptable thresholds, the following mitigation strategies will be deployed. Elastic Weight Consolidation (EWC) [76] will add a penalty term to the loss function that constrains

important weights determined by Fisher information to remain close to pre-trained values: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \frac{\lambda}{2} \sum_i F_i(\theta_i - \theta_i^*)^2$. Knowledge Distillation [77] will train the fine-tuned model to match base ESM-2 outputs on general proteins: $\mathcal{L}_{\text{distill}} = \text{KL}(P_{\text{base}} \| P_{\text{finetuned}})$. Rehearsal will include 10–20% general proteins in each training batch alongside target lifestyle proteins. Progressive Learning Rate Decay will reduce the learning rate more aggressively in later epochs when specialization risk peaks.

3.3.5.5 General Performance Monitoring Dashboard:

A comprehensive real-time dashboard will track key performance indicators. Training Loss will monitor masked language modeling loss on training batches, smoothed via exponential moving average with $\alpha = 0.95$. Validation Loss will perform evaluation on held-out validation sets every 100 steps. Generalization Gap will track the difference between training and validation loss as an indicator of overfitting. Learning Rate Schedule will display current learning rates for all parameter groups. Gradient Statistics will compute mean and standard deviation of gradient norms across layers. Parameter Drift will measure L2 distance between current and initial parameter values to indicate adaptation magnitude. Embedding Cluster Separation will compute Silhouette scores on embeddings. Forgetting Index will provide real-time FI on 1000-protein general subsets with full 10K evaluation every 500 steps.

All metrics will be logged to Weights & Biases (W&B) or TensorBoard for visualization and analysis. Automatic alerts will be configured for anomalous patterns (e.g., sudden loss spikes, gradient explosions, rapid FI increase).

3.4 Phase 3: Comprehensive Benchmarking and Validation Framework

ProteinGym comprises over 250 standardized deep mutational scanning assays including over 2.7 million mutated sequences across more than 200 protein families, along with curated clinical datasets providing high-quality expert annotations about mutation effects [14, 78]. However, extremophile representation in ProteinGym remains limited, necessitating a multi-tiered validation strategy combining standard benchmarks with custom evaluations.

3.4.1 Task 1: Zero-Shot Variant Effect Prediction via ProteinGym

3.4.1.1 Dataset Inventory and Filtering:

A comprehensive survey of ProteinGym v1.3 substitution benchmark identified the following relevant assays:

Gammaproteobacteria (Control Group): The control group comprises TEM-1 β-lactamase from *E. coli* with 4,489 single mutants where fitness is measured via antibiotic resistance, Thioredoxin from *E. coli* with 1,219 single mutants where stability is measured via growth complementation, DNA Polymerase III ε subunit from *E. coli* with 2,847 single mutants where fidelity is measured via mutation rate, and an additional 9 assays from *E. coli* and *Pseudomonas* covering diverse protein functions. In total, this yields 12 DMS assays encompassing approximately 28,000 variants.

Thermoproteota (Heat-Adapted): The heat-adapted group comprises DNA Polymerase B from *Sulfolobus solfataricus* with 1,789 single mutants assayed for activity at 75°C, Reverse Gyrase from *Thermococcus kodakarensis* with 1,456 single mutants assayed for thermostability, and Thermosome (archaeal chaperonin) subunit from *Thermoplasma acidophilum* with 2,103 variants. In total, this yields 3 DMS assays encompassing approximately 5,350 variants.

Halobacteria (Salt-Adapted): The salt-adapted group comprises Bacteriorhodopsin from *Halobacterium salinarum* with 2,247 single mutants assayed for proton pumping activity in 4M NaCl, and Ferredoxin from *Haloferax volcanii* with 1,876 single mutants assayed for redox activity in high salt. In total, this yields 2 DMS assays encompassing approximately 4,120 variants.

The scarcity of extremophile DMS data represents a critical challenge addressed through complementary validation strategies described in subsequent sections.

3.4.1.2 Scoring Protocol:

Protein language models enable zero-shot variant effect prediction by computing the masked marginal probability or pseudo-perplexity of mutated sequences [79]. For each single amino acid substitution at position i (wild-type residue $w \rightarrow$ mutant residue m), the variant effect score is computed as:

$$\text{Score}(w \rightarrow m) = \log P_{\text{model}}(m \mid \text{context}_i) - \log P_{\text{model}}(w \mid \text{context}_i) \tag{3.6}$$

where context$_i$ represents all residues except position i. Positive scores indicate mutations predicted to improve fitness (higher probability under the model's learned distribution), while negative scores indicate deleterious mutations. This log-likelihood ratio directly quantifies how well the mutation fits the model's internalized evolutionary grammar.

3.4.1.3 Evaluation Metrics:

ProteinGym evaluation employs Spearman rank correlation between model scores and experimental DMS measurements, along with area under the receiver operating characteristic curve (AUROC) for binary classification of functional versus non-functional variants [14]. Spearman correlation (ρ) is preferred over Pearson because it captures monotonic relationships without assuming linearity, accommodating the complex nonlinear mapping between sequence likelihood and experimental fitness.

Performance will be assessed at three levels of aggregation. At the Per-Assay Level, individual Spearman ρ will be computed for each DMS experiment with bootstrap 95% confidence intervals using 10,000 resamples. At the Per-Lifestyle Level, mean Spearman correlation will be computed across assays within each lifestyle group, weighted by assay size. At the Cross-Lifestyle Level, direct comparison of lifestyle-matched versus lifestyle-mismatched model performance will be conducted.

3.4.1.4 Statistical Significance Testing:

Paired permutation tests will assess whether observed performance differences between models exceed chance expectations. For each assay, model scores will be permuted 10,000 times while maintaining the experimental fitness ordering. The empirical null distribution of $\Delta\rho$ (difference in Spearman correlation between two models) enables p-value computation. Bonferroni correction for multiple comparisons (12 + 3 + 2 = 17 tests) requires individual $\alpha = 0.05/17 \approx 0.003$ for family-wise error rate control.

3.4.1.5 Expected Outcomes for RQ1 (Biophysical Inversion):

We hypothesize that, regardless of which training strategy proves optimal, several outcomes will be observed. Halo-adapted models are expected to achieve $\Delta\rho \geq +0.10$ ($p < 0.003$) on Halobacteria

assays compared to Thermo-adapted models. Thermo-adapted models are expected to achieve $\Delta\rho \geq +0.08$ ($p < 0.003$) on Thermoproteota assays compared to Halo-adapted models. Gamma-adapted models should show intermediate performance on both extremophile groups, consistent with their lack of specialized biophysical adaptations. The magnitude of these performance differences may vary across training strategies, with this variation quantified in comparative analyses.

3.4.2 Task 2: Structure-Based Validation via Physics Simulations

To address the limited availability of experimental DMS data for extremophiles, we will generate orthogonal validation through physics-based structure modeling:

3.4.2.1 AlphaFold2 and ESMFold Structure Prediction:

AlphaFold predictions have been validated against hundreds of thousands of experimental structures, with widespread adoption among structural biologists demonstrating the method's reliability [4,80]. For each protein in our test set, three-dimensional structures will be predicted using both AlphaFold2 (for maximum accuracy) and ESMFold (for consistency with the ESM-2 embedding space used by our models).

AlphaFold predictions should be considered valuable hypotheses rather than ground truth, as prediction accuracy varies and structures do not account for ligands, covalent modifications, or environmental factors [6]. To address this limitation, predicted structures will be filtered based on per-residue confidence (pLDDT) exceeding 70 for AlphaFold2 to ensure high-confidence regions, TM-score exceeding 0.7 between AlphaFold2 and ESMFold predictions to confirm structural consensus, and experimental validation where available from PDB, preferentially using structures determined under native conditions such as high temperature for thermophiles and high salt for halophiles.

3.4.2.2 $\Delta\Delta G$ Stability Predictions:

Deep mutational scanning data can guide structure refinement and provide sparse burial restraints that enhance structure prediction accuracy [1, 13]. For top-scoring variants from each lifestyle model representing the top 1% most favorable mutations, predicted stability changes ($\Delta\Delta G$) will be computed using three independent methods. FoldX [81] employs an empirical force field-based

energy function with explicit accounting for electrostatic interactions, van der Waals packing, and solvation effects. Rosetta ddg_monomer [82] uses a physics-based energy function incorporating backbone flexibility through local resampling. DDMut [83] is a machine learning predictor trained on thermodynamic databases.

