Developing machine learning and deep learning models for predicting the thermal stability of proteins

By

Surabhi Singh

MT22207

Under the Supervision of Dr. N. Arul Murugan

Indraprastha Institute of Information Technology, Delhi

May, 2024

Developing machine learning and deep learning models for predicting the thermal stability of proteins

By

Surabhi Singh

MT22207

Submitted

in partial fulfilment of the requirements for the degree of Master of

Technology

To

Indraprastha Institute of Information Technology Delhi Month, 2024

CERTIFICATE

This is to certify that the thesis titled " Developing machine learning and deep learning models for predicting the thermal stability proteins" being submitted by Surabhi Singh (MT22207) to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology, is an original research work carried out by her 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.

May, 2024, Dr. N. Arul Murugan

Department of Computational Biology

Indraprastha Institute of Information Technology,

Delhi New Delhi 110 020

Acknowledgments

I would like to express my heartfelt gratitude to my supervisor, Dr. N. Arul Murugan, and his team for their consistent support and motivation throughout this project. Their invaluable guidance and encouragement have been instrumental in the completion of this work. I would also like to extend my sincere thanks to Prateek Paul and Rudrakshi Chauhan for their constant support and guidance. Their insights and assistance have been crucial in navigating the challenges encountered during the course of this project.

TABLE OF CONTENTS

Abstract

The accurate prediction of protein stability temperatures is essential for numerous applications in bioinformatics and biotechnology. In this study, we utilized a multifaceted computational approach to develop predictive models for protein stability temperatures and determine which modeling technique yields the best results. We began by utilizing the Pfeature package to compute different set of 16 features for a dataset comprising 31,470 protein sequences. These features encompassed various aspects, including amino acid composition, physicochemical properties, and structural characteristics. Subsequently, the dataset was standardized using StandardScaler to prepare it for analysis. Next, we employed an array of modeling techniques, including Artificial Neural Networks (ANN), Linear Regression, Decision Trees, and Random Forests, to establish predictive models. Each model was trained on the concatenated dataset of protein features and evaluated using standard regression metrics such as root mean square error (RMSE), mean absolute error (MAE), and R^2 score. Furthermore, we utilized MODELLER for homology modeling to generate three-dimensional structures for a subset of 15,000 proteins, selected based on sequence similarity. The Graphein package facilitated the analysis of protein structures by computing various types of bonds within the proteins. Additionally, using Amber Tools, we computed various energy components for each protein structure, including bond energy, angle energy, and solvation energy. These energy values were integrated into a dataset alongside the corresponding stability temperatures. Finally, we assess the accuracy of the different modeling techniques by evaluating their predictive accuracy using the aforementioned regression metrics. By systematically assessing the performance of each model, we endeavored to identify the most effective approach for predicting protein stability temperatures. This comprehensive computational study offers valuable insights into the prediction of protein stability temperatures, offering a sturdy foundation for future research in this domain.

Chapter 1

Introduction

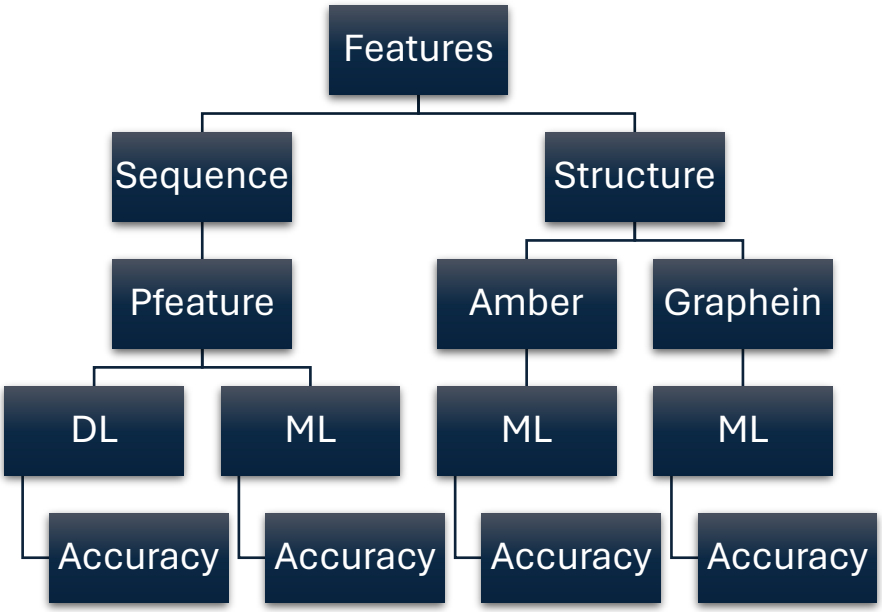
Thermophilic proteins, adapted to high temperatures, exhibit remarkable stability owing to a combination of structural and compositional adaptations. These proteins typically display greater hydrophobicity, facilitating tighter packing of amino acids and reducing water interactions. They often feature shortened or fewer loops, minimizing flexibility, along with fewer cavities, limiting water penetration into the core. Furthermore, upon oligomerization, thermophilic proteins bury more surface area, enhancing stability. Amino acid substitutions,

