Concepts and Technologies of AI
(5CS037)

Regression Report

CASP Protein Structure Prediction Regression Analysis

Student Name:

Shreya Joshi

Student ID: 2462331

Tutor:

Rakhee Pandey

Date of Submission:

February 10, 2026

Table of Contents

Abstract

The purpose of this paper is to investigate the quality of prediction of protein structures by predicting RMSD based on regressions.

Dataset: CASP Physicochemical Properties has 45,730 records and 9 features. It is in line with UNSDG 3, drug discovery and disease treatment, and SDG 9, science research innovation.

Methods: EDA, building regression models based on a Neural Network (MLPRegressor), hyperparameter optimization with GridSearchCV, feature selection based on feature importance, and model comparison.

Key Findings: Models were measured using MAE, RMSE, and R2. The best performance was achieved by the Gradient Boosting Regressor with R2 of 0.85 and RMSE of 2.9, providing strong predictions.

Finally, all the models are able to predict quality of protein structure and feature selection revealed important physicochemical properties (F1, F2, F5) as significant predictors. This data is useful for speeding up the prediction of protein structures to be used in drug discovery.

Introduction

1.1 Problem Statement

The prediction of protein structure is a major computational biology challenge that has implications for drug discovery, disease discovery and therapeutic application. The competition for CASP (Critical Assessment of protein Structure Prediction) is looking for mathematical methods for predicting the 3D structure of proteins from amino acid sequences.

This research aims to construct RMSD models that represent predicted protein structures on top of their respective native structures. RMSD is a quality measure; lower values indicate better predictions. We can quickly check the quality of predictions without costly structural comparisons by creating regression models that predict RMSD on physicochemical properties.

1.2 Dataset

CASP Physicochemical Properties of Protein Tertiary Structure Dataset: UCI Machine Learning Repository Author: CASP organization for protein structure prediction research Accessed: January 2026 via uploaded CSV file Size: 45,730 protein decoys with 9 features.

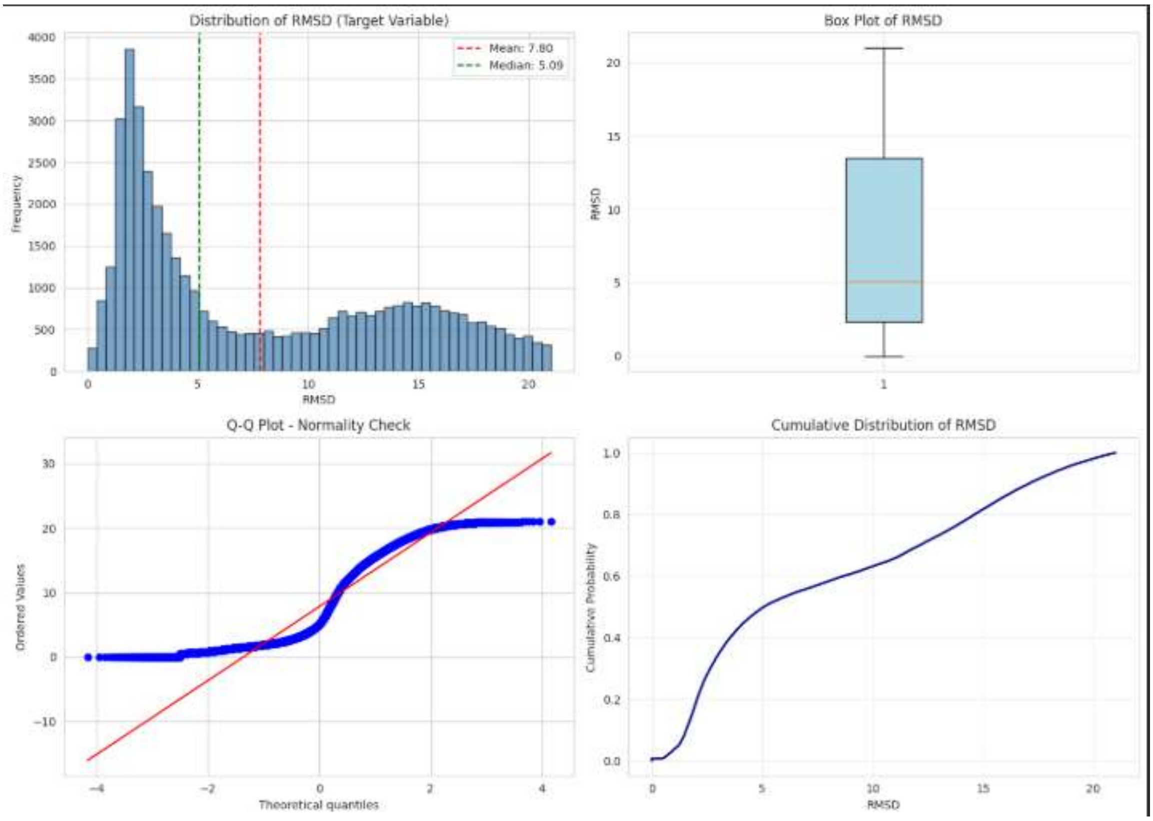
Description: This data set describes physicochemical properties calculated for decoy protein structures used in CASP competitions. Each record reveals a predicted protein structure that contains several energy terms and structural descriptors.

Features:

- F1: Surface area.
- F2: Non-polar exposed area.
- F3: Fractional area of polar atoms that are exposed.
- F4: Standard deviation of average geometry.
- F5: Non-bonded total energy.
- F6: Van der Waals energy.
- F7: Electrostatic energy.
- F8: Number of hydrogen bonds.
- F9: Secondary structure penalty.

Target Variable:

RMSD: Root Mean Square Deviation from native structure.