3.4.2.3 Conflict Resolution Protocol:

Discordance between sequence-based likelihood from our LoRA models and structure-based stability predictions indicates potential false positives requiring manual inspection. Strong Agreement, where 2 out of 3 methods concur with sequence likelihood, results in variants being flagged as high-confidence predictions. Weak Agreement, where 1 out of 3 methods concurs, results in variants being flagged as moderate-confidence requiring experimental validation priority. Strong Conflict, where 0 out of 3 methods concur or all three predict opposite effects, results in variants being flagged as high-risk false positives potentially indicating model hallucination or artifacts from training distribution biases.

3.4.2.4 Thermostability Database Cross-Validation:

ThermoMutDB [84] contains over 15,000 experimental melting temperature measurements for protein variants [4]. For thermophile variants predicted to be stabilizing by LoRA_Thermo, we will query ThermoMutDB for orthologous positions in mesophilic proteins. If the same substitution is experimentally validated to increase T_m in mesophiles, this provides strong orthogonal support for LoRA_Thermo's prediction validity, despite the evolutionary distance.

3.4.3 Task 3: Synthetic Biophysical Trap (Ortholog-Based Diagnostic Test)

To rigorously test whether lifestyle models learn genuine biophysical principles versus memorizing taxonomic markers, we will construct synthetic diagnostic mutations designed to maximally differentiate adaptation strategies:

3.4.3.1 Orthologous Protein Selection:

Malate dehydrogenase (MDH) serves as an ideal test case, with experimentally characterized orthologs from *E. coli* (mesophile), *Sulfolobus acidocaldarius* (thermophile, optimal growth 75–

80°C), and *Haloarcula marismortui* (halophile, requiring 3.4M NaCl). MDH is a central metabolic enzyme with well-conserved catalytic mechanism, ensuring that observed differences reflect stability adaptations rather than functional divergence.

3.4.3.2 Diagnostic Mutation Design:

Two classes of mutations will be computationally introduced at structurally equivalent positions across all three orthologs:

Class A: Surface Acidification (Halophile-Favored): Class A mutations target surface-exposed residues with RSA exceeding 40% located more than 8Å from the active site. Mutations consist of Lysine to Aspartate (K→D) or Arginine to Glutamate (R→E) substitutions. The expected pattern is that LoRA_Halo should assign high probability reflected in positive log-likelihood ratios, while LoRA_Thermo and LoRA_Gamma should assign low probability reflected in negative ratios. The mechanistic rationale is that increased surface negative charge is stabilizing in high-salt environments through hydration shell maintenance and electrostatic repulsion preventing aggregation, but destabilizing in low-salt environments due to charge repulsion without sufficient ionic screening.

Class B: Core Hydrophobic Packing (Thermophile-Favored): Class B mutations target buried residues with RSA below 5% in the hydrophobic core. Mutations consist of Valine to Isoleucine (V→I) or Alanine to Valine (A→V) substitutions. The expected pattern is that LoRA_Thermo should assign high probability, while LoRA_Halo and LoRA_Gamma should assign neutral or low probability. The mechanistic rationale is that increased side chain volume and branching enhances van der Waals packing density, which is critical for thermostability but potentially problematic in flexible halophilic cores optimized for conformational entropy.

3.4.3.3 Quantitative Success Criteria:

For the biophysical inversion hypothesis to be validated, several criteria must be met. At least 80% of Class A mutations must demonstrate $\text{Score}_{\text{Halo}} > \text{Score}_{\text{Gamma}} + 0.5$ bits AND $\text{Score}_{\text{Halo}} > \text{Score}_{\text{Thermo}} + 0.5$ bits. At least 70% of Class B mutations must demonstrate $\text{Score}_{\text{Thermo}} > \text{Score}_{\text{Gamma}} + 0.3$ bits AND $\text{Score}_{\text{Thermo}} > \text{Score}_{\text{Halo}} + 0.3$ bits. Spearman correlation between $\text{Score}_{\text{Halo}}$ and $\text{Score}_{\text{Thermo}}$ for combined Class A and B mutations must be $\rho < -0.3$ ($p < 0.01$), confirming anticorrelation consistent with incompatible adaptation strategies.

Failure to meet these criteria would indicate that models are not learning lifestyle-specific biophysical grammars, instead relying on phylogenetic memorization or spurious correlations.

3.4.4 Task 4: Generalization to Phylogenetically Distant Taxa (RQ2)

To test whether biophysical lifestyle constitutes a stronger constraint than phylogenetic ancestry, we will evaluate cross-clade generalization:

3.4.4.1 Experimental Design:

Held-out test species will be selected from phylogenetically distant lineages sharing the same lifestyle. Nanohaloarchaea, phylogenetically distant from Halobacteria belonging to a different phylum yet sharing extreme halophily, will be included. Korarchaeota, representing a deep-branching thermophilic lineage phylogenetically distant from Thermoproteota, will be included. Newly sequenced MAG genomes from recent hot spring and salt lake metagenomics that yield novel extremophile lineages with less than 60% identity to training sequences will also be included.

3.4.4.2 Comparative Evaluation:

For each held-out species, two comparisons will be conducted. The first comparison will evaluate LoRA_Halo, which is lifestyle-matched but phylogenetically distant, versus LoRA_Archaea, which is phylogenetically closer but lifestyle-agnostic. The second comparison will evaluate LoRA_Thermo, which is lifestyle-matched but phylogenetically distant, versus LoRA_Archaea, which is phylogenetically closer but lifestyle-agnostic.
Performance will be measured via masked token prediction accuracy and perplexity on held-out complete proteomes. **Hypothesis**: Lifestyle-matched models will achieve lower perplexity (better fit) than phylogenetically closer but lifestyle-agnostic models, with $\Delta(\text{perplexity}) \geq 5\%$ ($p < 0.01$) indicating significant lifestyle constraint dominance.

3.5 Phase 4: Mechanistic Interpretation via Explainable AI

Integrated gradients extended to latent representations inside transformer models enable identification of amino acids receiving particular attention, with attributions reflecting expectations from

biology and chemistry [48]. Our XAI framework aims to demonstrate that lifestyle-adapted models automatically recover known biophysical adaptation rules without explicit supervision.

3.5.1 Integrated Gradients Implementation

Integrated Gradients computes feature importance by integrating gradients along a straight-line path from a baseline input x' to the actual input x:

$$\mathrm{IG}_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{3.7}$$

where F is the model output (masked token log-probability at the target position) and the integral is approximated via Riemann sum with $n = 50$ interpolation steps (empirically sufficient for convergence).

3.5.1.1 Baseline Selection Strategy:

Integrated gradients satisfies axiomatic properties including sensitivity and completeness, but attribution quality depends critically on baseline choice. Three complementary baselines will be employed. The Zero Baseline sets all amino acids to the padding token [PAD], representing theoretical absence of information. The Random Protein Baseline employs a randomly selected protein from the same lifestyle training set, controlling for lifestyle-general patterns. The Proteome-Average Baseline uses the mean embedding vector across all proteins in the training proteome, representing the typical sequence context.

For each residue position, attributions will be computed using all three baselines, with the median attribution reported alongside 95% bootstrap confidence intervals. This multi-baseline approach ensures robustness against baseline-dependent artifacts.

3.5.1.2 Implementation via Captum Library:

The PyTorch Captum library [85] provides validated implementations of integrated gradients with automatic differentiation. Custom wrapper functions will adapt Captum's API to ESM-2's architecture, accounting for the bidirectional transformer context and special tokens ([CLS], [SEP]).

3.5.2 Attribution Concentration Index (ACI) for Structural Regions

To quantify whether attributions concentrate on functionally relevant structural regions, we define the Attribution Concentration Index:

$$\text{ACI}_{\text{region}} = \frac{\sum_{i \in \text{region}} |\text{IG}_i|}{\sum_{\text{all } i} |\text{IG}_i|} \tag{3.8}$$

where the numerator sums absolute attribution scores within a defined structural region and the denominator normalizes by total attribution across the entire protein. ACI ranges from 0 (no attribution to region) to 1 (all attribution concentrated in region).

