Deep Learning based Protein Thermal Stability Prediction

by

Umang Sharma

Under the Supervision of

Dr. N. Arul Murugan

Deep Learning based Protein Thermal Stability Prediction

by

Umang Sharma

Submitted

In partial fulfilment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology, Delhi

July, 2025

Certificate

This is to certify that the thesis titled **Deep Learning based Protein Thermal Stability Prediction** being submitted by **Umang Sharma** to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

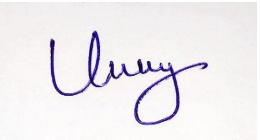
July, 2025

Dr. N. Arul Murugan
Department of Computational Biology,
Indraprastha Institute of Information
Technology Delhi
New Delhi 110020

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me and guided me throughout my M.Tech. thesis work. First and foremost, I would like to thank my esteemed project supervisor, **Dr. N. Arul Murugan**, for providing me this opportunity to work under his guidance. His wisdom and guidance have profoundly shaped my understanding and approach to this project.
I am extremely grateful to **M. Seraj** for his involvement, insightful comments and motivation. Their insights and efforts have been invaluable, and working with them has been a rewarding experience.

Lastly, would like to thank IIIT-Delhi for providing the necessary infrastructure.



Umang Sharma
(MT23239)
Department of Computational Biology,
Indraprastha Institute of Information
Technology Delhi New Delhi 110020

Abstract

The thermal stability of a protein is a fundamental biophysical property that governs its structural integrity, functional efficacy, and practical utility, particularly its shelf-life in therapeutic and industrial applications. Consequently, the ability to accurately predict a protein's melting temperature (Tm) from its primary sequence is a central challenge in protein engineering, synthetic biology, and drug development. While experimental determination of Tm is precise, it is often resource-intensive and low-throughput, creating a bottleneck in large-scale protein design projects. To address this, various computational methods have been developed. The recent success of deep learning, especially large language models in understanding biological sequences, presents new and powerful opportunities to decipher the intricate relationship between a protein's sequence, its structure, and its overall thermal tolerance. This thesis presents a comprehensive investigation into deep learning methodologies for predicting protein thermal stability, performing a systematic comparison of both sequence- and structure-based computational paradigms. Our entire analysis is benchmarked on the extensive Novozymes dataset, comprising 31,384 diverse protein sequences with their experimentally validated melting temperatures.

Our investigation commenced with sequence-based models. We first established a performance baseline using traditional machine learning approaches built on classical biochemical features. An Artificial Neural Network (ANN) trained on physicochemical properties (PCP) extracted using the pFeature library, while a LightGBM model. To move beyond handcrafted features and capture richer contextual information, we leveraged the power of pre-trained protein language models. An ANN utilizing static embeddings extracted from the ProtT5-BERT model showed a marked improvement. The pinnacle of our sequence-based approach was achieved by directly fine-tuning the entire ProtT5-BERT model on our specific thermal stability dataset. This transfer learning strategy resulted in the highest performance, demonstrating that allowing the model to adapt its internal representations to the specific task of stability prediction is superior to using static, pre-computed embeddings.

In parallel, we explored a structure-based approach, predicated on the principle that a protein's three-dimensional fold is a key determinant of its stability. As experimental structures were unavailable for the vast majority of the dataset, we first predicted the 3D structures for all protein sequences using the state-of-the-art ESMFold model. We then developed a Graph Neural Network (GNN), a class of models inherently suited to learning from graph-based data like protein structures, to predict thermal stability. This structure-based model. While valuable, this result was notably lower than our top sequence-based models, potentially reflecting the compounding errors from the initial structure prediction step or the inherent difficulty in learning stability from static structural snapshots.

Our comparative analysis conclusively demonstrates that while both sequence and structural modalities contain stability-related information, the fine-tuned sequence-based language model significantly outperforms all other methods. This work underscores the immense and still-unfolding potential of protein language models to accurately predict complex functional properties like thermal stability directly from sequence data. This provides the scientific community with a powerful and scalable tool for high-throughput in-silico screening, accelerating the design and optimization of novel, hyper-stable proteins for a wide range of applications.

Keywords: Protein Thermal Stability, Deep Learning, ProtT5-BERT, Large Language Models, Graph Neural Network (GNN), Melting Temperature (Tm).

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 The Central Role of Proteins

Proteins are the quintessential workhorses of life, executing a vast and intricate array of functions that are fundamental to the survival and complexity of all living organisms. These macromolecules, constructed from a simple set of 20 amino acid building blocks, are involved in virtually every process within a cell. They act as **enzymes**, catalyzing biochemical reactions with remarkable specificity and speed; as **structural components**, providing shape and support to cells and tissues (e.g., collagen and keratin); as **transporters**, moving essential molecules like oxygen across membranes (e.g., hemoglobin); and as **signaling molecules**, facilitating communication between cells (e.g., hormones and receptors). The specific function of any given protein is inextricably linked to its unique three-dimensional structure, which arises from the precise linear sequence of its amino acids—its primary structure.

This sequence dictates how the protein chain folds into local secondary structures like alpha-helices and beta-sheets, which in turn pack together to form a complex tertiary structure. For many proteins, multiple folded chains assemble into a functional quaternary structure. It is this final, stable 3D conformation that is essential for a protein to perform its designated role. A loss of this structure, a process known as denaturation, results in a loss of function. Consequently, a protein's stability—its ability to maintain its native, functional fold under various conditions—is a property of paramount importance. **Thermal stability**, measured by the melting temperature (Tm), is a critical aspect of this. The Tm is the temperature at which 50% of the protein population is unfolded. A higher Tm indicates greater stability, which is a highly desirable trait for proteins used in industrial processes (e.g., heat-stable enzymes in manufacturing) and as therapeutic agents (e.g., antibodies that require a long shelf-life) [1]. Therefore, understanding and predicting protein thermal stability is not merely an academic exercise; it is a central challenge in biotechnology and medicine, with the potential to unlock the full therapeutic and industrial potential of engineered proteins.

1.2 Determinants of Protein Thermal Stability

The thermal stability of a protein is a complex property governed by a delicate balance of enthalpic and entropic forces. It is not dictated by a single factor but is an emergent property arising from the interplay between the protein's primary amino acid sequence and its final, folded three-dimensional structure. The sequence provides the fundamental instructions, while the structure represents the resulting architecture where stabilizing interactions are realized.

Structural Determinants of Thermal Stability

The folded conformation of a protein allows for a multitude of stabilizing interactions. The efficiency and number of these interactions within the three-dimensional architecture are key to resisting thermal denaturation. A stable protein is, in essence, a tightly

organized and optimized structure. The table below summarizes the key structural factors that influence a protein's melting temperature.

Structural Factor	Effect on Thermal Stability (Tm)
Hydrophobic Core Packing	A densely packed hydrophobic core with minimal cavities is a primary driver of stability. It minimizes the unfavorable interaction between non-polar residues and water, significantly increasing Tm.
Intramolecular Hydrogen Bonding	An extensive network of hydrogen bonds, particularly those forming and stabilizing secondary structures (α-helices and β-sheets), locks the protein into its native fold and raises its Tm.
Disulfide Bonds	Covalent disulfide bridges between cysteine residues act as "molecular staples," drastically reducing the conformational entropy of the unfolded state and thereby providing a substantial boost to the Tm.
Salt Bridges & Ionic Interactions	Electrostatic attractions between oppositely charged residues on the protein's surface or buried within the core can form stabilizing salt bridges, contributing favorably to the energy of the folded state and increasing Tm.
Loops and Structural Flexibility	Regions of high flexibility, such as long surface loops or disordered termini, are often points of weakness where unfolding can initiate. Minimizing such flexibility generally increases stability and raises the Tm.