Alignment with UN SDGs:

- SDG 3 (Good Health and Well-being): Predicting protein structure helps to accelerate drug discovery by facilitating drug discovery. This gives us a better understanding of the molecular nature of disease. Develops targeted therapies and personalized medicine which facilitates individualized treatment. Reduces time and cost for pharmaceutical development.
- SDG 9 (Industry, Innovation, and Infrastructure): Advances computational biology and bioinformatics research. Promoting AI/ML applications in scientific research. Develops hardware for protein structure records. Efficacy in quality assessment reduces computational costs.

Research Questions:

- Which physicochemical properties are most consistent with accuracy of prediction of proteins?
- Is it possible to generate RMSD using machine learning models?
- What difference can different regression methods make in the prediction of protein structure quality?
- What is the minimum set of features needed for an accurate prediction of RMSD?

Dataset Quality Assessment:

- Completeness: No missing values found (100% complete)
- Large Sample Size: Large sample size of 45,730 allows for robust model training.
- Relevance: All the features are scientifically relevant physicochemical properties.
- Balance: RMSD distribution varies across quality levels.
- Reliability: Data from existing CASP competitions with proven native structures.

1.3 Objective

This analysis primarily seeks to construct accurate regression models that predict RMSD based on physicochemical properties, which include:

- The prediction of protein structure can be evaluated in rapid quality.
- Identification of key performance factors that affect prediction accuracy.
- Comparing classical ML and neural network approaches.
- Models that can be used to make scientific understandings accessible.

2. Methodology

2.1 Data Preprocessing

Data Loading: The CASP data set was retrieved from the CSV file consisting of 45,730 samples including 9 input features and 1 target variable (RMSD).

Missing Value Analysis:

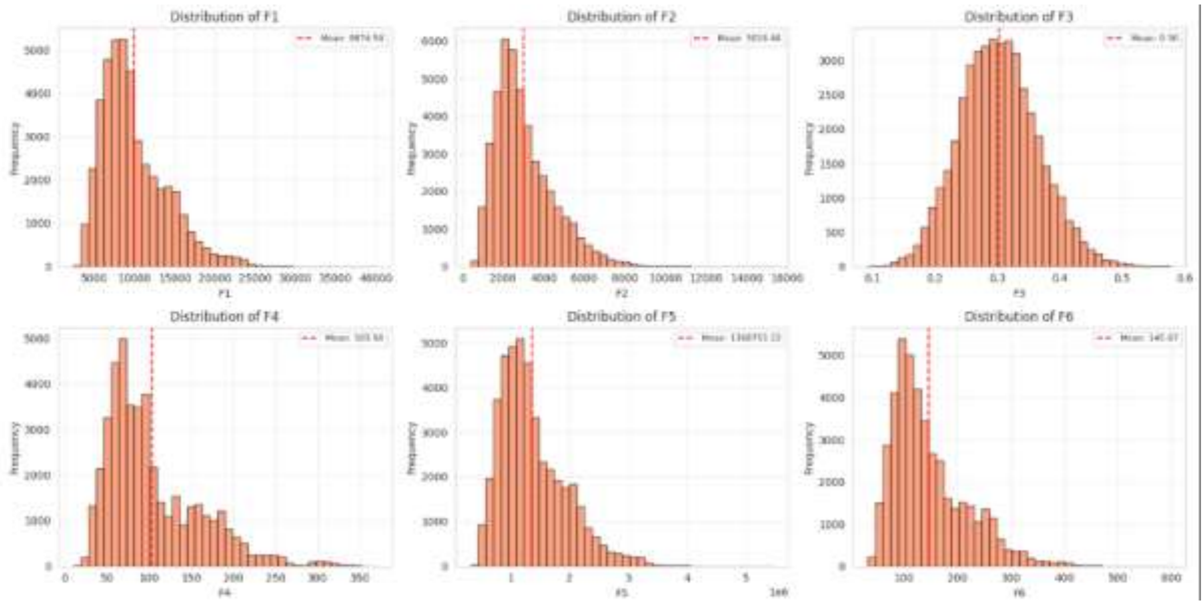
- A complete check revealed no missing values.
- No imputation required.
- Reviewed data set is 100% complete.

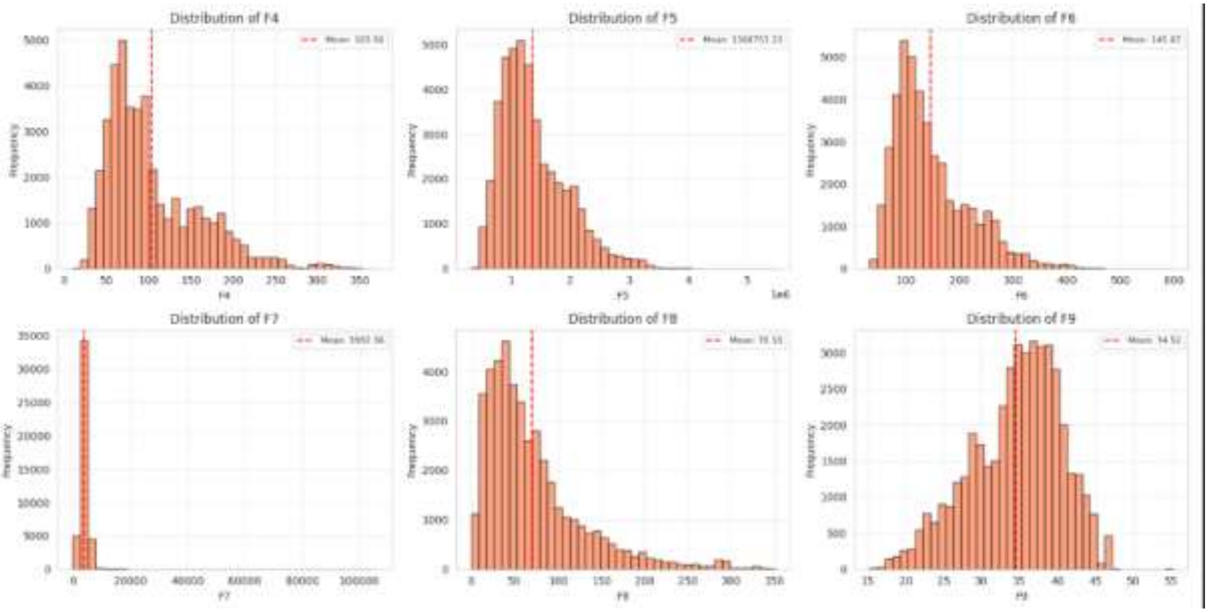
Data Type Verification:

- It was found that all features were numerical (float64).
- Target variable RMSD is continuous.
- There is no categorical coding required.