including an increased occurrence of proline residues and a decrease in thermolabile residues, contribute to their stability. Additionally, thermophilic proteins tend to have higher helical content, more polar surface area, and an increased propensity for forming hydrogen and salt bridges, all of which enhance structural integrity. Through these adaptations, thermophilic proteins maintain functional stability even in extreme thermal environments, providing valuable insights into protein stability and adaptation. Thermally, stable proteins exhibit the ability to maintain their structural integrity and functionality even under elevated temperatures, making them valuable in several industries such as manufacturing, pharmaceuticals, and protein engineering. The assessment of a protein's thermal stability commonly relies on determining its melting temperature (Tm)[1]. However, with the rapid expansion of protein sequences, current databases like ProThermDB containing approximately 120,000 records, with fewer than 10,000 providing Tm information, face challenges in keeping up. In contrast, databases like UniProtKB/Swiss-Prot and UniProtKB/TrEMBL (Release 2023_01) (reference) boast over 200 million protein sequence entries. Determining a protein's Tm in the laboratory involves intricate procedures including protein expression, purification, and specialized instrumentation utilization. Differential Scanning Calorimetry (DSC), Circular Dichroism (CD) spectroscopy, and Differential Scanning Fluorimetry (DSF) are usual procedure used for this purpose. DSC measures heat capacity by controlling temperature, CD spectroscopy analyses structural changes via circularly polarized light, and DSF exploits hydrophobicity changes upon heating. In cases where screening numerous protein scaffolds is necessary, computational prediction may be more feasible than experimental methods, especially considering expression and purification challenges. Due to the impracticality of experimentally measuring Tm values for the vast number of proteins lacking this data, computational methods are increasingly vital for predicting protein melting temperatures[3], closing the gap between protein sequence information and thermal stability data. Stability stands as a crucial property across all proteins and biological macromolecules, ensuring their functionality in varied environmental conditions. Protein engineering often targets enhancing stability, particularly thermal stability, to bolster functionality. Disease-related variations commonly lead to reduced stability, necessitating thorough measurements. Protein stability encompasses the capacity to uphold structural and functional integrity amidst changing environments [2]. Thermal stability, in particular, is pivotal due to its implications in biological processes. Proteins with lower thermal stability are prone to aggregation, potentially leading to loss of function or even toxic aggregate formation. Factors influencing thermal stability include amino acid composition, protein-protein interactions, post-translational modifications, and the presence of ligands. Understanding thermal stability is vital across biotechnology and food science, where proteins endure temperature fluctuations during cultivation, processing, and storage. Thermal stability is commonly quantified through denaturation or melting temperature (Tm), indicative of the temperature at which 50% of the protein loses its native structure or via the area under the melting curve. Traditionally, measuring protein stability demanded significant effort and yielded limited data, prompting the development of computational solutions. Predictors for

protein stability broadly fall into two categories: those assessing overall protein stability and those evaluating the impact of sequence alterations. Various prediction methods utilize diverse descriptors, including sequence physicochemical features[5], organism living temperature, salt bridges, and statistical potentials. Recent advancements include the collection of cellular stability data through limited proteolysis and mass spectrometry across organisms like Escherichia coli, Homo sapiens, Saccharomyces cerevisiae, and Thermus thermophilus. Additionally, the release of the Meltome Atlas provides thermal stability data for approximately 48,000 proteins across 13 organisms. Melting profiles are obtained by heating cells or lysates at different temperatures, followed by centrifugation to remove precipitates. The soluble fraction undergoes trypsin digestion, and peptides are analyzed using liquid chromatography-tandem mass spectrometry (LC-MS/MS) to derive melting temperatures. These advancements not only expand our comprehension of protein stability but also offer valuable insights for various applications, from biotechnology to food science, where proteins encounter diverse temperature conditions[4]. In Prostab2, data were sourced from two primary databases: ProTstab and the Meltome Atlas. After excluding entries containing ambiguous amino acids, the dataset comprised 3,500 records from four species in ProTstab and 31,413 records from the Meltome Atlas across thirteen species. This culminated in a final dataset comprising 34,913 records, meticulously curated for clarity and accuracy.In this project, data were sourced from Prostab2, and Pfeature was employed to generate features from protein sequences, including amino acid composition, dipeptide frequencies, and other relevant characteristics. These features served as inputs for developing both machine learning and deep learning models to predict protein thermal stability. Following feature extraction, Modeller was used to perform homology modeling to generate three-dimensional (3D) structures of proteins in PDB format. These PDB structures were then subjected to graph-based representation using the Graphein tool, enabling the creation of graphical representations of protein structures. These representations were utilized as features for machine learning models to predict protein stability. Additionally, Amber was utilized to compute the minimization energies of the generated PDB structures, providing further insights into the stability of the proteins. Machine learning models were then applied to predict protein stability based on these computed energies.The project adopted a comprehensive approach, integrating various methodologies from data collection to feature extraction, model development, and evaluation. By combining machine learning, deep learning, homology modeling, graph-based representation, and energy minimization techniques, the project aimed to improve the accuracy of protein stability prediction. Evaluation of each technique's performance was conducted to identify the most effective approach for predicting protein stability. To assess the accuracy of each model, several evaluation metrics were calculated, including the R2 score, mean square error, and mean absolute error. These metrics offer insights into the predictive performance of the models, allowing for a comprehensive assessment of their effectiveness in predicting protein stability.