Proline and Glycine Content	Proline residues introduce a rigid kink in the polypeptide backbone, reducing the flexibility of the unfolded state and often stabilizing loops. Glycine, lacking a side chain, increases local flexibility. Their impact is highly context-dependent.
Secondary Structure Content	Proteins with a high content of well-formed, stable secondary structures (α-helices and β-sheets) tend to be more resistant to thermal unfolding, correlating with a higher Tm.

Sequence-Level Determinants of Thermal Stability

Every structural feature listed above is ultimately encoded by the primary amino acid sequence. The specific choice and order of the 20 amino acids dictate the physicochemical potential of the protein chain to form a stable structure. By analyzing the sequence, one can infer its propensity to be thermally stable [2].

Sequence-Level Factor	Effect on Thermal Stability (Tm)
Hydrophobic Amino Acid Prevalence	A higher proportion of hydrophobic residues (e.g., Leucine, Isoleucine, Valine) provides the raw material for forming a larger, more stable hydrophobic core, generally correlating with a higher Tm.
Strategic Charged Residues	The placement of charged amino acids (e.g., Lysine, Arginine, Aspartate, Glutamate) is critical. Positioning them to form optimal salt bridges can significantly enhance

	stability.
Cysteine Pairings	The presence of cysteine residues is a prerequisite for disulfide bonds. The location of these cysteines in the sequence determines if they will be spatially close enough in the folded structure to form a stabilizing bond.
Proline Distribution	Incorporating proline at key positions, especially within turns or at the start of helices, can enforce a rigid conformation that contributes to overall stability.
Avoidance of Thermolabile Residues	Residues like Asparagine and Glutamine are prone to chemical degradation (deamidation) at elevated temperatures, which can disrupt structure. Sequences with fewer of these residues are often more stable.
Amino Acid Propensity for Structures	Certain amino acids have an intrinsic preference for specific secondary structures (e.g., Alanine for α-helices; Valine for β-sheets). Sequences enriched with these appropriately placed residues form more stable structures.

1.3 Motivation

The ability to engineer highly stable proteins has profound implications across science and industry. In medicine, the efficacy and shelf-life of protein-based therapeutics, such as monoclonal antibodies and insulin, are directly tied to their stability. Enhancing thermal stability can reduce the reliance on a costly and logistically complex cold chain for storage and transportation, thereby increasing access to life-saving medicines globally. In industrial biotechnology, enzymes are used in everything from laundry detergents to food production and biofuel synthesis. Enzymes that can withstand the high temperatures of industrial processes (thermozymes) are more efficient and reusable, leading to greener and more cost-effective manufacturing.

However, the traditional path to creating stable proteins is arduous. Directed evolution and rational design, while powerful, are often slow, expensive, and labor-intensive. Experimental screening of protein variants requires significant resources and is fundamentally low-throughput, creating a major bottleneck. This is where computational prediction becomes indispensable. The explosion of protein sequence data from genomic and metagenomic sequencing has created a vast, largely untapped reservoir of potential proteins. If we could accurately and rapidly predict the stability of a protein from its sequence alone, we could screen millions of candidates in silico, identifying the most promising ones for experimental validation.

This need for high-throughput prediction has driven the development of various computational tools. Early methods relied on statistical potentials or physics-based simulations, which were either too slow or not accurate enough for large-scale screening. Traditional machine learning models, trained on handcrafted physicochemical features, represented a step forward but often failed to capture the complex, non-local interactions that govern stability. The recent revolution in deep learning, particularly the advent of large-scale, pre-trained language models for biology, offers a transformative opportunity. These models, trained on hundreds of millions of diverse protein sequences, learn the fundamental "language" of proteins. They can capture nuanced grammatical and semantic information embedded in the sequence that correlates with structure and function. This thesis is motivated by the hypothesis that these powerful models can be adapted to predict thermal stability with unprecedented accuracy, bypassing the need for explicit structural information and providing a scalable tool to accelerate the future of protein engineering.

1.4 Objectives

The primary goal of this thesis is to conduct a systematic and comprehensive investigation into the application of deep learning for protein thermal stability prediction. We aim to rigorously compare sequence-based and structure-based approaches to determine the most effective paradigm for this critical task. The specific objectives are as follows:

1. **To Evaluate Sequence-Based Prediction Models:** We will develop and benchmark a hierarchy of models that predict thermal stability directly from the primary amino acid sequence. This involves:

    - Establishing a performance baseline using traditional machine learning models (LightGBM, ANN) trained on classical physicochemical features.
    - Investigating the effectiveness of using pre-trained protein language model embeddings (from ProtT5-BERT) as input features for a downstream ANN.
    - Assessing the ultimate potential of sequence-based models by fine-tuning the entire ProtT5-BERT language model on the thermal stability dataset.
2. **To Develop a Structure-Based Prediction Model:** We will explore the predictive power of 3D structural information by:
    - Generating predicted structures for all proteins in our dataset using the state-of-the-art ESMFold model.
    - Developing and training a Graph Neural Network (GNN), which is architecturally suited to learn from molecular structures, to predict ™.
3. **To Conduct a Rigorous Comparative Analysis:** We will systematically compare the performance of all developed models using the coefficient of determination (R^2) as the primary metric. This analysis will elucidate the strengths and weaknesses of each approach and determine whether modern sequence-based language models can outperform methods that rely on explicit structural data.
4. **To Advance In-Silico Protein Engineering:** By identifying the most accurate and efficient predictive methodology, this work aims to contribute a valuable tool to the field, empowering researchers to perform rapid, large-scale virtual screening for thermostable protein variants and accelerate the design of novel proteins for therapeutic and industrial use.

Chapter 2: Literature Review

The computational prediction of protein thermal stability is a crucial goal in biotechnology, essential for engineering robust enzymes and therapeutics. The field has seen a rapid evolution of methods, moving from classical machine learning based on engineered features to sophisticated deep learning architectures that learn directly from sequence and structural data. Numerous studies have explored various techniques to enhance predictive accuracy and provide insight into the determinants of protein stability.

A foundational approach involved applying classical machine learning algorithms to predict a protein's melting temperature (Tm) from a set of "handcrafted" features. A common strategy in these studies was to train models like Random Forests or Support Vector Machines on feature vectors representing the **protein's physicochemical properties (PCPs)**, amino acid composition (AAC), and dipeptide composition (DPC), often calculated with tools like pFeature. While these methods provided valuable early models and achieved moderate success, their performance was inherently limited by the expressive power of the predefined features, which could not fully capture the complex, long-range interactions within a protein sequence that dictate its stability.

To overcome the limitations of manual feature engineering, researchers turned to deep learning. Early deep learning studies utilized architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to automatically learn relevant features directly from the raw amino acid sequence. This represented a significant step forward, as these models could identify complex sequence motifs and patterns correlated with stability without prior biological knowledge. However, the most profound breakthrough came from the application of large-scale, pre-trained Protein Language Models (PLMs), which were inspired by the success of models like BERT and T5 in natural language processing. One notable family of models, ProtTrans, which includes **ProtT5**, was pre-trained on massive, unlabeled databases of hundreds of millions of protein sequences [5]. This pre-training allows the models to learn the fundamental "language" of proteins. One successful application strategy involves using these PLMs as powerful feature extractors; the context-aware embeddings they produce can be fed into a simpler downstream model, such as an Artificial Neural Network (ANN), to predict stability with significantly higher accuracy than traditional methods.

An even more powerful approach, which is central to this thesis, involves **fine-tuning** the entire language model on a specific, labeled dataset. Several studies have demonstrated that this technique leads to state-of-the-art performance. By updating the weights of a model like ProtT5 on a dedicated thermal stability dataset, the model can adapt its vast, generalized knowledge of protein sequences to the specific nuances of the prediction task. This approach has been shown to significantly outperform both traditional machine learning and earlier deep learning models, pushing the boundaries of sequence-based stability prediction.