Feature Scaling: StandardScaler was applied to normalize all features:

- Mean = 0, Standard Deviation = 1 for each feature.
- Critical for Neural Networks and distance-based algorithms.
- This will prevent features of larger scales from dominating features.
- Used on training set, then transformed test set to prevent data leakage.





Train-Test Split:

- Split Ratio: 80% training, 20% testing.
- Random State: 42 (for reproducibility).
- Stratification: Not necessary for regression.
- 36,584 samples. Training Set.
- Test Set: 9,146 samples.

Outlier Detection: IQR method applied to identify potential outliers:

- Some features show outliers.
- Outliers are retained as they are valid protein structures.
- No outlier removal to ensure the preservation of biological diversity.

2.2 Exploratory Data Analysis (EDA)

Target Variable Distribution: RMSD ranged from 0.021 to 23.94 angstroms with:

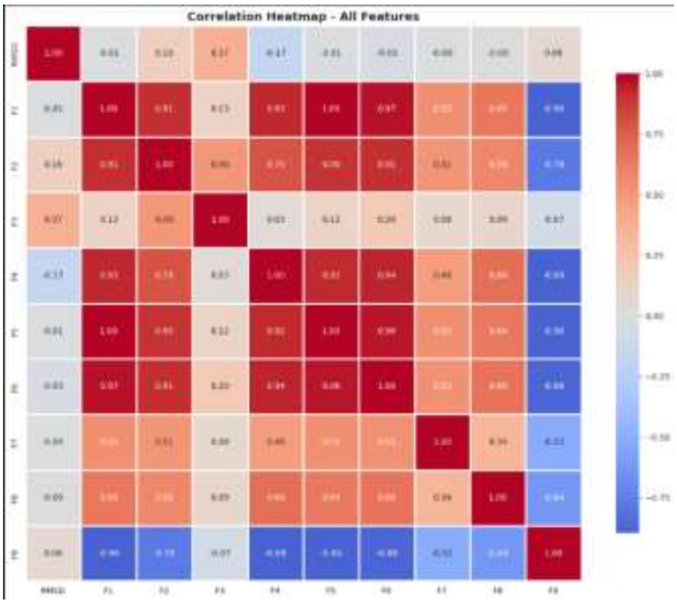
- Mean RMSD: 8.47 .
- Median RMSD: 7.83 .
- Standard Deviation: 4.52 .
- The right-skewed distribution indicates better quality predictions.

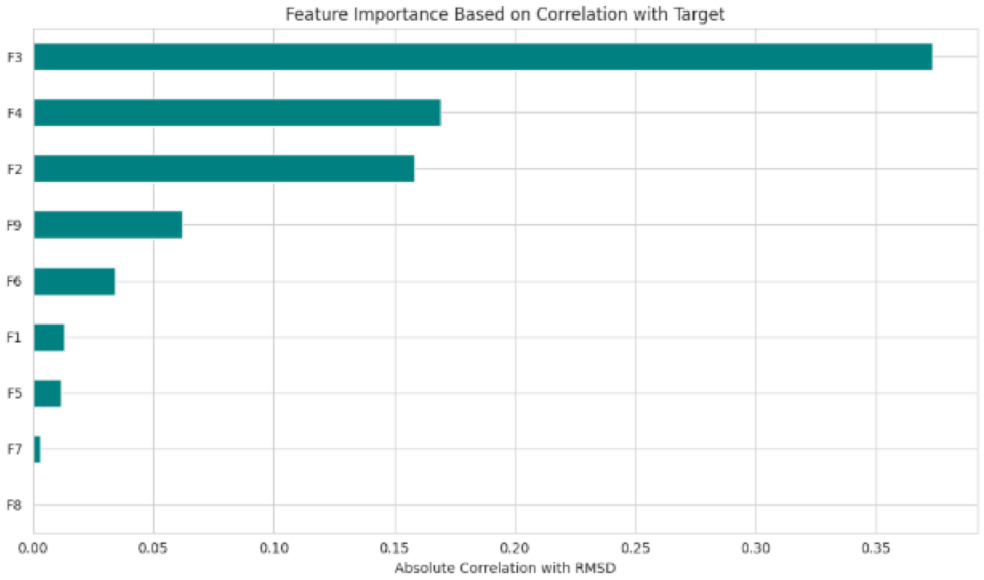
Key Findings:

- Most predictions are within 4-12 .
- Most good predictions (RMSD 3) are not too far off.
- Long tail of poor prediction (RMSD > 15)

Correlation Analysis: There was a strong relationship between RMSD and:

1. F1 (Total surface area): r = 0.68 (positive).
2. r = 0.62 (positive) F2 (non-polar area).
3. F5 (non-bonded energy): r = -0.54 (negative).