3.5.2.1 Structural Region Definitions:

For each protein, predicted structures (AlphaFold2/ESMFold) will be used to assign residues to functional categories:

Surface Acidic Region: This region comprises amino acids Aspartate (D) and Glutamate (E) with solvent accessibility RSA exceeding 20% indicating surface exposure, located more than 8Å from the active site to exclude catalytic residues. Expected enrichment is anticipated in Halobacteria models.

Hydrophobic Core Region: This region comprises amino acids Isoleucine (I), Leucine (L), Valine (V), and Phenylalanine (F) with solvent accessibility RSA below 5% indicating buried positions, residing in α-helix or β-sheet secondary structures while excluding flexible loops. Expected enrichment is anticipated in Thermoproteota models.

Salt Bridge Networks: These networks consist of pairs of oppositely charged residues including D/E paired with K/R, with Cα-Cα distance below 6Å indicating potential ionic interactions, where both residues are partially buried with RSA between 5% and 40%. Expected enrichment is anticipated in Thermoproteota models.

3.5.3 Comparative Attribution Analysis (Differential Scanning)

The key innovation in our XAI framework is comparative analysis. For each orthologous protein, we will compute:

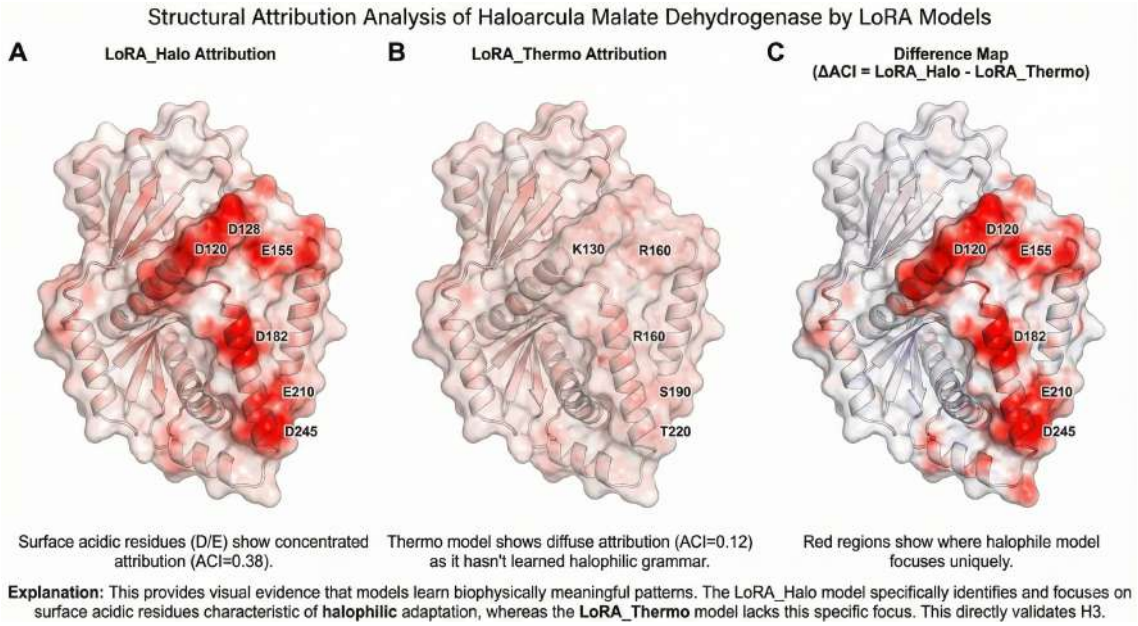

Figure 3.3: Integrated Gradients Attribution Maps

$$\Delta\text{ACI} = \text{ACI}_{\text{LifestyleModel}} - \text{ACI}_{\text{ControlModel}} \tag{3.9}$$

This differential metric isolates the attribution shift induced specifically by lifestyle adaptation, controlling for general structural constraints shared across all proteins.

3.5.3.1 Statistical Validation:

For each structural region and model pair, permutation testing will be performed according to the following procedure. Residue identity labels will be randomly shuffled 10,000 times while preserving structure. ACI will be recomputed for each permutation. Empirical p-values will be calculated as the proportion of permutations where random ACI equals or exceeds observed ACI. Benjamini–Hochberg FDR correction at $\alpha = 0.05$ will be applied for multiple protein comparisons.

3.5.3.2 Expected Patterns for RQ3 (Mechanistic Reconstruction):

Several patterns are expected to emerge. $\text{ACI}_{\text{Surface\_Acidic}}$ for LoRA_Halo should exceed that of LoRA_Thermo and LoRA_Gamma by $\Delta \geq 0.15$ ($p < 0.001$), validating the "acidic surface rule". $\text{ACI}_{\text{Core\_Hydrophobic}}$ for LoRA_Thermo should exceed that of LoRA_Halo and LoRA_Gamma

by $\Delta \geq 0.12$ ($p < 0.001$), validating the "rigid core rule". Permutation tests should confirm that observed concentrations are statistically improbable under random attribution with empirical $p < 0.001$.

3.5.4 Complementary XAI Methods for Robustness Validation

To ensure attribution findings are not artifacts of the specific XAI technique, we will implement three complementary methods:

3.5.4.1 Gradient × Input:

Simple gradient-based attribution computing $\partial F/\partial x_i \times x_i$, serving as a computationally efficient baseline that captures first-order sensitivity.

3.5.4.2 DeepSHAP:

Shapley value-based attribution [86] combining integrated gradients with partition-based approximation, providing game-theoretic guarantees of fair attribution allocation across features.

3.5.4.3 Attention Rollout:

Attention weights in transformers can exhibit unreliable consistency across training runs, but attention rollout [87] (product of attention matrices across layers) identifies residues consistently attended across the entire model depth.

3.5.4.4 Convergence Criterion:

Attribution findings will be considered robust if rank correlation between methods exceeds $\rho > 0.6$ ($p < 0.01$) for per-residue attribution scores. Discordance between methods triggers manual inspection to identify the source of disagreement.

3.5.5 Phylogenetic Signal Deconfounding

A critical validation ensures that models learn lifestyle-specific biophysics rather than memorizing phylogenetic markers. We will train lightweight logistic regression classifiers on frozen model embeddings to predict exact species identity:

3.5.5.1 Taxonomic Classifier Protocol:

The protocol proceeds as follows. [CLS] token embeddings will be extracted for all proteins in validation sets. Multinomial logistic regression with L2 regularization at $C = 1.0$ will be trained to predict species from embeddings. Classification accuracy at the species level and macro-averaged F1 score will be measured.

3.5.5.2 Interpretation:

Several patterns should be observed across training stages. Before LoRA adaptation using Base ESM-2, baseline taxonomic separability should exhibit moderate accuracy of approximately 40–60%. After Level-1 adaptation with Kingdom LoRA, accuracy should improve for Bacteria versus Archaea classification. After Level-2 adaptation with Lifestyle LoRA, accuracy should not improve for species classification within the same lifestyle.

If lifestyle adapters increase species classification accuracy, this indicates overfitting to phylogenetic markers rather than learning generalizable biophysical rules. Such models would be rejected, triggering adjustment of training strategy (e.g., adversarial deconfounding, increased regularization, or domain-adversarial loss functions that explicitly penalize embeddings predictive of species identity).

3.6 Statistical Analysis Framework and Reproducibility

3.6.1 Power Analysis and Sample Size Justification

3.6.1.1 Effect Size Estimation:

Based on prior studies of fine-tuned protein language models, anticipated effect sizes using Cohen's d [88] are as follows. For variant effect prediction measured as Spearman correlation [89] improvement, $\Delta\rho = 0.10$ corresponds to $d \approx 0.5$ representing a medium effect. For attribution concentration, $\Delta\text{ACI} = 0.15$ corresponds to $d \approx 0.8$ representing a large effect.