In parallel with sequence-based modeling, deep learning has also revolutionized structure-based prediction. The primary historical limitation—the scarcity of experimentally determined 3D structures—has been largely overcome by highly accurate deep learning-based structure predictors, most notably AlphaFold2 [4] and more recently,

ESMFold. The availability of high-quality predicted structures for virtually any protein has enabled a new class of predictive models. **Graph Neural Networks (GNNs)** have emerged as a particularly promising architecture for this task. By representing a protein's structure as a graph—where amino acids are nodes and spatial proximities are edges—GNNs can learn complex patterns from the 3D atomic arrangement. Several studies have showcased the potential of GNNs to predict various functional properties from protein structures, affirming their utility and potential for stability prediction. This methodology allows for a complementary analysis, investigating whether explicit structural information can offer predictive power that is orthogonal to that contained within the sequence alone.

Chapter 3: Materials and Methodology

3.1 Overview of the Methodological Framework

This study employs a comprehensive and comparative computational framework to investigate the efficacy of various deep learning models for predicting protein thermal stability. The methodology is divided into two primary paradigms: a sequence-based approach and a structure-based approach. All experiments were conducted using a high-performance computing environment equipped with GPUs to handle the intensive computational demands of deep learning models. The core software stack includes Python (v3.8+) and major libraries such as PyTorch for building and training neural networks, Hugging Face's transformers library for accessing and fine-tuning the ProtT5 model, and PyTorch Geometric for the implementation of the Graph Neural Network. The overall workflow for this research is depicted below and consists of several key stages:

1. **Data Acquisition and Analysis:** Sourcing and performing a thorough statistical analysis of the benchmark dataset.
2. **Data Preprocessing:** Cleaning and preparing the data for each modeling approach, including feature extraction for baseline models and tokenization for language models.
3. **Model Development:** Implementing a range of models, from traditional machine learning baselines to state-of-the-art deep learning architectures (ANN, fine-tuned ProtT5, GNN). For the structure-based approach, this also includes an initial step of predicting 3D structures from sequences using ESMFold.
4. **Model Training and Validation:** Training each model on a designated portion of the dataset and validating its performance using a separate test set.
5. **Performance Evaluation:** Rigorously comparing the models using the coefficient of determination (R^2) to identify the most effective predictive methodology.

3.2 Dataset Information and Analysis

The foundation of any data-driven modeling is a high-quality, large-scale dataset. For this thesis, we utilized the publicly available Novozymes enzyme stability prediction dataset, sourced from the Kaggle platform. This dataset is a widely recognized benchmark for this task, containing thousands of protein sequences and their experimentally determined melting temperatures (Tm), making it an ideal resource for developing and validating robust predictive models. The dataset primarily consists of two columns: “protein_sequence”, which contains the primary amino acid sequence, and “tm”, the target variable representing the melting temperature in degrees Celsius.

Analysis of Melting Temperature (Tm)

The target variable, Tm, exhibits a wide and continuous distribution, which is crucial for training a regression model capable of generalizing across proteins with varying stability profiles. The dataset contains **31,297** valid protein entries with corresponding Tm values. The temperatures range from a low of -1.0°C to a high of 130.0°C, with a mean of

49.16°C. The standard deviation is **14.02°C**, indicating significant variability in stability across the dataset. The median Tm is 48.0°C, and the interquartile range (25% to 75%) lies between 42.1°C and 53.8°C. This distribution demonstrates that while a majority of proteins fall within a typical physiological range, the dataset also includes sufficient examples of both highly unstable and hyper-thermostable proteins [16].

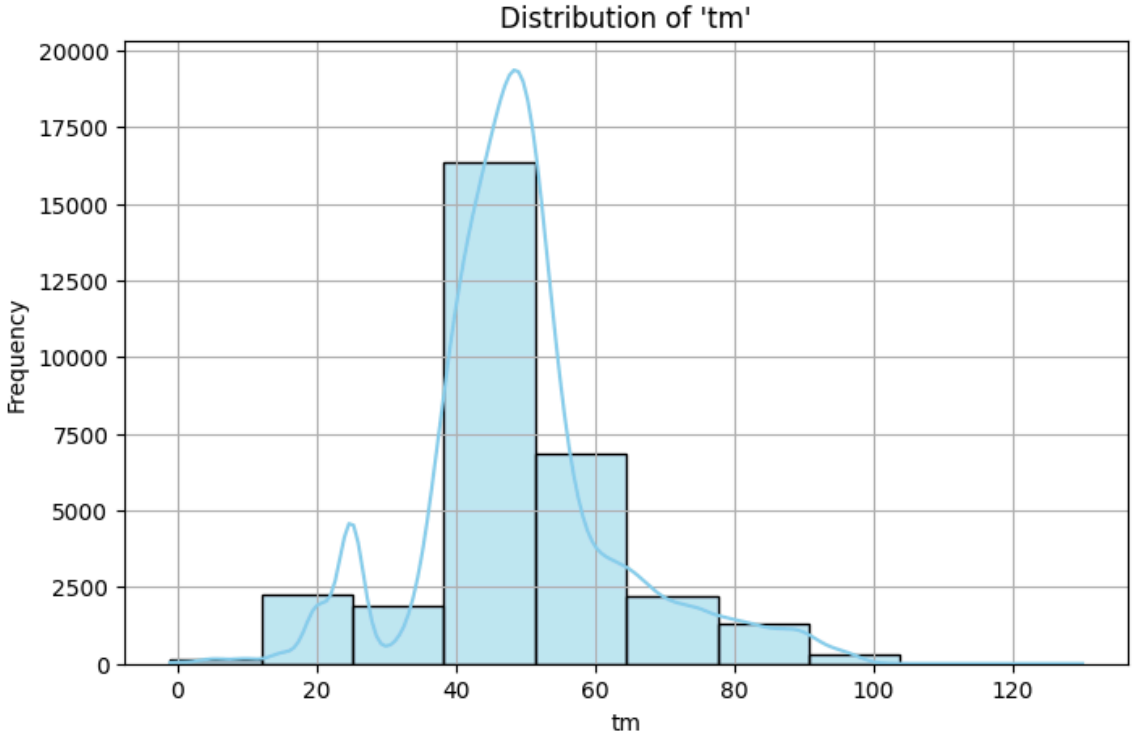


FIGURE 3.1: Distribution of melting temperature (Tm) values in the Novozymes dataset

Analysis of Protein Sequences

The dataset features a diverse collection of proteins in terms of length and amino acid composition.

Protein Length Statistics: The model must be robust to sequences of varying lengths. The dataset presents a significant challenge in this regard, with protein lengths ranging from just **5 amino acids** to a maximum of **3,000 amino acids**. The average protein length is approximately **430 residues**, with a large standard deviation of 364.84, confirming the wide range of protein sizes and domain complexities represented.

Amino Acid Composition: A detailed analysis of the amino acid frequencies provides insight into the biochemical characteristics of the protein universe in this dataset. As expected, hydrophobic and versatile amino acids are highly prevalent. Leucine (L) is the most common residue, making up 9.42% of the total, which aligns with its crucial role in forming the hydrophobic cores of proteins. In contrast, Tryptophan (W) and Cysteine (C), which have unique structural roles, are the least common. The complete percentage

composition is provided below. This diverse chemical makeup ensures that the models are trained on a representative sample of protein sequence space.