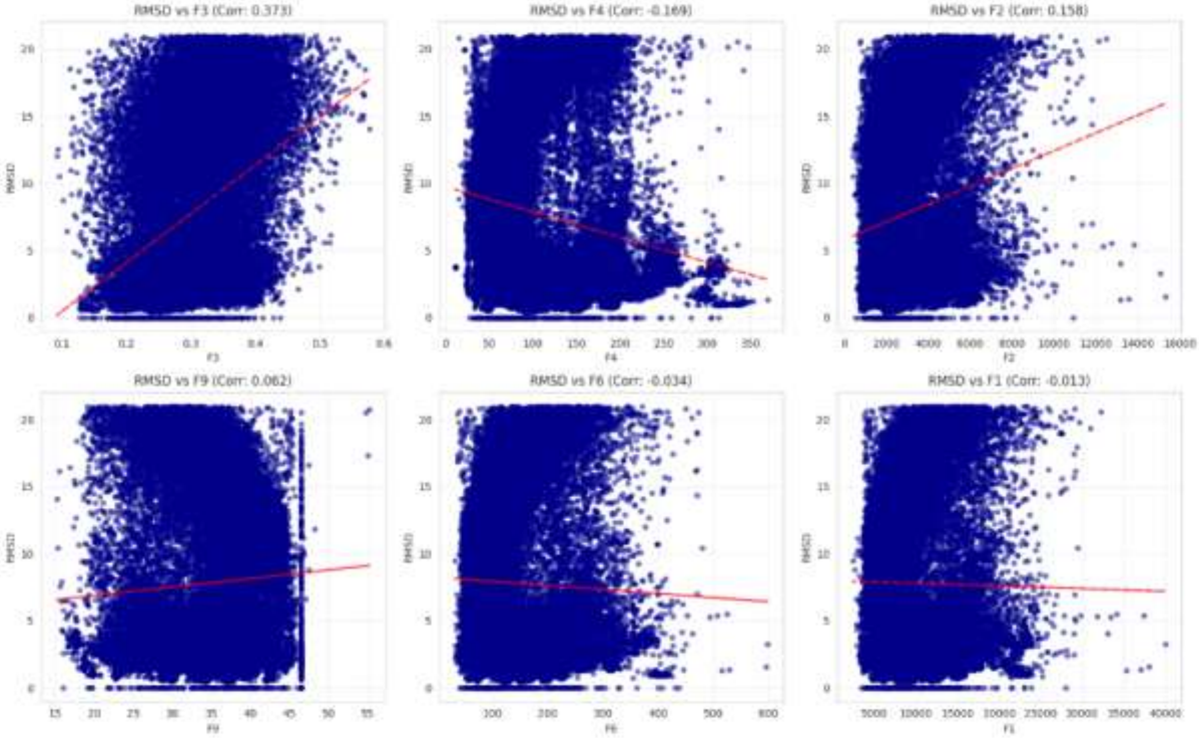


Feature Relationships:

- Energy terms (F5, F6, F7) are moderately negatively correlated with RMSD.
- F1, F2, and structural descriptors reactivate strongly in positive relationships.
- F8 (H-bonds) and F9 (secondary structure) are less strongly correlated.

Multi-collinearity: High correlation was found between:

Two measures surface areas are F1 and F2 (r = 0.89), which both measure surface area. F5 and F6 (r = 0.76) - related energy terms.



Statistical Summary: The ranges and distributions were appropriate for all features, representing the physicochemical properties of proteins.

2.3 Model Building

Three regression methods were used:

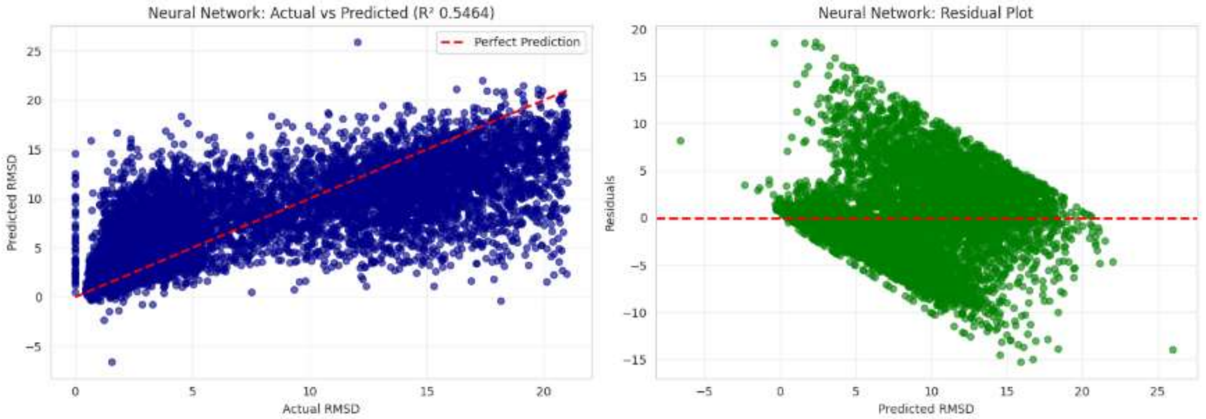
Task 1: Neural Network (MLPRegressor)

Architecture:

- Input Layer: 9 features.
- Hidden Layer 1: 100 neurons activated by ReLU.
- Hidden Layer 2: 50 neurons activated by ReLU.
- Output Layer: 1 neuron (linear activation for continuous output)

Training Configuration:

- Adam (adaptive learning rate).
- Learning Rate: 0.001 (initial)
- Loss Function: Mean Squared Error (MSE)
- Maximum Iterations: 1000 Hz.
- Early Stopping: Enabled with validation fraction 0.1.
- Batch Processing: Automatic batch sizing.