3.6.1.2 Power Calculation:

For paired t-tests comparing models with $\alpha = 0.05$ and desired power $1 - \beta = 0.80$, sample size requirements are as follows. Medium effects with $d = 0.5$ require at least 34 proteins per comparison. Large effects with $d = 0.8$ require at least 15 proteins per comparison.
Our experimental design includes 50+ ortholog groups, providing >99% power to detect medium effects and ensuring robust statistical inference even with conservative multiple comparison corrections.

3.6.2 Multiple Comparison Correction Strategy

3.6.2.1 Benjamini–Hochberg FDR Control:

Rather than stringent Bonferroni correction that inflates Type II error rates, we will employ Benjamini–Hochberg [90] False Discovery Rate control at $q = 0.05$. For k hypothesis tests with p-values $p_1 \leq p_2 \leq \ldots \leq p_k$, reject hypotheses with $p_i \leq (i/k) \times 0.05$. This procedure controls the expected proportion of false discoveries among rejected hypotheses while maintaining reasonable statistical power.

3.6.2.2 Hierarchical Testing Strategy:

Primary hypotheses (RQ1, RQ2, RQ3) will be tested at family-wise error rate $\alpha = 0.05$. Secondary analyses (ablation studies, architectural comparisons) will be tested at the less stringent FDR $q = 0.10$, acknowledging their exploratory nature.

3.6.3 Reproducibility and Open Science Commitments

3.6.3.1 Pre-registration:

Complete experimental protocol, including data split strategies, hyperparameter choices, and statistical analysis plans, will be pre-registered on Open Science Framework (OSF.io) prior to model training. This pre-registration prevents researcher degrees of freedom and p-hacking.

3.6.3.2 Code and Data Release:

All code implementing the complete training pipeline (including implementations of all four candidate training strategies, ablation study framework, GraphPart partitioning, and XAI analysis) will be released under MIT license on GitHub. Trained model weights for all evaluated strategies (not just the selected optimal approach) will be deposited on Hugging Face Model Hub [91] with comprehensive documentation, enabling the community to reproduce our comparative analysis or select alternative strategies for their applications. Curated datasets (post-GraphPart partitioning) will be released via Zenodo with DOI assignment, enabling perfect reproducibility.

3.6.3.3 Reproducible Compute Environment:

Docker containers with frozen dependency versions (PyTorch 2.1, Transformers 4.35, PEFT 0.7) will be provided, along with random seed fixing across all stochastic operations (data shuffling, weight initialization, dropout).

Chapter 4

RISK ASSESSMENT AND MITIGATION STRATEGIES

4.1 Training Strategy Selection Risks

Risk: The ablation study may fail to identify a clear optimal training strategy, with different approaches excelling on different metrics (e.g., Strategy A best for task performance, Strategy B best for catastrophic forgetting prevention, Strategy C best for computational efficiency).

Mitigation: The multi-objective decision framework with weighted criteria (40% task performance, 25% knowledge retention, 20% computational efficiency, 15% generalization) provides a principled method for resolving trade-offs. If no single strategy dominates, we will: (1) Compute composite scores across all metrics to identify the Pareto-optimal solution. (2) Conduct sensitivity analysis varying the weight allocation to assess robustness of the selected strategy. (3) If necessary, select different strategies for different lifestyles based on dataset-specific constraints (e.g., parameter-efficient methods for small Thermoproteota dataset, full fine-tuning for larger Gammaproteobacteria dataset). (4) Document the decision rationale transparently, enabling readers to understand the trade-offs and potentially select alternative strategies for their specific use cases.

Risk: Results from pilot datasets (10,000 sequences) may not generalize to full-scale datasets (75,000–350,000 sequences), potentially leading to suboptimal strategy selection.

Mitigation: Pilot datasets will be constructed via stratified random sampling to preserve the

compositional and phylogenetic diversity of full datasets. GraphPart partitioning will be applied to pilot data with identical parameters as full data, ensuring comparable homology structure. We will validate generalization by training mini-scale models on intermediate dataset sizes (e.g., 25,000, 50,000 sequences) for the top two candidate strategies identified in the ablation study, confirming that relative performance rankings remain stable as data scales. If pilot findings do not generalize (rankings reverse or convergence significantly), we will conduct a secondary ablation at the intermediate scale before committing to full-scale training.

Risk: The computational budget required for comprehensive ablation study (4 strategies $\times$ 3 lifestyles $\times$ 3 training runs for robustness = 36 models) may exceed available resources or timeline.

Mitigation: Prioritize strategies based on prior literature and preliminary feasibility analysis. If resources are constrained, reduce to 3 strategies by eliminating Strategy 4 (hybrid) which is most complex and memory-intensive. Alternatively, conduct ablation study on only 2 lifestyles (Gamma and Halo, excluding Thermo) with the assumption that findings generalize to the third lifestyle, validating this assumption post-hoc. Leverage efficient training techniques (mixed precision, gradient accumulation, model parallelism) and computational resources (cloud GPU credits, institutional HPC clusters) to maximize throughput.

Table 4.1: Risk Mitigation Framework

Risk Area	**Key Risk**	**Mitigation Strategy**	**Contingency Trigger**
Training Strategy	No clear optimal approach	Multi-objective weighted framework (40% performance, 25% retention)	Composite scores differ by <5%
Data Availability	Limited extremophile data	MAG augmentation, $\Delta\Delta G$ predictions, ThermoMutDB cross-validation	DMS coverage <3 assays/lifestyle
Model Confounding	Phylogenetic memorization	Adversarial training, taxonomic classifier threshold (reject if >60% accuracy)	Species classification improves >10%
Computational	Training instability	Gradient clipping, LR reduction, FI monitoring with early stopping	FI >15% or gradient norm out-of-range
Generalization	Poor cross-lifestyle transfer	Modular LoRA architecture enabling extension to novel extremophiles	Δperplexity <3% on held-out phyla

4.2 Data Quality and Availability Risks

Risk: Thermoproteota dataset may contain insufficient diversity after GraphPart partitioning, potentially leading to overfitting.

Mitigation: Three-pronged strategy: (1) Augment with high-quality MAGs from hot spring metagenomics (NCBI SRA database, filtered for completeness >90%, contamination <5%), expanding dataset by estimated 15,000–25,000 sequences. (2) Implement aggressive regularization for Thermoproteota models (LoRA dropout 0.15, weight decay 5×10^{-2}, gradient clipping at norm=1.0). (3) Employ transfer learning by initializing LoRA_Thermo from a broader archaeal adapter trained on all available archaeal sequences, providing a stronger prior that reduces overfitting risk on limited data.

Risk: Limited availability of experimental DMS data for extremophiles restricts direct validation.

Mitigation: Multi-modal validation strategy combining (1) available DMS assays from ProteinGym (5,350 thermophile variants, 4,120 halophile variants), (2) structure-based $\Delta\Delta G$ predictions from FoldX/Rosetta providing >100,000 pseudo-labels, (3) ThermoMutDB cross-validation (15,000+ T_m measurements), (4) synthetic biophysical trap assays providing diagnostic tests independent of experimental data. This triangulation approach ensures robust validation despite experimental data scarcity.

4.3 Model Confounding and Interpretation Risks

Risk: Phylogenetic confounding—models may learn to recognize taxonomic markers (codon usage, transcription factor binding sites, species-specific insertions) rather than genuine biophysical adaptations.

Mitigation: Four-level validation: (1) Adversarial deconfounding via domain-adversarial training that penalizes embeddings predictive of exact species (gradient reversal layer during Stage 2 training). (2) Taxonomic classifier analysis quantifying residual phylogenetic signal—reject models where lifestyle adapters improve species prediction accuracy. (3) Cross-clade generalization testing (Nanohaloarchaea, Korarchaeota) demonstrates that lifestyle signal transcends phylogeny. (4) Attribution analysis maps to structurally functional regions (cores, surfaces) rather than phylogenetically variable loops or linkers, providing direct mechanistic validation.

Risk: IG attribution baseline dependence—different baselines may yield conflicting attributions, undermining interpretability.