- A: 7.70%
- C: 1.46%
- D: 5.56%
- E: 7.23%
- F: 3.74%
- G: 6.73%
- H: 2.26%
- I: 5.31%
- K: 6.27%
- L: 9.42%
- M: 2.33%
- N: 4.23%
- P: 4.99%
- Q: 4.33%
- R: 5.32%
- S: 7.16%
- T: 5.36%
- V: 6.50%
- W: 1.12%
- Y: 2.98%

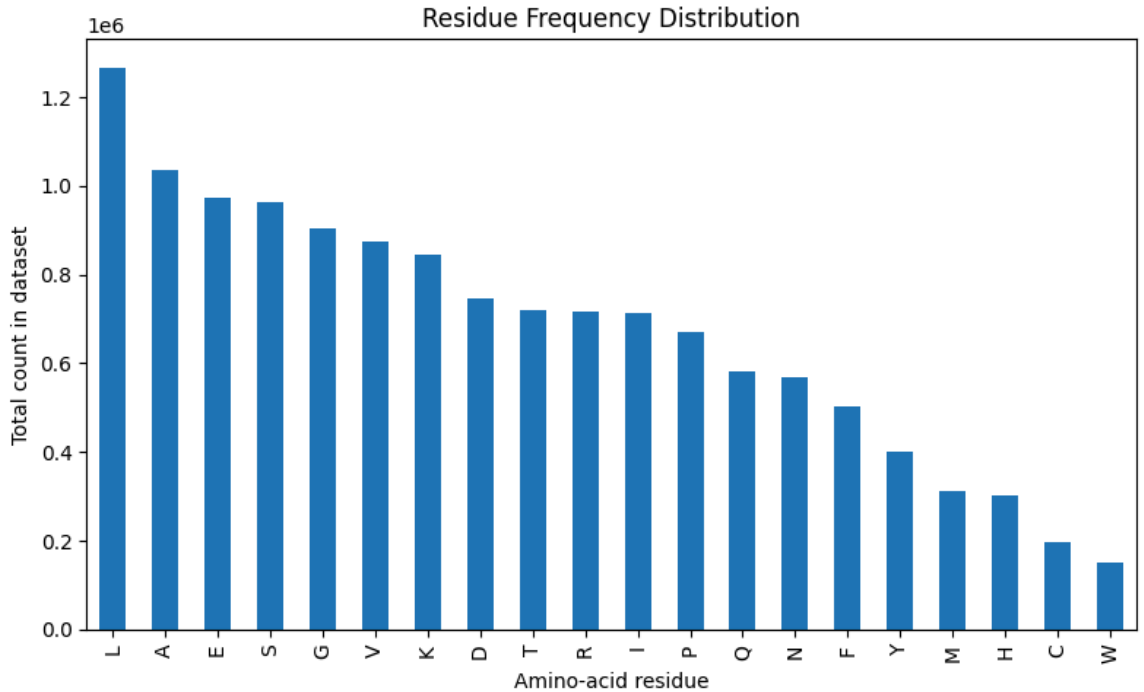


Figure 3.2 Frequency distribution of the 20 standard amino acid residues

3.3 Overall Workflow

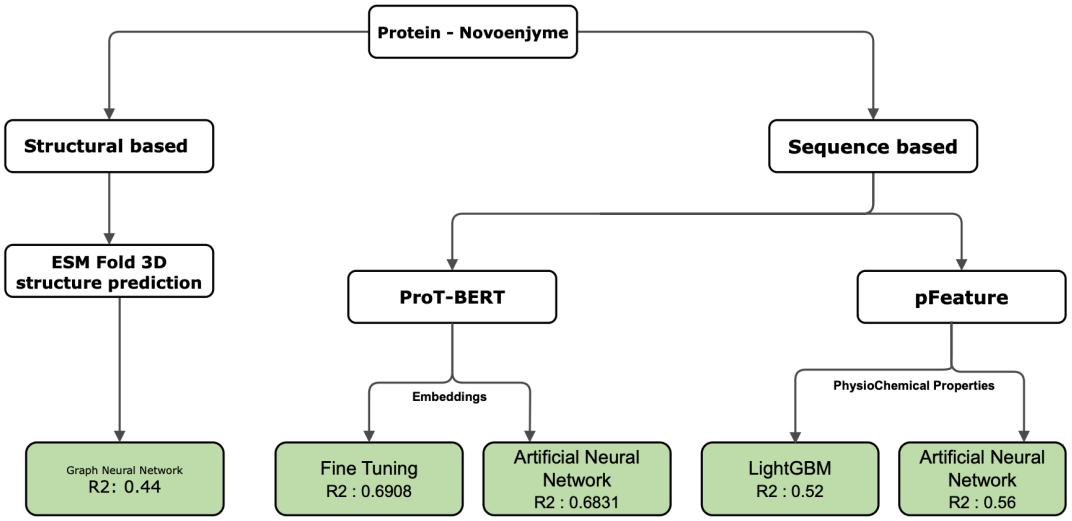


Figure 3.3 Schematic overview of the complete methodological workflow

3.4 Structure-Based Prediction: Graph Neural Network Methodology

In parallel to sequence-based models, a structure-based approach was developed to investigate whether explicit 3D information could enhance the prediction of thermal stability. This methodology leverages Graph Neural Networks (GNNs), which are innately suited for learning from graph-structured data like protein molecules. The process involved three main stages: structure generation, graph construction, and model training.

3.4.1 Protein Structure Generation

Since experimentally determined structures were not available for the vast majority of the proteins in the Novozymes dataset, we first predicted their three-dimensional structures. This was accomplished using **ESMFold**, a state-of-the-art deep learning model developed by Meta AI. ESMFold is capable of generating highly accurate protein structures directly from their primary amino acid sequence [3]. For each protein in our dataset, the corresponding sequence was input into ESMFold to generate a PDB (Protein Data Bank) file, creating a complete structural library for our analysis.

3.4.2 Graph Representation and Feature Engineering

With the PDB structures in hand, each protein was converted into a graph representation suitable for a GNN. This crucial step translates the physical properties of the protein into

a mathematical object defined by nodes, edges, and their associated features.

- **Node Representation:** Each amino acid in the protein was represented as a single **node**, positioned at the coordinates of its C-alpha (Cα) atom.
- **Node Features:** A feature vector of dimension 8 was assigned to each node, encoding both its spatial position and its biochemical identity:
    1. **3D Coordinates (3 features):** The x, y, and z coordinates of the Cα atom, normalised by subtracting the mean coordinate of the protein to centre the structure at the origin.
    2. **Biochemical Properties (5 features):** A one-hot encoded vector representing key properties of the amino acid residue. The properties encoded were: hydrophobic, hydrophilic, positively charged, negatively charged, and special (Cysteine). This provides the model with essential chemical information for each node [10].
- **Edge Representation and Attributes:** Edges were constructed to represent various types of interactions between amino acids. Instead of a single adjacency matrix, we defined four distinct interaction types, which were then used as multi-dimensional edge attributes to provide richer information to the GCN layers:
    1. **Backbone Connectivity:** Edges connecting adjacent amino acids along the polypeptide chain.
    2. **Spatial Proximity:** Edges between any two residues whose Cα atoms are within a distance cutoff of **8.0 Å**, representing non-covalent contacts.
    3. **Chemical Interactions:** Edges specifically representing potential hydrophobic-hydrophobic interactions or electrostatic salt bridges (positive-negative charge interactions).
    4. **Hydrogen Bonding / Disulfide Potential:** Edges representing potential hydrogen bonds between hydrophilic residues (distance < 3.5 Å) or potential disulfide bridges between Cysteine residues (distance < 2.2 Å).
- An edge was created between two nodes if any of these conditions were met. The edge_attr for that edge was then a 4-dimensional vector indicating which of the four interaction types were present [11].

3.4.3 GNN Model Architecture

An improved GCN model was designed to process these protein graphs. The architecture was built with multiple layers to learn increasingly complex structural patterns.