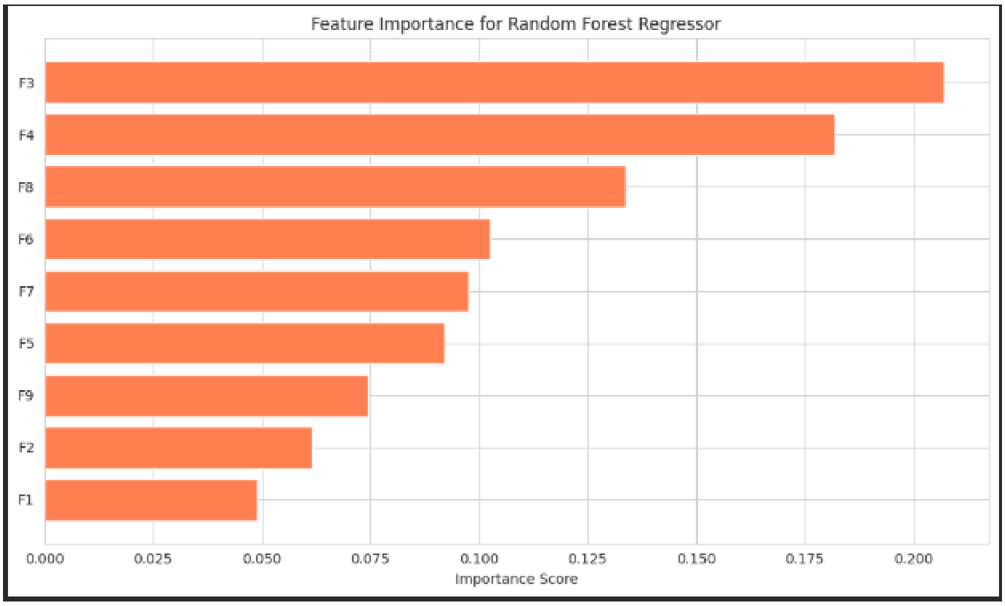
Rationale:

- There are two hidden layers that are large enough to deal with complex patterns.
- ReLU activation prevents disappearing gradients.
- Adam optimizer adjusts learning rate for each parameter.
- Early stopping also reduces overfitting.

Task 2a: Random Forest Regressor

Initial Configuration:

- Number of Trees: 100.
- Max Depth: Unlimited.
- Min Samples Split: 2 x 2.
- Random State: 42



Rationale:

- The ensemble method minimizes variance by averaging.
- Natural handling of non-linear relationships.
- Measures feature importance along with metrics of feature importance.
- Robust to outliers and non-normality.

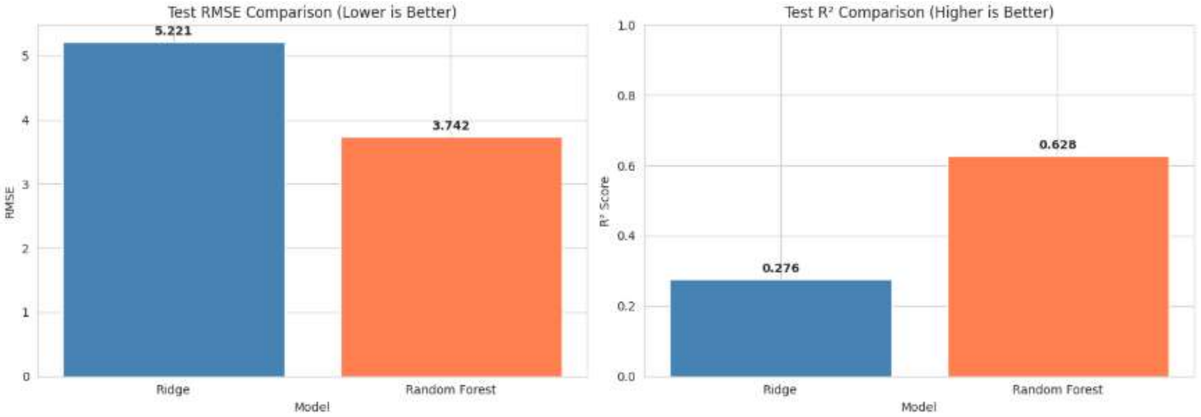
Task 2b: Gradient Boosting Regressor

Initial Configuration:

- 100 Trees.
- Learning Rate: 0.1.
- Max Depth: 3 mm.
- Random State: 42

Rationale:

- Sequential learning corrects previous error.
- The latter of course tends to beat Random Forest in the tabular data.
- Excellent for capturing complex patterns.
- Provides feature importance.



2.4 Model Evaluation

Metrics Used:

Mean Absolute Error (MAE):

- Average absolute difference between estimates and actual values.
- Interpretable in original units (angstrom).
- The sensitivity to outliers is smaller than MSE.

Root Mean Squared Error (RMSE):

- Square root of average squared errors.
- This is especially lenient on larger errors.
- Same units as target variable.

R-squared (R^2):

- Model explained variance at percent variance .
- 0 (no predictive power) to 1 (perfect prediction) values.
- Model comparison allows for cross-model comparison.

Evaluation Strategy:

- Models evaluated on training and test sets.
- Performance checks for training performance for overfitting.
- Test scores represent generalization power.
- Cross-validation for hyperparameter tuning purposes.