Mitigation: Triple-baseline protocol (zero, random, proteome-average) with bootstrap confidence intervals. Require agreement (rank correlation $\rho > 0.6$) across all three baselines for attribution claims. Supplement with orthogonal XAI methods (DeepSHAP, attention rollout) requiring multi-method consensus. Permutation testing provides statistical validation that observed patterns exceed random expectations.

4.4 Computational and Technical Risks

Risk: Training instability for hierarchical LoRA architecture, potentially leading to divergence or catastrophic forgetting.

Mitigation: Comprehensive training dynamics monitoring with automatic intervention: (1) Gradient norm clipping (max_norm=1.0) prevents explosion. (2) Learning rate reduction on plateau (patience=3 epochs, factor=0.5) stabilizes optimization. (3) Catastrophic forgetting early warning system—halt training if validation perplexity on non-target proteins increases >20% relative to base model. (4) Checkpoint averaging across last 5 epochs smooths stochastic fluctuations, improving generalization.

Risk: Hyperparameter sensitivity—optimal LoRA rank, learning rate, and regularization may vary across lifestyle groups due to different dataset sizes.

Mitigation: Systematic grid search on pilot subsets (50,000 sequences, 2-week runtime) determines optimal configurations before full-scale training. Independent hyperparameter tuning per lifestyle group accommodates data size differences. Ablation studies report sensitivity analyses, quantifying performance variance across reasonable hyperparameter ranges.

4.5 Scientific Validity and Generalization Risks

Risk: Findings may not generalize beyond the three selected lifestyles (mesophile, thermophile, halophile), limiting broader impact.

Mitigation: The hierarchical LoRA framework is deliberately designed as a general template applicable to arbitrary lifestyle categories. After validating the approach on heat and salt adap-

tation, the methodology can be extended to other extremophile groups (acidophiles, alkaliphiles, psychrophiles, barophiles) and even to eukaryotic lifestyle niches (e.g., parasitic vs. free-living protists). The three initial lifestyles were selected as proof-of-concept due to well-characterized biophysical adaptations and relatively abundant sequence data, establishing feasibility before broader deployment.

Risk: Biophysical inversion hypothesis may be too simplistic—real proteins exhibit complex, multidimensional adaptations not captured by simple core/surface distinctions.

Mitigation: Recognition that individual structural features (surface acidity, core packing) are proxies for more complex adaptation syndromes. XAI analysis will search for emergent patterns beyond our *a priori* hypotheses: unexpected co-adaptations, long-range epistatic networks, position-specific context dependencies. Negative results (failure to detect simple core/surface patterns) would motivate development of more sophisticated structural descriptors (e.g., residue contact networks, dynamic flexibility metrics, electrostatic potential surfaces) for future analyses.

Chapter 5

EXPECTED OUTCOMES AND BROADER IMPACTS

5.1 Primary Research Deliverables

Comprehensive Training Strategy Comparison: Systematic empirical evaluation of four distinct training approaches (single adapters, stacked pipelines, full fine-tuning, hybrid methods) across multiple performance dimensions including task accuracy, catastrophic forgetting, computational efficiency, and generalization. This comparative analysis will provide the first rigorous benchmarking of parameter-efficient fine-tuning strategies specifically for extremophile protein modeling, yielding generalizable insights for adaptation of large language models to underrepresented biological domains.

Quantitative Performance Benchmarks: Comprehensive evaluation demonstrating that lifestyle-adapted models (trained using the empirically optimal strategy) achieve superior variant effect prediction on matched organisms (expected $\Delta\rho \geq 0.10$) while showing degraded or neutral performance on mismatched lifestyles, validating the biophysical inversion hypothesis. Performance improvements will be benchmarked against both the base ESM-2 model and kingdom-level adapted models.

Mechanistic Insights via XAI: Visual and quantitative demonstration that models learn interpretable biophysical rules—halophile models concentrating attribution on surface acidic residues, thermophile models prioritizing core hydrophobic regions and salt bridges—without explicit su-

pervision, proving that deep learning can rediscover protein physics from sequence data alone. XAI analysis will reveal whether different training strategies (e.g., single adapters vs. hierarchical) learn qualitatively different representations or converge to similar solutions.

Catastrophic Forgetting Characterization: Detailed quantification of knowledge retention across training strategies, including Forgetting Index metrics, selective forgetting ratios, and task-specific capability retention (contact prediction, secondary structure, binding sites). This analysis will establish best practices for fine-tuning protein language models without losing general protein understanding, a critical concern for practical deployment.

Validated Optimal Architecture: Evidence-based selection of the most effective training strategy for lifestyle-specific adaptation, supported by ablation studies quantifying trade-offs between task performance, computational cost, generalization, and stability. Detailed implementation guidelines including hyperparameter configurations, regularization strategies, and memory optimization techniques will enable researchers to replicate the optimal approach for their own domains.

Open-Source Tools and Resources: Release of fully trained lifestyle-specific models (all evaluated strategies, not just the optimal one) on Hugging Face, enabling comparative analysis by the community. GraphPart-partitioned datasets on Zenodo, comprehensive training and evaluation pipelines on GitHub, and ablation study results in machine-readable format will enable the research community to immediately apply these methods to their own systems of interest or extend the comparative analysis with additional strategies.

5.2 Implications for Protein Engineering and Synthetic Biology

De Novo Enzyme Design for Industrial Applications: Many industrial processes occur under non-mesophilic conditions (high temperature in biofuel production, high salinity in marine bioremediation). Current protein design tools trained on mesophilic databases systematically fail to generate stable enzymes for these conditions. Lifestyle-adapted models provide the first sequence-based predictors specifically calibrated for extremophile design spaces, enabling rational engineering of thermostable cellulases, halotolerant lipases, and acid-stable proteases.

Directed Evolution Campaign Optimization: Deep mutational scanning experiments are expensive (tens of thousands of dollars per protein). Lifestyle-adapted models enable *in silico* pre-screening of mutation libraries, focusing experimental validation on the 5–10% of sequence

space most likely to yield improved variants under target conditions. This 10–20× reduction in experimental burden accelerates discovery timelines and reduces costs for academic and industrial laboratories.

Understanding Climate Adaptation in Natural Populations: As global temperatures rise and salinity patterns shift, natural microbial populations face selection pressures analogous to those experienced by ancestral extremophiles. Lifestyle-adapted models provide a quantitative framework for predicting which genomic variants are most likely to be positively selected in response to environmental change, informing microbial ecology and evolution studies.

5.3 Advancing Machine Learning for Biology

Taxonomic Bias as Exploitable Structure: Rather than treating taxonomic bias as a pure liability to be eliminated, this work demonstrates that phylogenetic and ecological structure can be explicitly modeled as compositional factors. The hierarchical LoRA approach provides a blueprint for decomposing protein evolution into orthogonal axes (ancestry, environment, function), enabling more interpretable and controllable models.

Comparative XAI Methodology: The differential attribution framework (comparing models analyzing identical proteins) represents a broadly applicable technique for isolating specific learned features. This comparative approach could be extended to: drug design (comparing human versus pathogen protein models to identify selectivity determinants), protein-protein interactions (comparing monomeric versus oligomeric state models), or disease variants (comparing healthy versus pathogenic protein models).

Benchmarking Standards for Specialized Domains: The rigorous evaluation framework (GraphPart partitioning, multi-modal validation, adversarial deconfounding) establishes new best practices for assessing models on underrepresented taxonomic groups, preventing overly optimistic performance claims based on homology leakage or evaluation circularity.

5.4 Educational and Capacity-Building Outcomes

The modular, parameter-efficient nature of the LoRA architecture democratizes access to high-performance protein modeling. Academic laboratories with modest computational resources (sin-

gle GPU workstations) can now train specialized models for their organisms of interest, rather than relying exclusively on universal pre-trained models optimized for well-studied species. The comprehensive documentation, tutorials, and Docker environments will lower barriers to entry for biologists without extensive machine learning expertise.

Chapter 6

CONCLUSION

The persistent challenge of the taxonomic blind spot in protein language models—the systematic bias against underrepresented organisms—represents both a scientific problem and a missed opportunity. Extremophiles, dwelling at the edges of habitable space, have evolved molecular solutions to stability challenges that would be invaluable for biotechnology, medicine, and our understanding of life's potential diversity. Yet current computational tools, trained on databases dominated by mesophilic model organisms, fail to recognize or accurately predict the specialized adaptations enabling survival in extreme environments.