Chapter 2

Work Flow



METHOD/data preprocessing

Dataset

The melting temperature data for 31470 proteins was gathered from ProTstab2.in prostab2 data collect from in prostab2 Data was gathered from the amalgamation of the ProTstab and Meltome Atlas datasets, focusing on the stability of proteins within the proteomes of human, mouse, and zebrafish. Within this data set, randomly divided the records into training and blind test sets

Method 1: pfeatues based features

The Pfeature is a package developed for computing broad spectrum of protein from sequence along with the structure of the proteins or peptides. Pfeature comprises five influential modules for calculating features, such as composition-based, binary-profile based, evolutionary information based, structure-based, and pattern-based features.

Utilize the Pfeature web server to compute 16 features of proteins .

Features generate –

Amino-acid Composition(acc),

Dipeptide Composition (dpc),

Atom Composition (atc),

Bond Composition (btc),

Physico-Chemical Properties n (pcp),

Binary Profile of amino acid (AAB)

Amino-acid index Composition (AAI),

Conjoint Triad Calculation (CTC),

Distance Distribution of Residues (DDR),

Pseudo Amino-acid Composition (PAAC),

Quasi-sequence order (QSO),

Repetitive Residue Information (RRI),

Shannon Entropy (SEP),

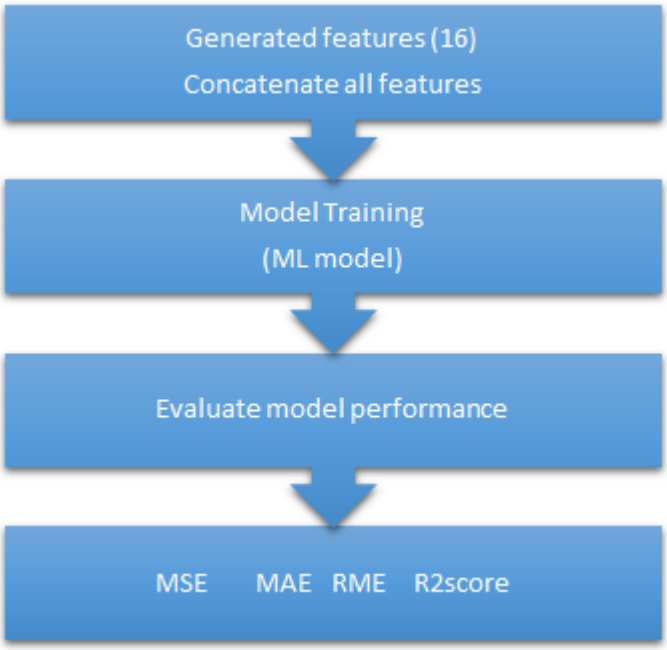
Shannon Entropy of Residue Level (SER),

Sequence Order Coupling (SOC),

Shannon Entropy of Physicochemical Property (SPC)

Start by concatenating all the features to construct a dataset with dimensions 31470 rows × 1511 columns. Next, standardize the dataset using Standard Scaler to normalize the features. . This dataset serves as the foundation for implementing machine learning models to predict stability temperatures. Various regression models, such as Linear Regression, Decision Trees, or Random Forests, can be utilized for training. Once trained, the model's performance can be evaluated using metrics like Root mean square error, Mean Squared Error, Mean Absolute Error and R2 score.

MSE	MAE	RMSE	R2score
$\mathrm{RMSE} = \sqrt{\frac{\sum_{t=1}^{N} (y_t - x_t)^2}{N}}.$	$\mathrm{MAE} = \frac{\sum_{t=1}^{N} \lvert y_t - x_t \rvert}{N}$	$\mathrm{MSE} = \frac{1}{N} \sum_{t=1}^{N} (y_t - x_t)^2$	$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$



Method 2: pfeatues based features(ANN)

The Pfeature is a package developed for computing broad spectrum of protein from sequence as along with the structure of the proteins or peptides. Pfeature comprises five influential modules for calculating features, such as composition-based, binary-profile based, evolutionary information based, structure-based, and pattern-based features.