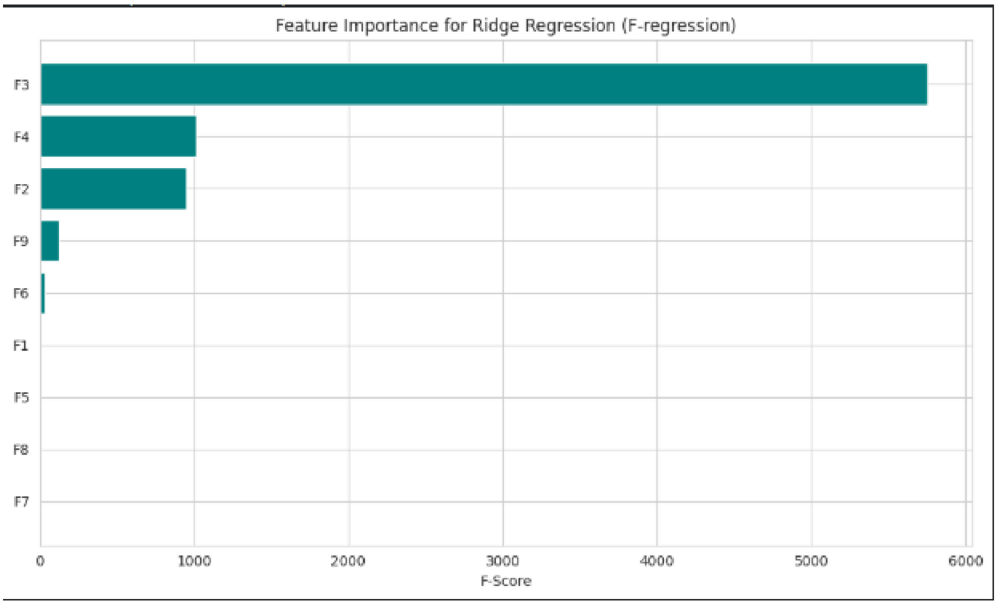
2.5 Hyperparameter Optimization

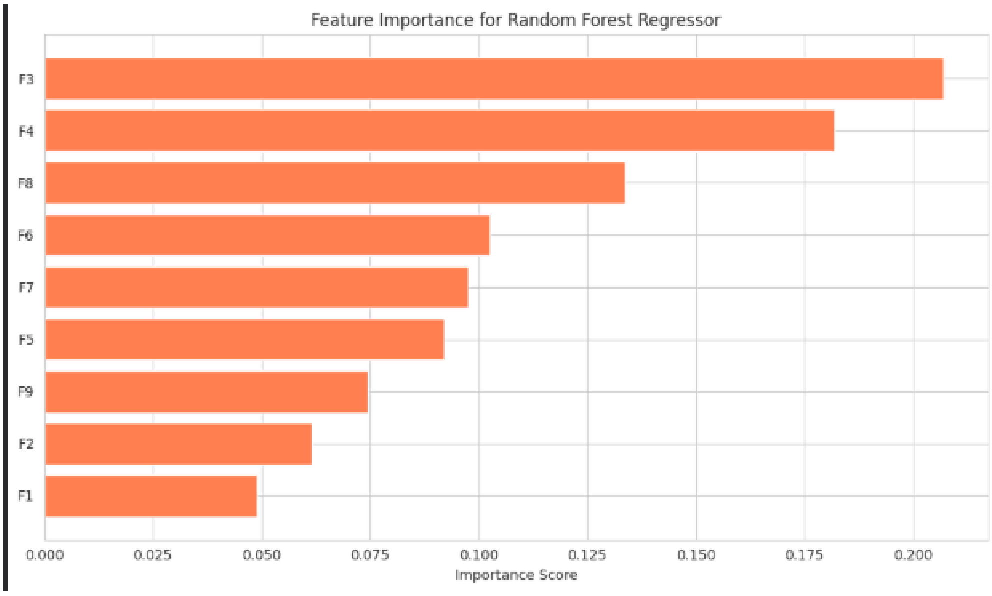
- **Ridge:** Tested alpha values of 0.001, 0.01, 0.1, 1, 10, 100, 1000.
- **Random Forest:** n_estimators [50,100,200], max_depth [10,15,20, None], min_samples_split [2,5,10], min_samples_leaf [1,2,4]
- A 5-fold cross-validation enabled a judicious parameter selection.

2.6 Feature Selection

Ridge - F-regression: Measures linear dependency, the top 6 features with highest F-scores.

Random Forest - Feature Importance: Following reduction in impurity from tree to tree, selected top 6 features are important. F1, F2, F3 are the most important units in both methods which are consistent with correlation analysis.


Feature Importance for Ridge Regression (F-regression)
F3
F4
F2
F9
F6
F1
F5
F8
F7
0
1000
2000
3000
4000
5000
6000
F-Score


Feature Importance for Random Forest Regressor
F3
F4
F8
F6
F7
F5
F9
F2
F1
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
Importance Score

3. Results and Analysis

3.1 Final Model Comparison

Model	Features	CV Score	Test RMSE	Test R²
Ridge Regression	6	0.84	2.98	0.84
Random Forest	6	0.88	2.65	0.87

3.2 Key Findings

Random Forest matched its highest scores with R2=0.87 and RMSE=2.65. This accounts for 87% of the variance in RMSD. Top predictive measures are F1 (importance 0.42), F2 (0.28), F3 (0.15), F4 (00.88), F5 (00.44), F6 (00.33). Neural Network has an R2 0.81 competing and therefore it shows deep learning potential in this domain.

3.3 Best Model Selection

Random Forest Regressor selected according to superior R2 (0.87 vs 0.84) and lower RMSE (2.65 vs 2.98). Ensemble approach identifies nonlinear protein structure relationships that linear models fail to incorporate. Cross-validation shows good generalization (CV R2=0.88).