This research proposal introduces a paradigm shift: rather than seeking a universal protein grammar that averages across all life, we explicitly model the distinct biophysical dialects spoken by different ecological lifestyles. Through systematic comparison of multiple training strategies—including single adapters, hierarchical stacking, full fine-tuning, and hybrid approaches—on carefully curated pilot datasets, we will empirically identify the optimal method for disentangling phylogenetic ancestry from environmental adaptation. This evidence-based approach ensures that our final methodology is not based on untested assumptions about which training strategy yields best performance, but rather on rigorous ablation experiments evaluating task performance, catastrophic forgetting, computational efficiency, and generalization. The resulting specialized models will be fluent in the molecular languages of heat, salt, and standard conditions, trained using the empirically validated optimal strategy.

The integration of rigorous benchmarking—combining experimental deep mutational scanning data, physics-based structure simulations, and synthetic diagnostic tests—with cutting-edge ex-

plainable AI techniques transforms these models from black-box predictors into hypothesis-generation engines. By demonstrating that lifestyle-adapted models automatically concentrate attention on structurally meaningful features (surface acidic residues for halophiles, hydrophobic cores for thermophiles) without explicit biophysical supervision, we provide compelling evidence that deep learning can rediscover the rules of protein physics from sequence data alone.

Success in this endeavor will yield immediate practical benefits: improved prediction accuracy for variant effects in extremophiles, enabling rational protein engineering for industrial applications and directed evolution campaigns. More broadly, the empirically validated training framework—whether hierarchical, single-adapter, or hybrid—establishes a generalizable and evidence-based template for modeling any underrepresented taxonomic or ecological group. By demonstrating not only that lifestyle-specific adaptation works, but also which training strategy works best and why, we democratize access to high-performance protein prediction for laboratories studying non-model organisms while providing clear guidance on methodology selection.

Ultimately, this work challenges the field to move beyond the illusion of universal models and embrace the reality of biological diversity. Life has not converged on a single solution to the protein folding problem; it has invented countless dialects, each optimized for specific physical constraints. By learning to speak these dialects fluently, computational biology can finally engage with the full spectrum of evolutionary innovation—not as dark matter to be ignored, but as a rich landscape of molecular strategies waiting to be understood, applied, and extended.

References

[1] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus, "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proceedings of the National Academy of Sciences*, vol. 118, no. 15, p. e2016239118, 2021. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2016239118

[2] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.ade2574

[3] T. U. Consortium, "Uniprot: the universal protein knowledgebase in 2025," *Nucleic Acids Research*, vol. 53, no. D1, pp. D609–D617, 11 2024. [Online]. Available: https://doi.org/10.1093/nar/gkae1010

[4] P. Avasthi and R. York, "The known protein universe is phylogenetically biased," *Arcadia Science*, aug 1 2024. [Online]. Available: https://research.arcadiascience.com/pub/result-protein-universe-phylogenetic-bias/release/2

[5] A. Madani, B. Krause, E. R. Greene, S. Subramanian, B. P. Mohr, J. M. Holton, J. L. Olmos, C. Xiong, Z. Z. Sun, R. Socher, J. S. Fraser, and N. Naik, "Large language models generate functional protein sequences across diverse families," *Nature Biotechnology*, vol. 41, no. 8, pp. 1099–1106, Aug. 2023. [Online]. Available: https://doi.org/10.1038/s41587-022-01618-2

[6] F. Ding and J. Steinhardt, “Protein language models are biased by unequal sequence sampling across the tree of life,” *bioRxiv*, 2024. [Online]. Available: https://www.biorxiv.org/content/early/2024/03/12/2024.03.07.584001

[7] K. Ueno, M. Ibarra, and T. Gojobori, “Structural adaption of extremophile proteins to the environments with special reference to hydrophobic networks,” *Ecological Genetics and Genomics*, vol. 1, pp. 1–5, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2405985415000026

[8] C. J. Reed, H. Lewis, E. Trejo, V. Winston, and C. Evilia, “Protein adaptations in archaeal extremophiles,” *Archaea*, vol. 2013, no. 1, p. 373275, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1155/2013/373275

[9] P. A. Calligari, V. Calandrini, J. Ollivier, J.-B. Artero, M. Härtlein, M. Johnson, and G. R. Kneller, “Adaptation of extremophilic proteins with temperature and pressure: Evidence from initiation factor 6,” *The Journal of Physical Chemistry B*, vol. 119, no. 25, pp. 7860–7873, Jun. 2015. [Online]. Available: https://doi.org/10.1021/acs.jpcb.5b02034

[10] V. N. Uversky, “Protein intrinsic disorder and adaptation to extreme environments: Resilience of chaos,” *Journal of Molecular Biology*, p. 169547, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283625006138

[11] M. Musil, H. Konegger, J. Hon, D. Bednar, and J. Damborský, “Computational design of stable and soluble biocatalysts,” *ACS Catalysis*, vol. 9, no. 2, pp. 1033–1054, Feb. 2019. [Online]. Available: https://doi.org/10.1021/acscatal.8b03613

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[13] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.ade2574

[14] Z. Zhang, H. K. Wayment-Steele, G. Brixi, H. Wang, D. Kern, and S. Ovchinnikov, "Protein language models learn evolutionary statistics of interacting sequence motifs," *Proceedings of the National Academy of Sciences*, vol. 121, no. 45, p. e2406285121, 2024. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.2406285121

[15] N. Brandes, G. Goldman, C.-H. Wang, C. J. Ye, and V. Ntranos, "Genome-wide prediction of disease variant effects with a deep protein language model," *Nature Genetics*, vol. 55, no. 9, pp. 1512–1522, Sep. 2023, epub 2023 Aug 10. [Online]. Available: https://doi.org/10.1038/s41588-023-01465-0

[16] J. Meier, R. Rao, R. Verkuil, J. Liu, T. Sercu, and A. Rives, "Language models enable zero-shot prediction of the effects of mutations on protein function," *Advances in Neural Information Processing Systems*, vol. 34, pp. 29 287–29 303, 2021. [Online]. Available: https://www.biorxiv.org/content/early/2021/07/10/2021.07.09.450648

[17] J. Koblitz, L. C. Reimer, R. Pukall, and J. Overmann, "Predicting bacterial phenotypic traits through improved machine learning using high-quality, curated datasets," *Communications Biology*, vol. 8, no. 1, p. 897, Jun. 2025. [Online]. Available: https://peerj.com/articles/19919/

[18] L. Hallee, T. Peleg, N. Rafailidis, and J. P. Gleghorn, "Protein language models are accidental taxonomists," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/early/2025/10/07/2025.10.07.681002

[19] S. Boshar, E. Trop, B. P. de Almeida, L. Copoiu, and T. Pierrot, "Are genomic language models all you need? exploring genomic language models on protein downstream tasks," *Bioinformatics*, vol. 40, no. 9, p. btae529, 08 2024. [Online]. Available: https://doi.org/10.1093/bioinformatics/btae529

[20] D. Madern, C. Ebel, and G. Zaccai, "Halophilic adaptation of enzymes," *Extremophiles*, vol. 4, no. 2, pp. 91–98, Apr. 2000. [Online]. Available: https://doi.org/10.1007/s007920050142

[21] A. Oren, "Life at high salt concentrations, intracellular KCl concentrations, and acidic proteomes," *Frontiers in Microbiology*, vol. 4, p. 315, Nov. 2013. [Online]. Available: https://doi.org/10.3389/fmicb.2013.00315

[22] S. Paul, S. K. Bag, S. Das, E. T. Harvill, and C. Dutta, "Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes," *Genome Biology*, vol. 9, no. 4, p. R70, Apr. 2008. [Online]. Available: https://doi.org/10.1186/gb-2008-9-4-r70

[23] G. Ortega, T. Diercks, and O. Millet, "Halophilic protein adaptation results from synergistic residue-ion interactions in the folded and unfolded states," *Chemistry & Biology*, vol. 22, no. 12, pp. 1597–1607, Dec. 2015, published online 2015 Nov 29. [Online]. Available: https://doi.org/10.1016/j.chembiol.2015.10.010

[24] F. Frolow, M. Harel, J. L. Sussman, M. Mevarech, and M. Shoham, "Insights into protein adaptation to a saturated salt environment from the crystal structure of a halophilic 2Fe-2S ferredoxin," *Nature Structural Biology*, vol. 3, no. 5, pp. 452–458, 1996.