Utilize the Pfeature web server to compute 16 features of proteins.

Features generate –

Amino-acid Composition(acc),

Dipeptide Composition (dpc),

Atom Composition (atc),

Bond Composition (btc),

Physico-Chemical Properties n (pcp),

Binary Profile of amino acid (AAB)

Amino-acid index Composition (AAI),

Conjoint Triad Calculation (CTC),

Distance Distribution of Residues (DDR),

Pseudo Amino-acid Composition (PAAC),

Quasi-sequence order (QSO),

Repetitive Residue Information (RRI),

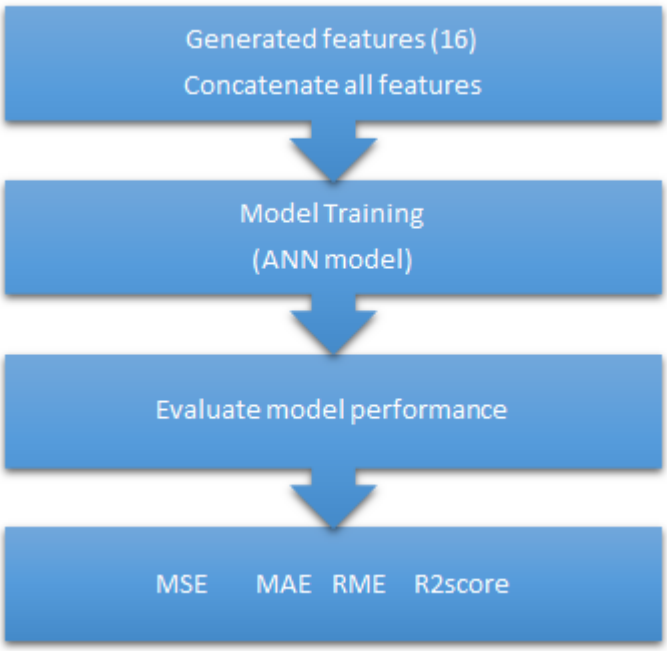
Shannon Entropy (SEP),

Shannon Entropy of Residue Level (SER),

Sequence Order Coupling (SOC),

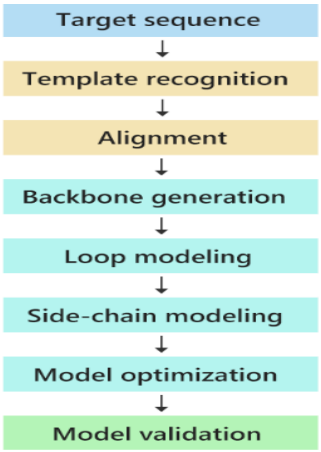
Shannon Entropy of Physicochemical Property (SPC)

Start by concatenating all the features to construct a dataset with dimensions 31470 rows × 1511 columns. Next, standardize the dataset using Standard Scaler to normalize the features. Then, proceed with the implementation of an Artificial Neural Network (ANN) for analysis. Upon training the ANN model, generate predictions (prediction value) and compare them with the actual values (y). Visualize the comparison by creating a plot showing prediction value against. Finally, evaluate the model's valuating performance using a range of metrics including root mean square error (RMSE), mean absolute error (MAE), R^2 score, and mean square error (MSE).



Method 3- Graph based Features

Homology modelling is One such computational methods for structure prediction method employed to infer the three-dimensional structure of a protein based on its amino acid sequence.



MODELLER is a software tool utilized for homology or comparative modelling of protein structures. In comparative modelling, the user supplies a sequence alignment of a target protein sequence with sequences of related proteins with established structures. Based on this alignment, MODELLER automatically generates a 3D model of the target protein.

Out of 31,470 protein sequences available, only 15,000 PDB files were generated using MODELLER. This selection was based on ensuring a minimum of 36% identity alignment between the target protein sequence and sequences of related proteins. excluding those with a radius of gyration exceeding 100 angstroms to eliminate loops. 12,860 remaining PDBs