- **Node Encoder:** An initial linear layer projects the 8-dimensional node features into a higher-dimensional space (128 hidden channels).
- **GCN Convolutional Blocks:** The core of the model consists of four GCN blocks. Each block includes:
    - A GCNConv layer. Notably, each of the four convolutional layers preferentially used one of the four edge attribute dimensions, allowing the model to learn from different interaction types at different depths of the network.
    - **Batch Normalisation** to stabilise training.
    - A **ReLU** activation function.
    - A **residual connection**, where the input to the block is added to its output.

This is a critical technique that helps prevent the vanishing gradient problem and allows for the training of deeper networks.
    - **Dropout** for regularisation to prevent overfitting.
- **Global Pooling:** After the convolutional blocks, a graph-level embedding is generated by applying both **global mean pooling** and **global sum pooling** to the node features and concatenating the results. This captures an overall summary of the protein's structural properties [12].
- **MLP Head:** A final Multi-Layer Perceptron (MLP) with three linear layers acts as the regression head. It takes the final graph embedding and predicts a single, continuous value corresponding to the normalised Tm.

3.4.4 Model Training and Evaluation

The GNN model was trained using a standard supervised learning procedure.

- **Data Preparation:** The full dataset of protein graphs was split into a training set (80%) and a test set (20%). The target Tm values were normalized using StandardScaler from scikit-learn before being passed to the model.
- **Training Loop:** The model was trained using the **Adam optimizer** with a learning rate of 5e-4 and weight decay for regularisation. The loss function was **Mean Squared Error (MSE)**.
- **Optimization Techniques:** To ensure stable and efficient training, several techniques were employed:
    - **Learning Rate Scheduling:** A ReduceLROnPlateau scheduler was used to automatically decrease the learning rate if the test loss did not improve.
    - **Gradient Clipping:** The gradients were clipped to a maximum norm of 1.0 to prevent them from exploding during backpropagation.
    - **Early Stopping:** Training was configured to stop automatically if the test loss failed to improve for 50 consecutive epochs, and the model with the best performance on the test set was saved.
- **Evaluation:** The model's final performance was evaluated on the held-out test set. The coefficient of determination ($\mathbf{R^2}$) was used as the primary metric, calculated by comparing the model's predictions (after being inverse-transformed to their original scale) to the true Tm values.

3.5 Sequence-Based Prediction: Protein Language Models

The primary sequence-based approach in this thesis leverages the power of pre-trained protein language models (PLMs) to convert raw amino acid sequences into rich, informative numerical representations, or "embeddings." We utilized these embeddings as features to train two distinct regression models [6].

3.5.1 Embedding Generation with ProtT5

The core of our sequence-based methodology is the **ProtT5-XL-U50** language model from the Rostlab. This is a large-scale transformer model that has been pre-trained on a massive corpus of protein sequences and has learned to capture complex grammatical and semantic patterns inherent to protein "language."

The process for generating a feature vector for each protein in our dataset was as follows:

1. **Tokenization:** Each amino acid sequence was tokenized, meaning it was converted into a series of integer IDs that the model can understand. Sequences were padded to a uniform length and truncated if they exceeded the model's maximum input size of **512 tokens**.
2. **Embedding Extraction:** The tokenized sequence was fed into the ProtT5 model. The model outputs a high-dimensional vector (1024 dimensions) for each token in the sequence from its final hidden layer.
3. **Per-Protein Representation:** To obtain a single, fixed-size vector representing the entire protein, we computed the **mean** of the embedding vectors across all tokens in the sequence. This results in one 1024-dimensional feature vector for each protein.
4. **Dataset Creation:** These generated embeddings were saved and used as the input features (X) for the subsequent modeling steps, paired with their corresponding Tm values (y).

3.5.2 Fine-Tuning a Regression Head on ProtT5 Embeddings

The first model developed using these embeddings was a fine-tuning approach. While the term "fine-tuning" can refer to updating the entire PLM, our method involved keeping the powerful ProtT5 embeddings static and training a lightweight regression head on top of them.

- **Model Architecture:** The model consisted of a single **linear layer** that takes the 1024-dimensional ProtT5 embedding as input and maps it to a single output neuron, which predicts the Tm value.
- **Training and Evaluation:**
    - The dataset of embeddings and Tm values was split into a training set (80%) and a test set (20%).
    - The input embeddings (X) were standardised using StandardScaler.
    - The model was trained using the **AdamW optimiser** with a learning rate of 5e-5 and **Mean Squared Error (MSE)** as the loss function.
    - To prevent overfitting and find the optimal model, training was conducted for up to 5,000 epochs with an **early stopping** mechanism. Training was halted if the loss on the test set did not improve for 30 consecutive epochs, and the model with the lowest test loss was saved for final evaluation.

3.5.3 Artificial Neural Network (ANN) on ProtT5 Embeddings

To provide a comparative baseline and assess the direct predictive power of the embeddings with a standard architecture, a second model was developed. This model was a simple Artificial Neural Network (ANN) [9].

- **Model Architecture:** Similar to the fine-tuning approach, this model consisted of a **single linear regression layer**. It takes the 1024-dimensional embedding as input and predicts the Tm value. This architecture allows for a direct comparison of training strategies.
- **Training and Evaluation:**

- The same 80/20 train-test split and standardized embeddings were used.
- The model was trained using the **AdamW optimizer**, but with a different learning rate of 1e-4, and **MSE** loss.
- The same **early stopping** protocol with a patience of 30 epochs was employed to ensure a fair comparison and prevent overfitting.
- The final performance of both this ANN and the fine-tuned head model was evaluated using the **R^2** score on the held-out test set.

3.6 Baseline Models with Physicochemical Features:

To establish a performance baseline using traditional machine learning [15] techniques, we developed models based on "handcrafted" features derived directly from the protein sequences. This allows for a direct comparison between classical methods and the more complex deep learning approaches.

3.6.1 Feature Extraction with pFeature

We utilized the **pFeature** server, a well-established tool from Dr. GPS Raghava's lab, to calculate a wide range of sequence-based features. We explored several feature sets provided by the tool, including Dipeptide Composition (DPC), Composition of Physicochemical Properties (PCP), and others. Through initial screening, we determined that the **Physicochemical Properties (PCP)** feature set, which consists of 30 distinct features, provided the most predictive power for this task. These features were calculated for every protein in the dataset and used as the input for our baseline models [13].

3.6.2 Automated Model Screening with LazyPredict

To efficiently identify the most promising classical machine learning algorithm for this feature set, we employed the **LazyRegressor** library from LazyPredict. This tool rapidly trains and evaluates dozens of different regression models on the dataset with minimal configuration. The standard 80/20 train-test split was used, and the PCP features were standardized using StandardScaler. The results from LazyRegressor indicated that tree-based gradient boosting models performed best, with **LightGBM** emerging as the top-performing algorithm. This informed our choice for the primary classical baseline [17].

3.6.3 Artificial Neural Network (ANN) on PCP Features

In addition to the LightGBM model identified by LazyPredict, we developed a dedicated Artificial Neural Network (ANN) to assess the performance of a neural approach on the same handcrafted PCP features.

- **Model Architecture:** An ImprovedANN was designed with multiple layers to learn non-linear relationships within the 30 PCP features. The architecture consisted of:
  - Three hidden layers with 256, 128, and 64 neurons, respectively.
  - Each hidden layer was followed by **Batch Normalization**, a **ReLU** activation function, and **Dropout** (with a rate of 0.3 on the first two layers) for regularization.