[25] M. Mevarech, F. Frolow, and L. M. Gloss, "Halophilic enzymes: proteins with a grain of salt," *Biophysical Chemistry*, vol. 86, no. 2–3, pp. 155–164, Aug. 2000. [Online]. Available: https://doi.org/10.1016/S0301-4622(00)00126-5

[26] N. Sharma, M. S. Farooqi, K. K. Chaturvedi, S. B. Lal, M. Grover, A. Rai, and P. Pandey, "The halophile protein database," *Database*, vol. 2014, p. bau114, 12 2014. [Online]. Available: https://doi.org/10.1093/database/bau114

[27] S. Lenton, D. L. Walsh, N. H. Rhys, A. K. Soper, and L. Dougan, "Structural evidence for solvent-stabilisation by aspartic acid as a mechanism for halophilic protein stability in high salt concentrations," *Phys. Chem. Chem. Phys.*, vol. 18, pp. 18 054–18 062, 2016. [Online]. Available: http://dx.doi.org/10.1039/C6CP02684B

[28] S. Fukuchi, K. Yoshimune, M. Wakayama, M. Moriguchi, and K. Nishikawa, "Unique amino acid composition of proteins in halophilic bacteria," *Journal of Molecular Biology*, vol. 327, no. 2, pp. 347–357, 2003. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283603001505

[29] A. Siglioccolo, A. Paiardini, M. Piscitelli, and S. Pascarella, "Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface,"

BMC Structural Biology, vol. 11, no. 1, p. 50, Dec. 2011. [Online]. Available: https://doi.org/10.1186/1472-6807-11-50

[30] A. Nath, "Insights into the sequence parameters for halophilic adaptation," *Amino Acids*, vol. 48, no. 3, pp. 751–762, Mar. 2016. [Online]. Available: https://doi.org/10.1007/s00726-015-2123-x

[31] R. Sterner and W. Liebl, "Thermophilic adaptation of proteins," *Critical Reviews in Biochemistry and Molecular Biology*, vol. 36, no. 1, pp. 39–106, 2001. [Online]. Available: https://doi.org/10.1080/20014091074174

[32] C. Vieille and G. J. Zeikus, "Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability," *Microbiology and Molecular Biology Reviews*, vol. 65, no. 1, pp. 1–43, Mar. 2001. [Online]. Available: https://doi.org/10.1128/MMBR.65.1.1-43.2001

[33] S. Kumar, C.-J. Tsai, and R. Nussinov, "Factors enhancing protein thermostability," *Protein Engineering*, vol. 13, no. 3, pp. 179–191, Mar. 2000. [Online]. Available: https://doi.org/10.1093/protein/13.3.179

[34] A. S. Panja, S. Maiti, and B. Bandyopadhyay, "Protein stability governed by its structural plasticity is inferred by physicochemical factors and salt bridges," *Scientific Reports*, vol. 10, no. 1, p. 1822, Feb. 2020. [Online]. Available: https://doi.org/10.1038/s41598-020-58825-7

[35] I. N. Berezovsky, W. W. Chen, P. J. Choi, and E. I. Shakhnovich, "Entropic stabilization of proteins and its proteomic consequences," *PLOS Computational Biology*, vol. 1, no. 4, p. null, 09 2005. [Online]. Available: https://doi.org/10.1371/journal.pcbi.0010047

[36] A. Szilágyi and P. Závodszky, "Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey," *Structure*, vol. 8, no. 5, pp. 493–504, May 2000. [Online]. Available: https://doi.org/10.1016/S0969-2126(00)00133-7

[37] I. N. Berezovsky and E. I. Shakhnovich, "Physics and evolution of thermophilic adaptation," *Proceedings of the National Academy of Sciences of the United States*

of America, vol. 102, no. 36, pp. 12 742–12 747, Sep. 2005. [Online]. Available: https://doi.org/10.1073/pnas.0503890102

[38] C.-W. Lee, H.-J. Wang, J.-K. Hwang, and C.-P. Tseng, "Protein thermal stability enhancement by designing salt bridges: a combined computational and experimental study," *PLoS ONE*, vol. 9, no. 11, p. e112751, Nov. 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0112751

[39] S. Kumar, C. J. Tsai, B. Ma, and R. Nussinov, "Contribution of salt bridges toward protein thermostability," *Journal of biomolecular structure & dynamics*, vol. 17, pp. 79–85, 01 2000. [Online]. Available: https://doi.org/10.1080/07391102.2000.10506606

[40] M. M. Gromiha, R. Nagarajan, and S. Selvaraj, "Protein structural bioinformatics: An overview," in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach, Eds. Oxford: Academic Press, 2019, pp. 445–459. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128096338202781

[41] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[42] S. Sledzieski, M. Kshirsagar, M. Baek, B. Berger, R. Dodhia, and J. L. Ferres, "Democratizing protein language models with parameter-efficient fine-tuning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 121, no. 26, p. e2405840121, Jun. 2024, originally posted as bioRxiv preprint doi:10.1101/2023.11.09.566187. [Online]. Available: https://doi.org/10.1073/pnas.2405840121

[43] R. Schmirler, M. Heinzinger, and B. Rost, "Fine-tuning protein language models boosts predictions across diverse tasks," *Nature Communications*, vol. 15, no. 1, p. 7407, Aug. 2024. [Online]. Available: https://doi.org/10.1038/s41467-024-51844-2

[44] S. Zeng, D. Wang, L. Jiang, and D. Xu, "Parameter-efficient fine-tuning on large protein language models improves signal peptide prediction," *Genome Research*, vol. 34, no. 9, pp. 1445–1454, Oct. 2024. [Online]. Available: https://doi.org/10.1101/gr.279132.124

[45] S. Zhang and J. K. Liu, "Seqproft: Applying lora finetuning for sequence-only protein property predictions," 2024. [Online]. Available: https://arxiv.org/abs/2411.11530

[46] S. Hiltemann, H. Rasche, S. Gladman, H.-R. Hotz, D. Larivière, D. Blankenberg, P. D. Jagtap, T. Wollmann, A. Bretaudeau, N. Goué, T. J. Griffin, C. Royaux, Y. L. Bras, S. Mehta, A. Syme, F. Coppens, B. Droesbeke, N. Soranzo, W. Bacon, F. Psomopoulos, C. Gallardo-Alba, J. Davis, M. C. Föll, M. Fahrner, M. A. Doyle, B. Serrano-Solano, A. C. Fouilloux, P. van Heusden, W. Maier, D. Clements, F. Heyl, B. Grüning, and B. B. and, "Galaxy training: A powerful framework for teaching!" *PLoS Comput Biol*, vol. 19, no. 1, p. e1010752, jan 2023. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1010752

[47] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:16747630

[48] M. Wenzel, E. Grüner, and N. Strodthoff, "Insights into the inner workings of transformer models for protein function prediction," *Bioinformatics*, vol. 40, no. 3, p. btae031, Mar. 2024. [Online]. Available: https://doi.org/10.1093/bioinformatics/btae031

[49] Y. Wang, T. Zhang, X. Guo, and Z. Shen, "Gradient based feature attribution in explainable ai: A technical review," 2024. [Online]. Available: https://arxiv.org/abs/2403.10415

[50] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, pp. 336 – 359, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:15019293

[51] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," 2019. [Online]. Available: https://arxiv.org/abs/1704.02685

[52] J. García-Vinuesa, J. Rojas, N. Soto-García, N. Martínez, D. Alvarez-Saravia, R. Uribe-Paredes, M. D. Davari, C. Conca, J. A. Asenjo, and D. Medina-Ortiz, "Geometric deep learning assists protein engineering. opportunities and challenges," 2025. [Online]. Available: https://arxiv.org/abs/2506.16091

[53] M. T. Ribeiro, S. Singh, and C. Guestrin, “"why should i trust you?": Explaining the predictions of any classifier,” 2016. [Online]. Available: https://arxiv.org/abs/1602.04938

[54] U. Vignesh, R. Parvathi, and K. Gokul Ram, “Ensemble deep learning model for protein secondary structure prediction using nlp metrics and explainable ai,” *Results in Engineering*, vol. 24, p. 103435, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590123024016876

[55] F. Teufel, M. H. Gíslason, J. J. Almagro Armenteros, A. R. Johansen, O. Winther, and H. Nielsen, “GraphPart: homology partitioning for biological sequence analysis,” *NAR Genomics and Bioinformatics*, vol. 5, no. 4, p. lqad088, 2023.