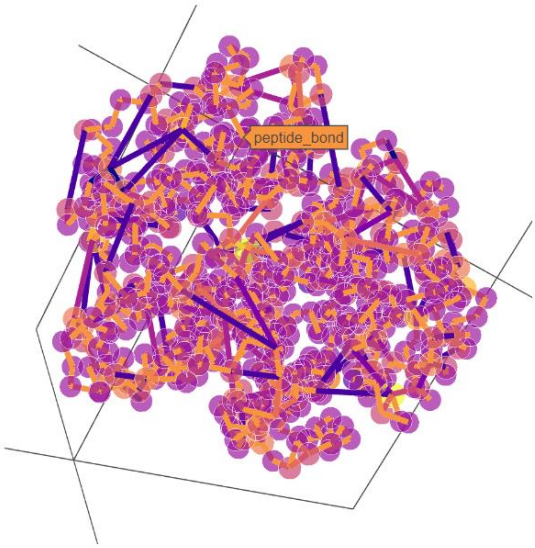
The Graphein package offers features to generate various graph-based representations of proteins. It's a Python library tailored for constructing graph and surface-mesh renditions of protein structures, primarily for computational analysis. It seamlessly connects with prevalent geometric deep learning libraries. Graphein aims to streamline network-centric and graph-theoretical assessments of protein structures, particularly in large-scale analyses.

Using Graphein with PDB structures as input, I computed various types of bonds within proteins. These include hydrogen bonds, peptide bonds, aromatic bonds, ionic bonds, single covalent bonds, double covalent bonds, double-ring covalent bonds, and ring-single covalent bonds.

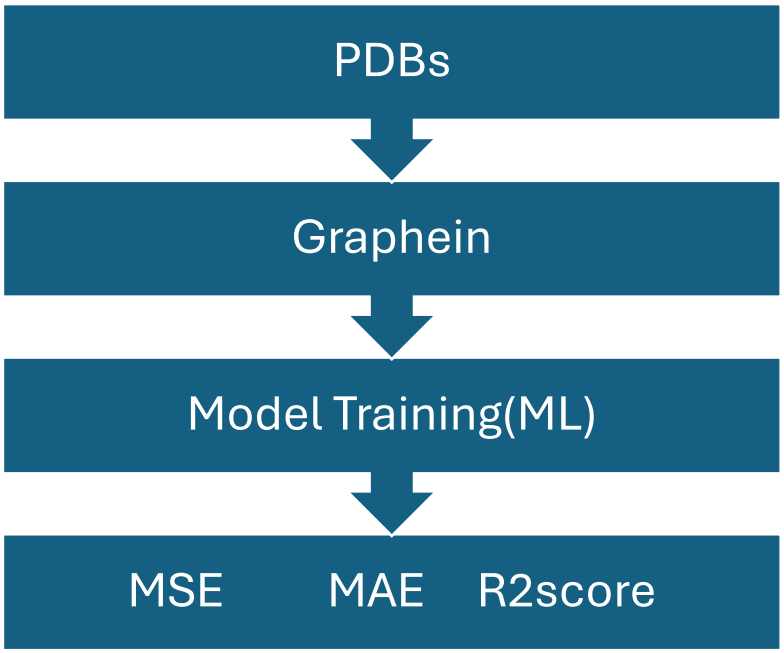
To predict the temperature at which a protein structure can remain stable, a dataset has been created with each column representing a specific type of bond and each row corresponding to a PDB name, along with the corresponding occurrences of each bond type and the melting temperature of the protein structure. This dataset serves as the foundation for implementing machine learning models to predict stability temperatures. Various regression models, like Linear Regression, Decision Trees, or Random Forests, can be utilized for training. Once

trained, the model's performance can be evaluated using metrics like Mean Squared Error, Mean Absolute Error and R2 score.

Graph-based representations of proteins



- lighter color of the node indicates a greater number of connections.
- The color of the edges represents the type of interaction



Method 4- minimization of energy based features

Amber Tools comprises a suite of open-source software programs designed for setting up, executing, and interpreting molecular simulations. Its functionalities span energy minimization, molecular dynamics simulations, free energy calculations, and beyond. Comprising various independently developed packages, Amber Tools offers flexibility for combined or standalone usage. Additionally, it supports molecular dynamics simulations using generalized Born solvent models or explicit water.

Utilizing Amber to compute the minimization energy of PBDS (presumably referring to potential energy surface points) in implicit solvent, the calculated energies typically encompass:

Bond Energy

Angle Energy

Dihedral Energy

van der Waals Energy (VDWAALS)

Electrostatic Energy (EEL)

Solvation Energy (EGB)