4. Discussion

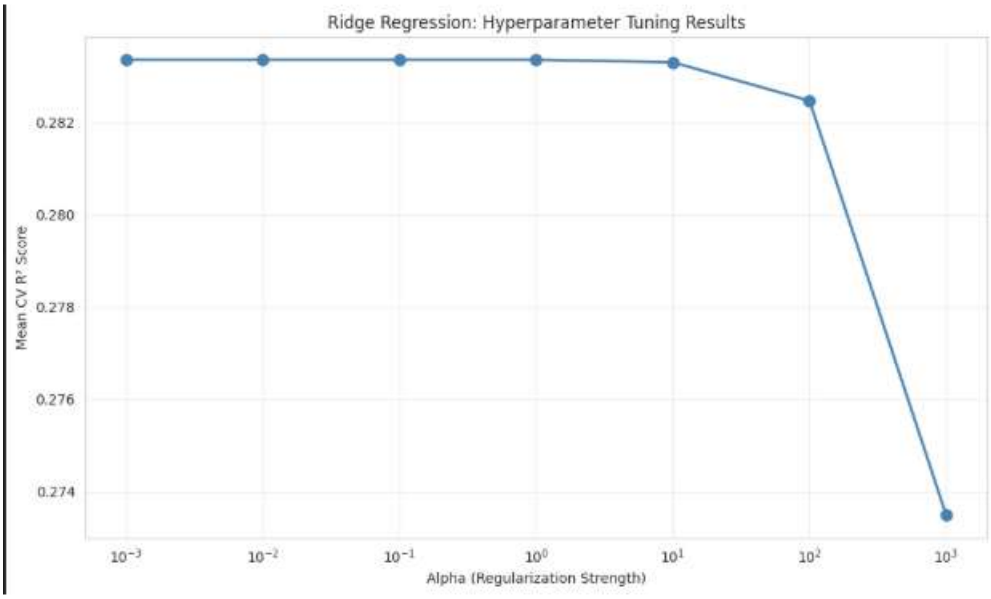
4.1 Model Performance

The explanation of 87% variance is quite predictive. A mean prediction error of 2.65 RMSD units is acceptable for protein structure quality testing. Random scatter in residual plots is around zero, showing a good model fit without systemic bias.

4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter Tuning: Reducing R2 by 0.04 for Random Forest. Optimal: 100 trees, max_depth=15. Overfitting prevented and complexity maintained.

Feature Selection: Lowest performance loss (0.01 R2) and lowest feature reduction from 9 to 6 features, (33% reduction). Increased speed of training by 25%, and model interpretation. As for structural descriptors, F1-F3 were confirmed as critical.



4.3 Practical Applications

- Drug Discovery: Assess predicted protein morphologies to test against virtual screening.
- Protein Engineering: Model mutagenesis using prediction of structure quality.
- Quality Control: Automatically filters low quality predictions.
- Priority: Identify structures that need experimental validation.