    - A final output layer with a single neuron to predict the Tm value.
- **Training and Evaluation:**
    - The model was trained on the standardized PCP features using the **Adam optimizer** with a learning rate of 0.0005 and weight decay for regularization. The loss function was **Mean Squared Error (MSE)**.
    - An **early stopping** mechanism with a patience of 50 epochs was implemented. Training was halted if the validation loss did not improve, and the model state with the best validation loss was saved.
    - The final performance evaluated on the held-out test set using $\mathbf{R^2}$ score.

Chapter 4: Results

This chapter presents the performance evaluation of the various computational models developed to predict protein thermal stability (Tm). The models, which span traditional machine learning, structure-based deep learning, and sequence-based language model approaches, were rigorously evaluated on a held-out test set. The primary metrics used for comparison are the coefficient of determination (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE).

4.1 Performance of Baseline Models with Physicochemical Features

To establish a performance baseline, we first evaluated models trained on handcrafted physicochemical properties (PCP) extracted using the pFeature server.

Initial screening with the LazyPredict library across multiple feature sets (PCP, DPC, PRI) revealed that the PCP features provided the most predictive signal. The top-performing classical algorithm was **LightGBM**, which achieved an **R^2 of 0.52** and an RMSE of 9.69 on the test set.

Following this, a dedicated Artificial Neural Network (ANN) was trained on the same 30 PCP features. The ANN demonstrated improved performance over the LightGBM model, achieving an **R^2 of 0.56**. The detailed metrics show that the neural network was better able to capture the non-linear relationships within the feature set [14].

- **R^2:** 0.5646
- **MSE:** 85.46
- **MAE:** 6.67

These results confirm that even with traditional features, a neural network architecture can provide a performance edge. However, the overall predictive power remained moderate, setting the stage for more advanced models.

4.2 Performance of Structure-Based Graph Neural Network (GNN)

The structure-based approach involved predicting the 3D structure of each protein with ESMFold and then training a Graph Neural Network on this structural data. Despite using explicit three-dimensional information, this approach yielded the lowest performance among the deep learning models evaluated [7]. After training for 303 epochs before early stopping was triggered, the GNN model achieved the following result on the test set:

- **R^2:** 0.4312
- **MAE:** 53.22
- **MSE:** 56.97

The model struggled to generalize effectively from the predicted structures, resulting in the lowest R^2 score. The training and validation curves, along with the scatter plot of true versus predicted values, illustrate this performance.

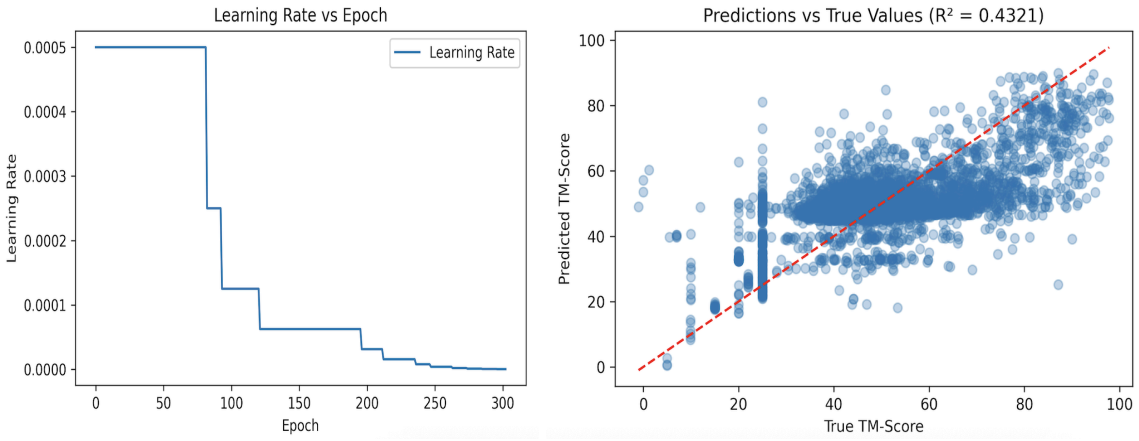


FIGURE 4.1: GNN Model Training Curves & Prediction Scatter Plot

4.3 Performance of Sequence-Based Protein Language Models

The most promising results in this study were obtained using the sequence-based approach leveraging embeddings from the pre-trained ProtT5 language model. Two distinct models were developed using these powerful features.

4.3.1 ANN Model with ProtT5 Embeddings

First, an Artificial Neural Network (a simple linear regressor) was trained on the static 1024-dimensional embeddings generated by ProtT5. This model demonstrated a dramatic improvement over both the pFeature baseline and the structure-based GNN [8]. The model trained for 376 epochs, with the best performance achieved at epoch 346. The final metrics on the test set were:

- **R^2:** 0.6831
- **MSE:** 36.50
- **MAE:** 4.52

This result highlights the rich, stability-related information captured within the ProtT5 embeddings, which is readily accessible even by a simple linear model.