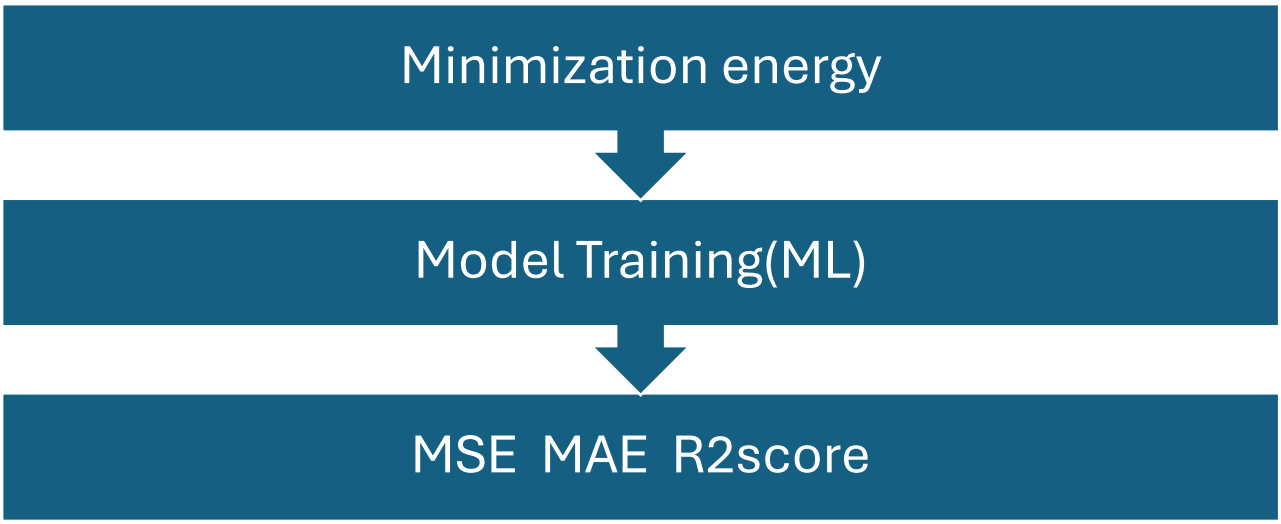
van der Waals Energy with Restraints (VDW, RESTRAINT)

Electrostatic Energy with Restraints (EEL, RESTRAINT)

Surface Energy (ESURF)

These energies collectively provide insights into the various interactions governing the system under study, including bond stretching, angle bending, dihedral rotation, nonbonded interactions, solvation effects, and any imposed restraints.

Utilizing Amber, we'll compute the Bond Energy, Angle Energy, Dihedral Energy, van der Waals Energy (VDWAALS), Electrostatic Energy (EEL), Solvation Energy (EGB), van der Waals Energy with Restraints (VDW, RESTRAINT), Electrostatic Energy with Restraints (EEL, RESTRAINT), and Surface Energy (ESURF) for each PDB file. Subsequently, we'll organize this data into a dataset where each row corresponds to a PDB file, including a column for the associated melting temperature. Then, employing regression models like Linear Regression, Decision Trees, or Random Forests, we'll train the model on the dataset. Finally, we'll evaluate the model's performance using metrics such as Mean Squared Error or Mean Absolute Error to gauge its ability to predict the melting temperature based on the provided energy values.



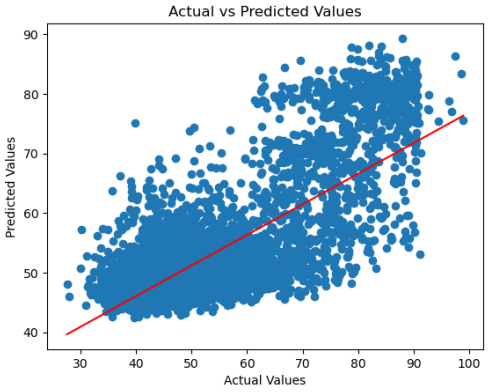
Chapter 3

RESULTS-

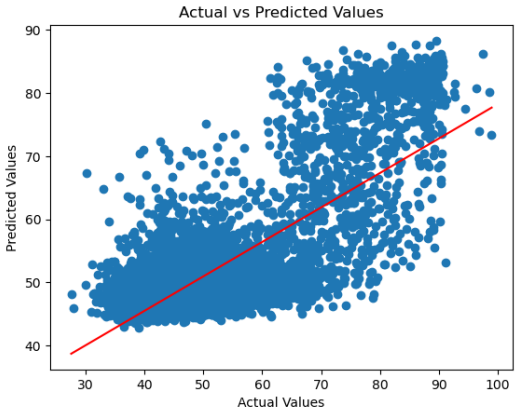
METHOD 1

MODEL	MSE	MAE	RMSE	R2
Gradient Boosting Regressor	55.15	5.56	7.42	0.53
Extreme Gradient Boosting (XG Boost)	55.79	5.54	7.47	0.53
Light Gradient Boosting	51.39	5.35	7.16	0.56
Random Forest Regressor	52.45	5.3	7.46	.55