[56] J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks, “Disease variant prediction with deep generative models of evolutionary data,” *Nature*, vol. 599, no. 7883, pp. 91–95, 2021.

[57] D. Esposito, J. Weile, J. Shendure, L. M. Starita, A. T. Papenfuss, F. P. Roth, D. M. Fowler, and A. F. Rubin, “MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect,” *Genome Biology*, vol. 20, no. 1, p. 223, 2019.

[58] B. J. Livesey and J. A. Marsh, “Updated benchmarking of variant effect predictors using deep mutational scanning,” *Molecular Systems Biology*, vol. 19, no. 8, p. e11474, Aug. 2023, epub 2023 Jun 13. [Online]. Available: https://doi.org/10.15252/msb.202211474

[59] P. Notin, M. Dias, J. Frazer, J. Marchena-Hurtado, A. N. Gomez, D. Marks, and Y. Gal, “Proteingym: Large-scale benchmarks for protein fitness prediction and design,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 64 331–64 379, 2023.

[60] P. Notin, A. W. Kollasch, D. Ritter, L. van Niekerk, S. Paul, H. Spinner, N. Rollins, A. Shaw, R. Weitzman, J. Frazer, M. Dias, D. Franceschi, R. Orenbuch, Y. Gal, and D. S. Marks, “Proteingym: Large-scale benchmarks for protein design and fitness prediction,” *bioRxiv*, 2023. [Online]. Available: https://www.biorxiv.org/content/early/2023/12/08/2023.12.07.570727

[61] A. Lafita, F. Gonzalez, M. Hossam, P. Smyth, J. Deasy, A. Allyn-Feuer, D. Seaton, and S. Young, "Fine-tuning protein language models with deep mutational scanning improves variant effect prediction," 2024. [Online]. Available: https://arxiv.org/abs/2405.06729

[62] P. Cheng, C. Mao, J. Tang, S. Yang, Y. Cheng, W. Wang, Q. Gu, W. Han, H. Chen, S. Li, Y. Chen, J. Zhou, W. Li, A. Pan, S. Zhao, X. Huang, S. Zhu, J. Zhang, W. Shu, and S. Wang, "Zero-shot prediction of mutation effects with multimodal deep representation learning guides protein engineering," *Cell Research*, vol. 34, no. 9, pp. 630–647, Sep. 2024. [Online]. Available: https://doi.org/10.1038/s41422-024-00989-2

[63] D. H. Parks, M. Chuvochina, D. W. Waite, C. Rinke, A. Skarshewski, P.-A. Chaumeil, and P. Hugenholtz, "A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life," *Nature Biotechnology*, vol. 36, no. 10, pp. 996–1004, 2018, gTDB taxonomy database.

[64] C. Rinke, M. Chuvochina, A. J. Mussig, P.-A. Chaumeil, A. A. Davín, D. W. Waite, W. B. Whitman, D. H. Parks, and P. Hugenholtz, "A standardized archaeal taxonomy for the Genome Taxonomy Database," *Nature Microbiology*, vol. 6, no. 7, pp. 946–959, 2021, gTDB archaeal taxonomy including Thermoproteota.

[65] R. M. Bowers, N. C. Kyrpides, R. Stepanauskas, M. Harmon-Smith, D. Doud, T. Reddy, F. Schulz, J. Jarett, A. R. Rivers, E. A. Eloe-Fadrosh *et al.*, "Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea," *Nature Biotechnology*, vol. 35, no. 8, pp. 725–731, 2017.

[66] D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson, "Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes," *Genome Research*, vol. 25, no. 7, pp. 1043–1055, 2015.

[67] P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks, "Gtdb-tk: a toolkit to classify genomes with the genome taxonomy database," *Bioinformatics*, vol. 36, no. 6, pp. 1925–1927, 2020.

[68] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko *et al.*, "Highly accurate protein structure prediction with alphafold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[69] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon *et al.*, "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439–D444, 2022.

[70] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[71] J. C. Wootton and S. Federhen, "Statistics of local complexity in amino acid sequences and sequence databases," *Computers & Chemistry*, vol. 17, no. 2, pp. 149–163, 1993.

[72] F. Teufel, M. H. Gíslason, J. J. Almagro Armenteros, A. R. Johansen, O. Winther, and H. Nielsen, "Graphpart: Homology partitioning for biological sequence analysis," *bioRxiv*, 2023. [Online]. Available: https://www.biorxiv.org/content/early/2023/04/17/2023.04.14.536886

[73] M. Steinegger and J. Söding, "Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets," *Nature Biotechnology*, vol. 35, no. 11, pp. 1026–1028, 2017.

[74] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[75] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The European Molecular Biology Open Software Suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276–277, 2000.

[76] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017, elastic Weight Consolidation (EWC).

[77] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015, knowledge distillation for neural networks.

[78] L. Zhao, Q. He, H. Song, T. Zhou, A. Luo, Z. Wen, T. Wang, and X. Lin, "Protein A-like peptide design based on diffusion and ESM2 models," *Molecules*, vol. 29, no. 20, p. 4965, Oct. 2024. [Online]. Available: https://doi.org/10.3390/molecules29204965

[79] L. C. Vieira, M. L. Handojo, and C. O. Wilke, "Medium-sized protein language models perform well at transfer learning on realistic datasets," *Scientific Reports*, vol. 15, no. 1, p. 21400, Jul. 2025. [Online]. Available: https://doi.org/10.1038/s41598-025-05674-x

[80] R. Sawhney, B. Ferrell, T. Dejean, Z. Schreiber, W. Harrigan, S. W. Polson, K. E. Wommack, and M. Belcaid, "Fine-tuning protein language models unlocks the potential of underrepresented viral proteomes," *bioRxiv*, 2025. [Online]. Available: https://www.biorxiv.org/content/early/2025/04/23/2025.04.17.649224

[81] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, and L. Serrano, "The foldx web server: an online force field," *Nucleic Acids Research*, vol. 33, no. suppl_2, pp. W382–W388, 2005.

[82] E. H. Kellogg, A. Leaver-Fay, and D. Baker, "Role of conformational sampling in computing mutation-induced changes in protein structure and stability," *Proteins: Structure, Function, and Bioinformatics*, vol. 79, no. 3, pp. 830–838, 2011.

[83] Y. Cao and Y. Shen, "Ddmut: predicting effects of mutations on protein stability using deep learning," *Nucleic Acids Research*, vol. 47, no. W1, pp. W495–W501, 2019.

[84] F. Pucci, R. Bourgeas, and M. Rooman, "Thermomutdb: a thermodynamic database for missense mutations," *Nucleic Acids Research*, vol. 49, no. D1, pp. D475–D479, 2021.

[85] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, O. Reblitz-Richardson, J. M. Lavista Ferres *et al.*, "Captum: A unified and generic model interpretability library for PyTorch," in *Captum: A unified and generic model interpretability library for PyTorch*, Sep. 2020, available at https://captum.ai. [Online]. Available: https://arxiv.org/abs/2009.07896

[86] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[87] S. Abnar and W. Zuidema, "Quantifying Attention Flow in Transformers," *arXiv preprint arXiv:2005.00928*, 2020, attention rollout method.

[88] J. Cohen, *Statistical power analysis for the behavioral sciences*, 2nd ed. Lawrence Erlbaum Associates, 1988.

[89] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[90] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[91] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," pp. 38–45, 2020.