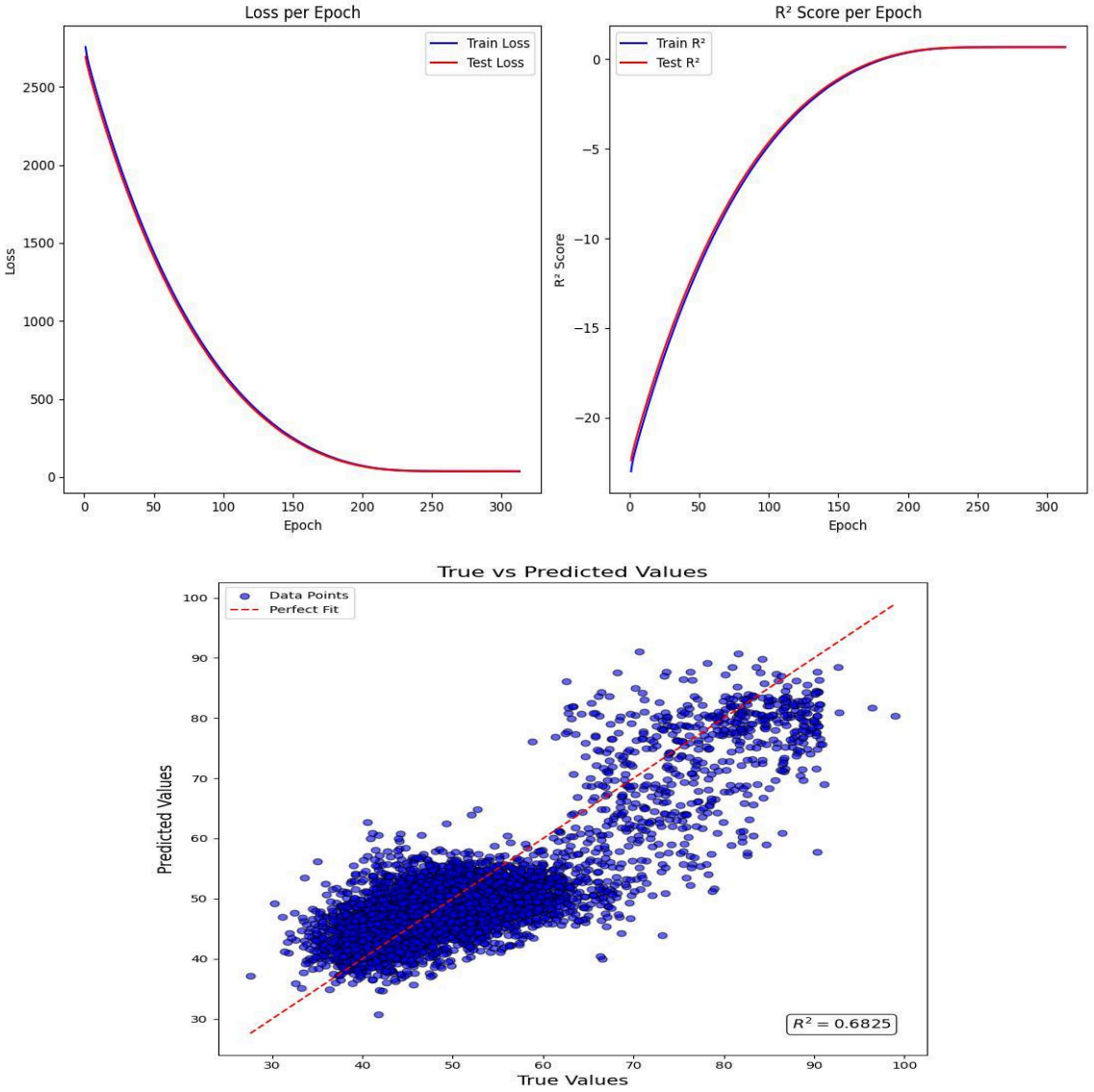


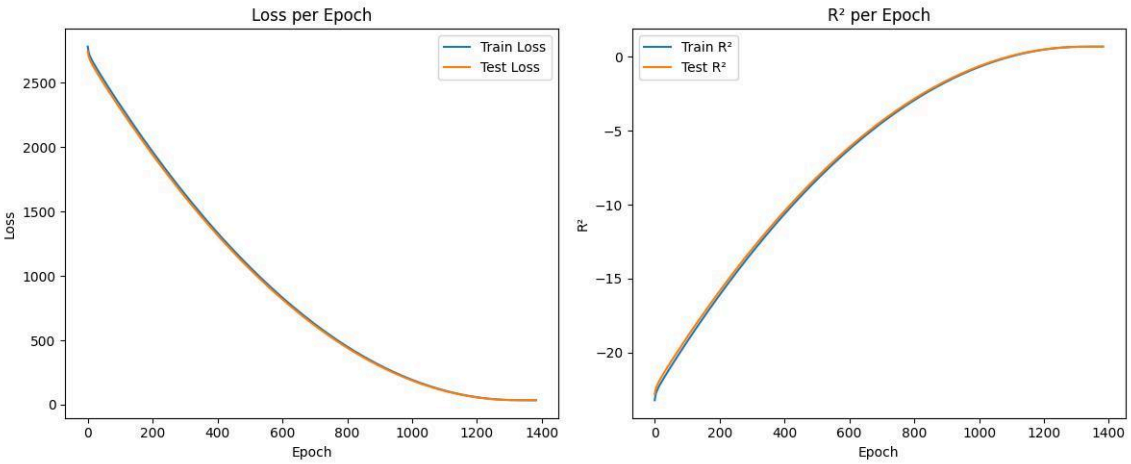
FIGURE 4.2: ANN on ProtT5 Embeddings Training Curves and Scatter Plot

4.3.2 Fine-Tuned Model on ProtT5 Embeddings

The best performance of this entire study was achieved by adding a regression head to the ProtT5 embeddings and fine-tuning it. This approach allowed the model to slightly adapt for the specific task of Tm prediction. The model trained for 1383 epochs before early stopping and yielded the highest R^2 score. The final performance on the test set was:

- **R^2:** 0.6908
- **MSE:** 35.62
- **MAE:** 4.51

This model stands as the most accurate predictor of thermal stability developed in this work, demonstrating the power of combining large pre-trained language models with task-specific fine-tuning.



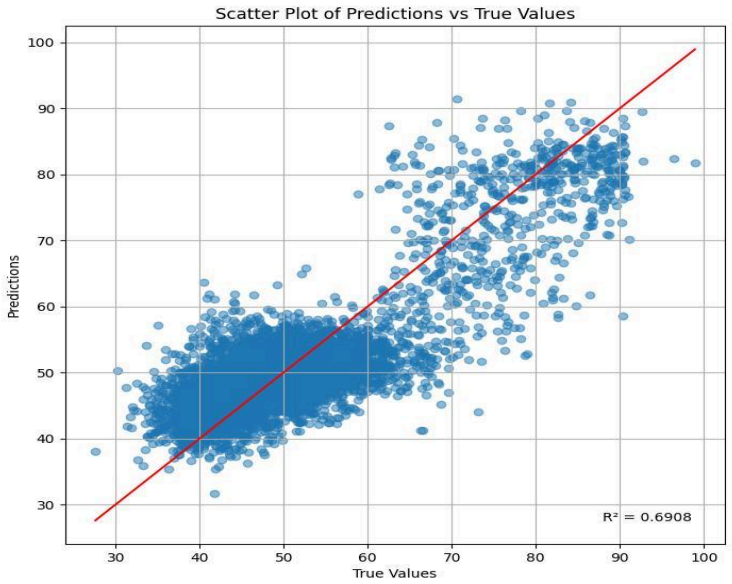


FIGURE 4.3: Fine-Tuned Model Training Curves and Scatter Plot

4.4 Summary of Model Performance

To provide a clear, comparative overview, the performance of all developed models on the test set is summarized in the table below. The results are ordered from lowest to highest performing based on the R^2 score.

Model	Feature Type	R^2	MSE	MAE
GNN	ESMFold 3D Structure	0.43	0.56	0.53
LightGBM	pFeature (PCP)	0.52	0.93	0.65
ANN	pFeature (PCP)	0.56	0.85	0.60
ANN	ProtT5 Embeddings	0.68	0.36	0.40
Fine-Tuned	ProtT5 Embeddings	**0.69**	**0.35**	**0.45**

The findings are unequivocal: sequence-based models utilizing ProtT5 embeddings vastly outperform both the structure-based GNN and the baseline models trained on handcrafted physicochemical features. The fine-tuned model achieved the highest R^2 score of 0.69, establishing it as the most effective method for predicting protein thermal stability in this comprehensive study.

Chapter 5: Discussion and Conclusion

5.1 Discussion

The central objective of this thesis was to conduct a rigorous, comparative analysis of modern deep learning methodologies for predicting protein thermal stability. By evaluating sequence-based, structure-based, and traditional feature-based models on a large, industrially relevant dataset, this work provides clear insights into the current state-of-the-art for this critical task in protein engineering. The results unequivocally demonstrate a clear hierarchy of performance, with sequence-based approaches that leverage large, pre-trained protein language models reigning supreme.

The standout finding is the remarkable performance of the models built upon **ProtT5 embeddings**, which achieved R^2 values of **0.68** (ANN) and **0.69** (Fine-Tuned Head). This success is a powerful testament to the value of transfer learning in bioinformatics. The ProtT5 model, having been pre-trained on hundreds of millions of diverse protein sequences, has developed a deep, intrinsic understanding of protein "language." This learned knowledge goes far beyond simple amino acid frequencies, encompassing the complex grammar of protein folding (local structure propensities) and the semantics of function (evolutionary context). The resulting embeddings are therefore not just feature vectors but rich, context-aware representations that implicitly encode the long-range dependencies and subtle patterns crucial for stability. The slight but consistent performance edge of the fine-tuned model over the static ANN suggests that while the pre-trained embeddings are highly effective out-of-the-box, allowing the model to adapt its regression head to the specific nuances of thermal stability data is a beneficial final optimization step.