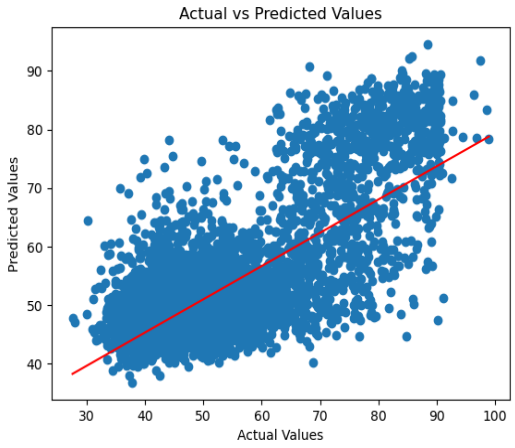
Plot of Actual VS Predicted Values



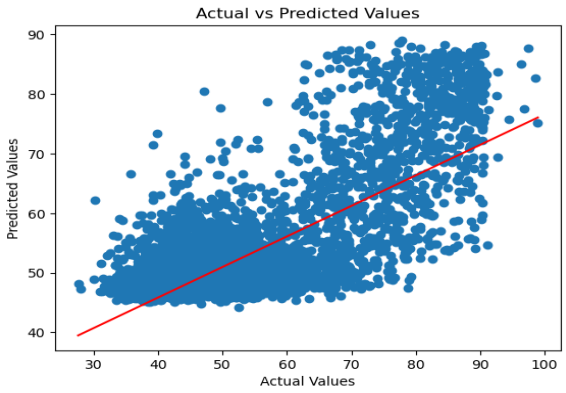
Random Forest Regressor



Light Gradient Boosting



Extreme Gradient Boosting

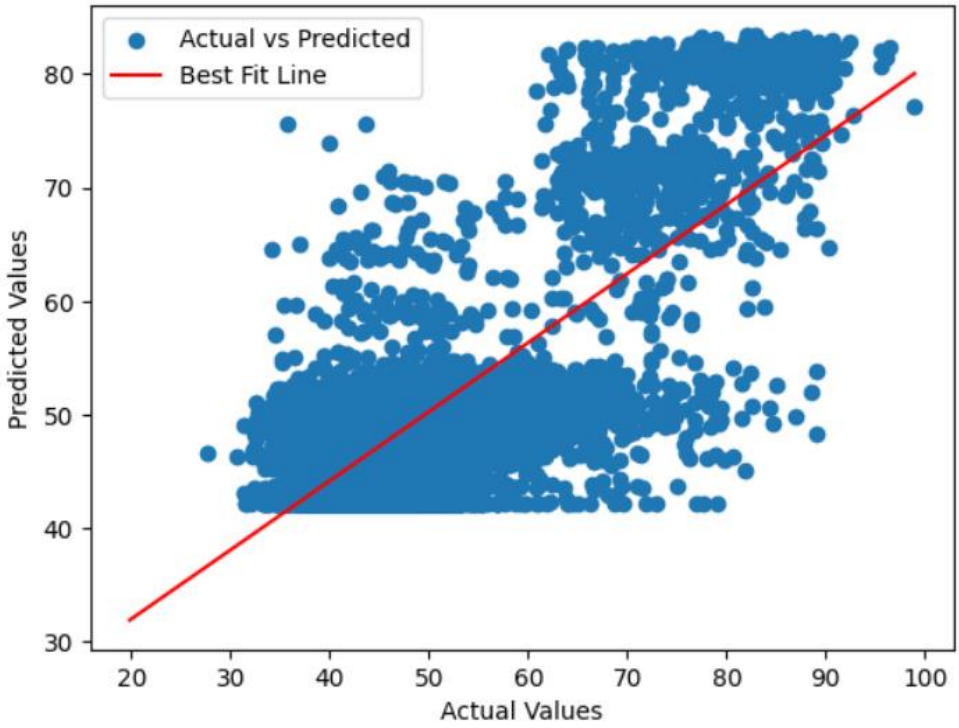


Gradient Boosting Regressor

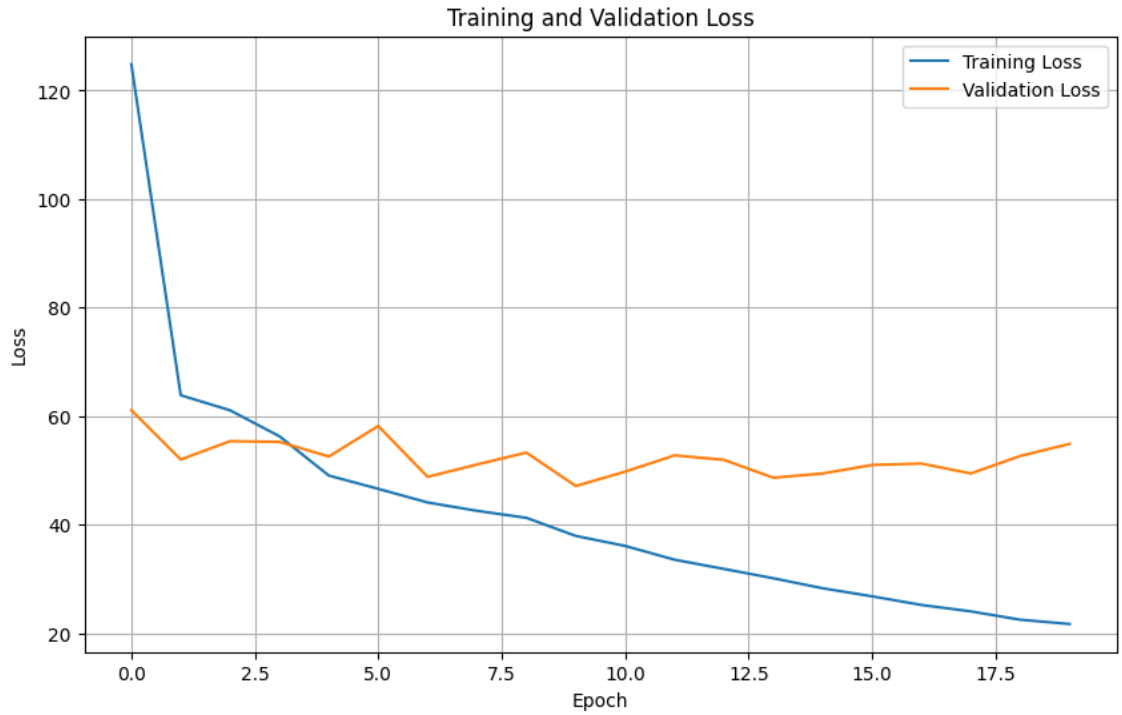
METHOD 2-

MODEL	MSE	MAE	RMSE	R2
ANN	47.181	5.03	6.86	0.593

Plot of Actual VS Predicted Values



Training loss VS Validation loss



METHOD 3-

MODEL	MAE	R2	MSE
Random Forest	7.67	0.20	108.061
Polynomial Regression	7.80	0.18	110.535
Ridge Regression	8.24	0.114	110.53
Linear Regressor	8.24	0.114	120.021

METHOD 4-

MODEL	MAE	R2	MSE
Xg boost	7.93	0.16	116.01
Polynomial Regression	8.335	0.11	124.04
Random forest	7.68	0.22	107.92
Light Gradient Boosting	7.75	0.20	111.49

Chapter 4

Conclusion

- The pfeatures on the ANN model produced the most promising results compared to other feature-based models, with an
- RMSE of 6.86
- MSE of 47.18
- R2 score of 0.5
- MAE of 5.03.

Comparison

I used data from Prostab2, which employed sequence-based features to train their machine learning model. Similarly, I generated sequence-based features (pfetures) and trained an artificial neural network model. As a result, I achieved better performance compared to Prostab2.

Errors	My model	ProsTab2
RMSE	6.86	9.09
R2score	0.59	0.58