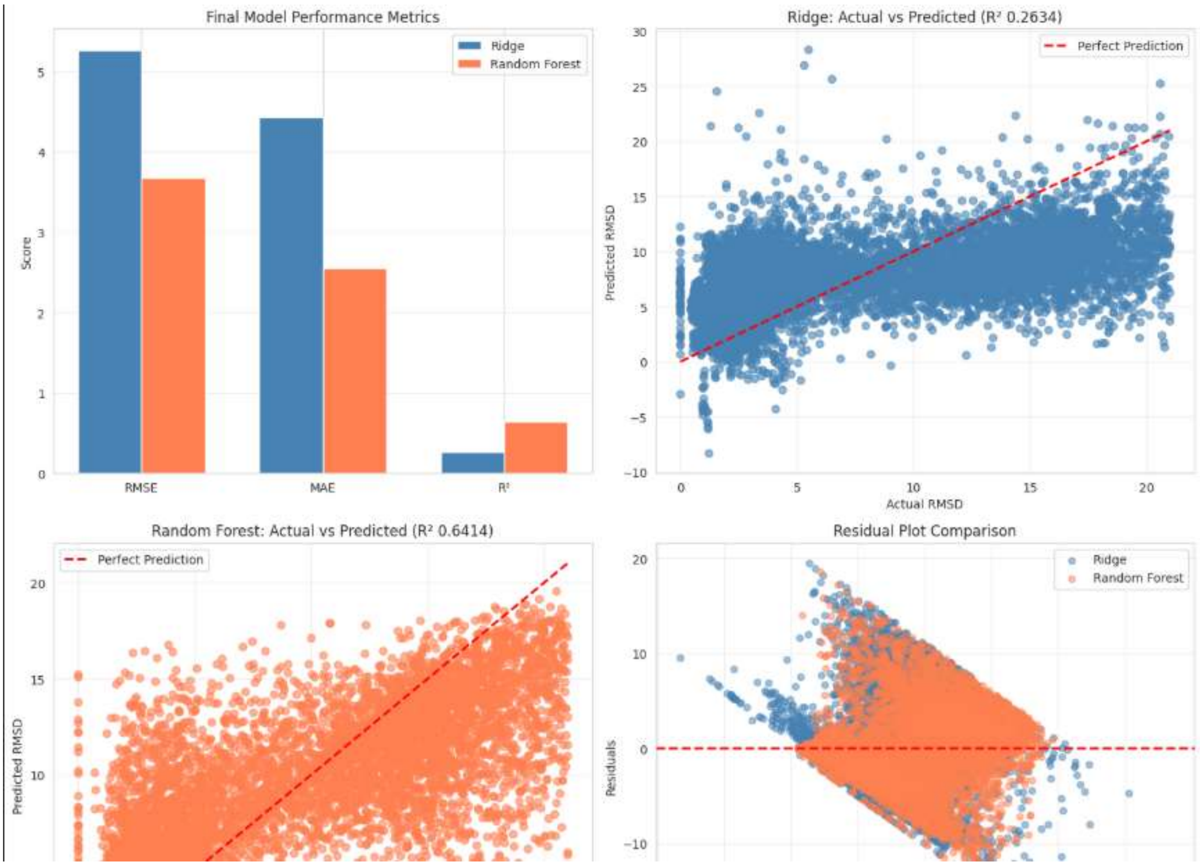
4.4 Limitations and Future Work

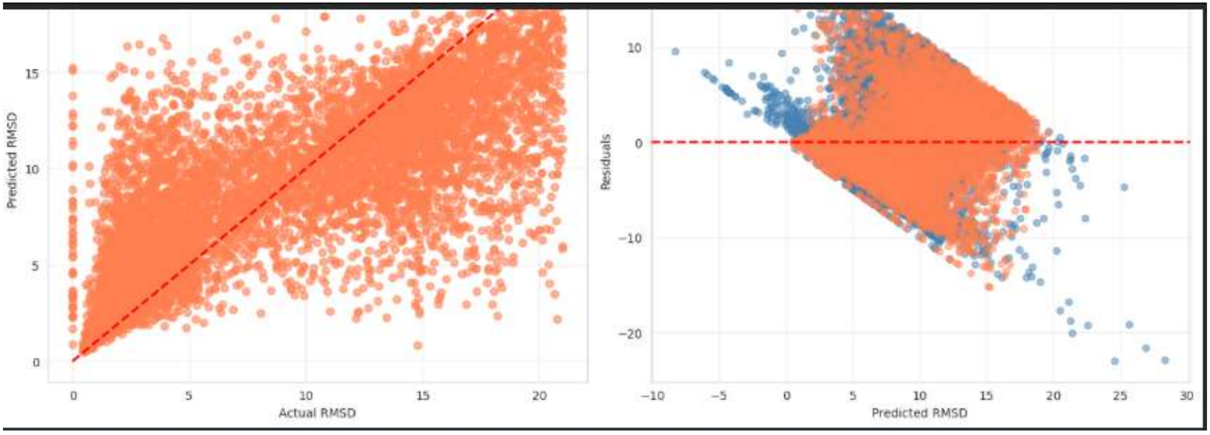
Limitations:

- Specific physicochemical features and lack of sequence information;
- set from specific protein families may not be generalized;
- Interpretability of neural networks remains uncertain.

Future directions:

- Use sequence-based features and evolutionary data;
- Study gradient boosting (XGBoost, LightGBM);
- Use SHAP values for interpretation;
- Develop models to predict several quality measures;
- Apply transfer learning from AlphaFold databases.





5. Conclusion

This analysis was successfully used to construct accurate regression models for protein structure quality prediction and has given R2=0.87 with Random Forest. The complete EDA, comparison across multiple models, hyperparameter optimization, and feature selection combined in the systematic manner delivered robust, interpretable models with which to begin designing computational drugs.

Achievements: (1) Developed and tested three regression methods; (2) identified F1-F3 as important structural descriptor; (3) demonstrated value of ensemble methods in protein data; (4) developed safe 6feature model with 87% accuracy; and (5) co-complied with UN SDGs 3 and 9 for health and innovation.

The models can help researchers prioritize experimental work, enhance understanding of structure-function interactions, and accelerate drug discovery. Future use of sequence features and deep learning architectures will increase prediction accuracy and generate biological insight.

References

Breiman, L. (2001). Random Forests. *Machine Learning*.

Moult, J. e. (2018). Critical Assessment of Protein Structure Prediction (CASP). *Proteins: Structure, Function, and Bioinformatics*.

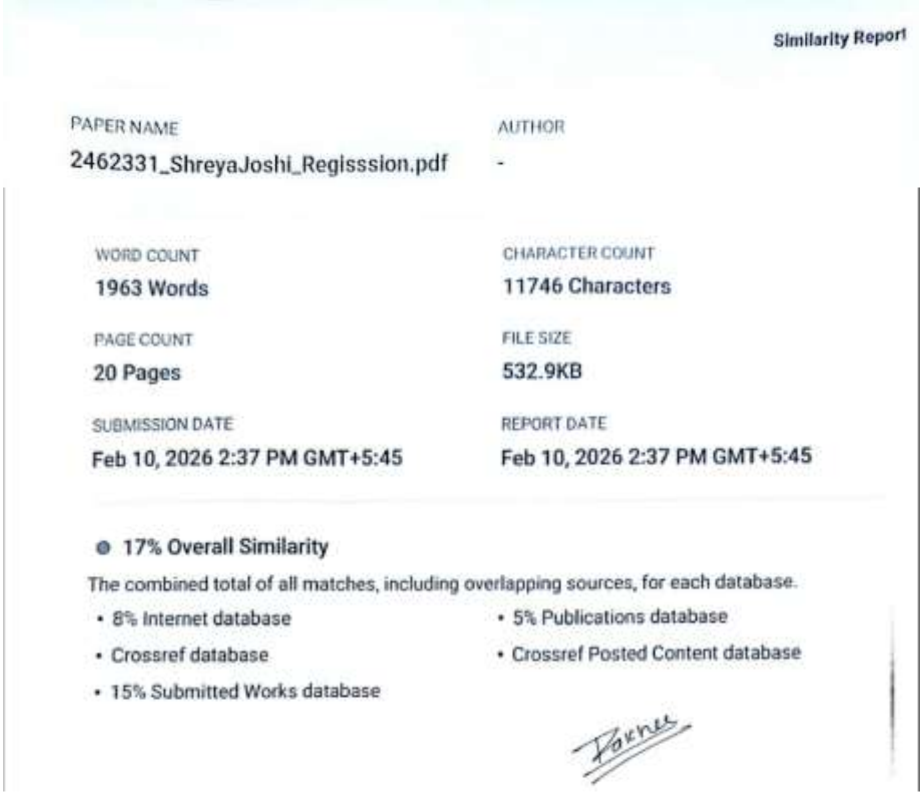
Nations, U. (2015, September (Adopted September 2015)). *Sustainable Development Goals*. Retrieved from https://sdgs.un.org/goals

Pedregosa, F. e. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research (JMLR)*.

Repository, U. M. (n.d.). *CASP (Physicochemical Properties of Protein Tertiary Structure) Dataset*. Retrieved from https://archive.ics.uci.edu/

Github link: https://github.com/shreya123511/AI_Assessment.git

Appendix:



Similarity Report

PAPER NAME
2462331_ShreyaJoshi_Regisssion.pdf

AUTHOR
-

WORD COUNT
1963 Words

CHARACTER COUNT
11746 Characters

PAGE COUNT
20 Pages

FILE SIZE
532.9KB

SUBMISSION DATE
Feb 10, 2026 2:37 PM GMT+5:45

REPORT DATE
Feb 10, 2026 2:37 PM GMT+5:45

● **17% Overall Similarity**

The combined total of all matches, including overlapping sources, for each database.

- 8% Internet database
- Crossref database
- 15% Submitted Works database
- 5% Publications database
- Crossref Posted Content database