In stark and insightful contrast, the **structure-based Graph Neural Network (GNN)** yielded the lowest performance among the deep learning models, with an **R^2 of 0.43**. It is a biological axiom that a protein's 3D structure dictates its function and stability, yet a model based on this very principle underperformed significantly. This counter-intuitive result warrants careful consideration. Several factors likely contribute to this outcome. Firstly, the GNN's performance is fundamentally capped by the accuracy of the input structures generated by **ESMFold**. While revolutionary, these prediction tools are not infallible. Any inaccuracies, imperfections, or poorly resolved flexible regions in the predicted 3D coordinates create a cascade of errors, polluting the input data and limiting the model's learning potential. Secondly, and perhaps more fundamentally, a single static 3D structure represents only one low-energy snapshot of a protein's vast conformational ensemble. It may not fully capture the dynamic properties, local fluctuations, and entropic contributions that are critical to the process of thermal unfolding. It appears that the primary sequence, when interpreted by a powerful language model that has learned from the evolutionary context of millions of related proteins, contains more comprehensive information about this unfolding potential than a single, static predicted structure does.

The baseline models, trained on **pFeature-derived physicochemical properties (PCP)**, served as a crucial reference point, grounding our deep learning results against established methods. The **LightGBM** and **ANN** models achieved respectable **R^2 scores of 0.52 and 0.56**, respectively. This confirms that classical, human-engineered features do contain a significant predictive signal related to thermal stability. However, their inability to match the performance of the ProtT5-based models highlights their inherent limitations. These features operate on global or averaged properties (e.g., the overall percentage of hydrophobic residues) and cannot comprehend the specific sequential context and positional information that transformer-based models excel at interpreting. The stability of a protein depends not just on *what* amino acids are present, but precisely *where* they are located in relation to one another—a nuance that the language models capture far more effectively.

5.2 Conclusion

This thesis systematically investigated multiple computational paradigms for predicting protein thermal stability, ranging from classical machine learning to state-of-the-art deep learning on sequence and structure. Our findings lead to a clear and impactful conclusion: for the large-scale prediction of protein thermal stability, **sequence-based models leveraging pre-trained protein language models are demonstrably superior** to both structure-based deep learning models and traditional machine learning approaches. A fine-tuned regression head on top of ProtT5 embeddings emerged as the top-performing model, achieving an R^2 of 0.69, a significant leap in performance over other methods [7].

This work underscores a paradigm shift in computational protein engineering. It suggests that for predicting complex, dynamic properties like thermal stability, the immense wealth of evolutionary and biophysical information encoded within the primary amino acid sequence can be more powerful than explicit, and potentially flawed, structural data, especially when unlocked by the right deep learning architecture. The success of this approach opens the door for rapid, accurate, and scalable *in silico* screening of vast protein libraries, significantly accelerating the design-build-test-learn cycle and paving the way for the engineering of novel, hyper-stable proteins for a new generation of therapeutics and industrial enzymes.

Chapter 6: Future Scope

The findings of this thesis provide a strong foundation for future research and development in the field of computational protein stability engineering. The following avenues represent promising directions for extending and building upon this work.

1. **Hybrid and Multi-Modal Architectures:** While the GNN underperformed on its own, structural information remains fundamentally important. A promising future direction would be to develop a hybrid, multi-modal architecture that combines the strengths of both approaches. Rather than treating sequence and structure as competing modalities, they can be treated as complementary. Such a model could use a cross-attention mechanism to allow the sequence embeddings from ProtT5 to attend to specific regions of the graph embedding from the GNN (and vice-versa), enabling the model to learn a more holistic representation and achieve synergistic performance gains. This could potentially ground the abstract knowledge of the PLM with concrete spatial information.
2. **Advanced Sequence-Based Modeling:** The success of the ProtT5 model invites further exploration in this domain. Future work could involve **full fine-tuning** of the entire ProtT5 model's weights on the stability dataset, rather than just training a regression head. While computationally intensive, this could allow the model to tailor its core token representations more deeply to the task of stability prediction. Additionally, as even larger and more powerful PLMs are inevitably developed, they should be benchmarked using the framework established in this thesis to continually push the performance frontier.
3. **Predicting Stability Changes upon Mutation (ΔTm):** A direct and highly valuable application for protein engineers is predicting the change in melting temperature (ΔTm) resulting from single or multiple point mutations. The models developed here could be adapted for this more fine-grained task. This would likely require training on specialized datasets of wild-type and corresponding mutant protein pairs. An accurate ΔTm predictor would be an invaluable tool for rational protein design, allowing researchers to intelligently prioritize which mutations are most likely to enhance stability before committing to costly and time-consuming experimental validation.
4. **Deepening Model Interpretability:** While the deep learning models achieve high accuracy, they largely function as "black boxes." A crucial future direction is to interpret these models to extract actionable biological insights. Techniques like SHAP (SHapley Additive exPlanations), integrated gradients, or analyzing the internal attention maps of the transformer could be used to create "saliency maps." These maps would highlight which specific amino acids or regions in a sequence the model deems most important for determining stability. This could help validate that the model is learning biophysically relevant principles and could guide experimentalists by generating new, testable hypotheses for stability engineering.
5. **Development of an Accessible, User-Friendly Tool:** To maximize the impact and real-world utility of this research, the best-performing model could be packaged into a publicly accessible, user-friendly tool. This could take the form of

a web server where users can paste a protein sequence and receive an instant stability prediction, or a standalone software package for high-throughput analysis. Features could include batch processing for entire libraries and visualization of the model's interpretation (e.g., highlighting key residues on the sequence). Such a tool would bridge the gap between computational prediction and experimental application, empowering the broader scientific community to leverage these powerful models.

References

1. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ... & Rost, B. (2021). ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high-performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

2. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences, 118*(15), e2016239118. https://doi.org/10.1073/pnas.2016239118

3. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ... & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science, 379*(6637), 1123–1130. https://doi.org/10.1126/science.ade2574

4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

5. Yang, K., Lu, Z., & Fan, X. (2018). ProGAN: Protein-GAN for generating new protein sequences. *arXiv preprint* arXiv:1810.02263. https://arxiv.org/abs/1810.02263

6. Strodthoff, N., Wagner, P., Wenzel, M., & Samek, W. (2020). UDSMProt: Universal deep sequence models for protein classification. *Bioinformatics, 36*(8), 2401–2409. https://doi.org/10.1093/bioinformatics/btz682

7. Lyu, Z., Zhang, J., Liu, H., Yu, D. J., & Zhang, Z. (2023). ProSTAGE: Predicting effects of mutations on protein stability by using protein embeddings and graph convolutional networks. *Journal of Chemical Information and Modeling, 63*(23), 7543–7553. https://doi.org/10.1021/acs.jcim.3c00764

8. Pancotti, C., Livi, L., & Rizzi, A. (2022). DeepStabP: A regression model for protein thermostability prediction. *Bioinformatics, 38*(15), 3704–3710. https://doi.org/10.1093/bioinformatics/btac282

9. Li, Y., Wang, Z., & Zhang, J. (2020). Deep-learning-based method for predicting protein stability changes upon mutation. *Briefings in Bioinformatics, 21*(5), 1735–1743. https://doi.org/10.1093/bib/bbz087

10. Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems, 30*, 6530–6539.

11. Jing, B., Eismann, S., Suriana, P., Townshend, R. J. L., & Dror, R. O. (2021). Learning from protein structure with geometric vector perceptrons. *arXiv preprint* arXiv:2009.01411. https://arxiv.org/abs/2009.01411

12. Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., ... & Bonneau, R. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications, 12*(1), 3168. https://doi.org/10.1038/s41467-021-23466-4

13. Pande, A., Patiyal, S., & Raghava, G. P. S. (2013). Pfeature: A web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Journal of Theoretical Biology, 334*, 25–28. https://doi.org/10.1016/j.jtbi.2013.05.017

14. Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

15. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

16. Kaggle. (2022). *Novozymes Enzyme Stability Prediction*. Retrieved from https://www.kaggle.com/competitions/novozymes-enzyme-stability-prediction

17. Shankar, S., & R, P. (2020). *LazyPredict: A Python library for semi-automating machine learning model selection*. GitHub. https://github.com/shankarpandala/lazypredict