MSE	47.18	82.75
MAE	5.03	6.93

Future Scope

Integrating 3D CNNs, leveraging large language models, and incorporating transformer attention layers to enhance model performance and capture intricate patterns in the data.

[1] Y. Yang, J. Zhao, L. Zeng, and M. Vihinen, “ProTstab2 for Prediction of Protein Thermal Stabilities,” *Int. J. Mol. Sci.*, vol. 23, no. 18, p. 10798, Sep. 2022, doi: 10.3390/ijms231810798.

[2] M. Li, H. Wang, Z. Yang, L. Zhang, and Y. Zhu, “DeepTM: A deep learning algorithm for prediction of melting temperature of thermophilic proteins directly from sequences,” *Comput. Struct. Biotechnol. J.*, vol. 21, pp. 5544–5560, 2023, doi: 10.1016/j.csbj.2023.11.006.

[3] B. Kumwenda, D. Litthauer, Ö. T. Bishop, and O. Reva, “Analysis of Protein Thermostability Enhancing Factors in Industrially Important *Thermus* Bacteria Species,” *Evol. Bioinforma.*, vol. 9, p. EBO.S12539, Jan. 2013, doi: 10.4137/EBO.S12539.

[4] F. Jung, K. Frey, D. Zimmer, and T. Mühlhaus, “DeepSTABp: A Deep Learning Approach for the Prediction of Thermal Protein Stability,” *Int. J. Mol. Sci.*, vol. 24, no. 8, p. 7444, Apr. 2023, doi: 10.3390/ijms24087444.

[5] A. D. Robertson and K. P. Murphy, “Protein Structure and the Energetics of Protein Stability,” *Chem. Rev.*, vol. 97, no. 5, pp. 1251–1268, Aug. 1997, doi: 10.1021/cr960383c.