University of Auckland

To What Extent Can the Mechanical Strength of β-Sheet-Based Protein Domains Be Predicted from Sequence and Structural Features Under Controlled Pulling Geometries?

Matthew Richard Chen Wei Lee

BE(Hons) / BSc

Department of Computer Science

February 2026

Contents

1 Abstract

The mechanical strength of β-sheet-rich protein domains is central to mechanobiology and biomaterials engineering, yet its predictability from sequence and structural descriptors remains uncertain. We curated 54 β-rich domains with reported maximum unfolding forces ($F_{\max}$) and engineered approximately 70 interpretable features from FASTA sequences, DSSP annotations, and PDB structures, capturing composition, topology, and packing properties.

Regularized linear models (Ridge and Lasso) and nonlinear ensemble methods (Random Forest and XGBoost) were evaluated using cross-validation in a low-sample, moderate-dimensional setting. Linear models failed to outperform a length-only baseline, indicating limited additive linear signal. In contrast, nonlinear models consistently explained approximately 30% of the variance in $F_{\max}$.

These results suggest that mechanostability is only partially predictable from coarse-grained descriptors and likely depends on nonlinear interactions among distributed structural and sequence features.

Code is publicly available at `https://github.com/matthue-lee/compsci_380`

2 Introduction

Proteins with β-sheet-rich architectures exhibit some of the highest known mechanical resistances among globular domains. Under externally applied force, typically measured via single-molecule force spectroscopy (SMFS) using atomic force microscopy (AFM) [7], these domains resist unfolding through shear-aligned hydrogen bond networks and densely packed strand topologies [8]. The resulting maximum unfolding force ($F_{\max}$) provides a quantitative measure of mechanostability, with implications for cellular mechanotransduction, extracellular matrix resilience, and the design of protein-based biomaterials.

Despite growing experimental measurements, predicting mechanical strength from sequence and structure remains unresolved. While features such as strand length, packing density, hydrophobic content, and topology are thought to contribute [10], the quantitative relationship between these descriptors and unfolding force is unclear. Moreover, available datasets are small, heterogeneous, and derived from varying experimental conditions, complicating statistical inference.

From a computational perspective, this problem can be framed as a supervised regression task: given sequence-derived, secondary-structure, and structural descriptors, can we predict $F_{\max}$ under controlled pulling geometries? However, this setting presents several challenges: low sample size ($n = 54$), moderate feature dimensionality ($p \approx 70$), multicollinearity among descriptors, and potential nonlinear feature interactions. These constraints necessitate careful model selection, regularization, and validation strategies

to avoid overfitting and inflated performance estimates.

In this study, we curate a focused dataset of β-rich protein domains from published articles, and engineer interpretable features from FASTA sequences, DSSP annotations, and PDB coordinates. Thenm we systematically evaluate both regularized linear models (Ridge and Lasso) and nonlinear ensemble methods (Random Forest and XGBoost). The objective is not only to maximize predictive performance, but to quantify the extent to which mechanostability is statistically predictable from broad descriptors alone.

Conducted within a six-week research window, this study emphasizes reproducible methodology and careful cross-validation to establish a realistic lower bound on predictive performance.

3 Data Collection

Collecting an adequate dataset proved to be a significant challenge. Initially a database, the Biomolecule Stretching Database (BSDB), promised simulated stretching values for some 17,000 proteins. However, with time this database has been lost. Many papers mention or record various stretching metrics, under different conditions, and reported differently. This meant that an option would have been to curate a dataset by hand, recording and regularizing between experimental set-ups. Due to time constraints, this wasn't feasible, and we had to look in another direction for the data.

MechanoProDB was mentioned in publications as a newer alternative to the BSDB, however when considering β-sheet rich proteins domains, the data was limited.

Accordingly, the primary dataset used for this research is a number of Fmax values as reported by Sulkowska et. al. [8]. This article highlighted 137 strong proteins, along with unfolding force, PDBID, and CATH classification. From this dataset, after cleaning and pruning, along with filtering down to only consider β-rich domains, left 54 samples. This number of samples is lower than desirable, and as such it largely informed the decisions on which models to implement.

4 Methods

4.1 Exploratory Analysis

After data collection, exploratory analysis was performed to understand variable distributions and relationships. This included pairwise plots, Pearson correlation analysis, numeric distributions, and relationships with $F_{\max}$.

The following features were available in the curated dataset:

- N (residues)

- $F_{\max}$ (ϵ Å^{-1}) (maximum unfolding force)
- L_n (Å) (mean β-strand length)
- L_m (Å) (mean segment length)
- L_f (Å) (longest β-strand length)
- Pattern (topology classification)
- CATH (structural classification)

L_f was shown to have direct correlation with N, so was dropped accordingly.

Numeric Distributions

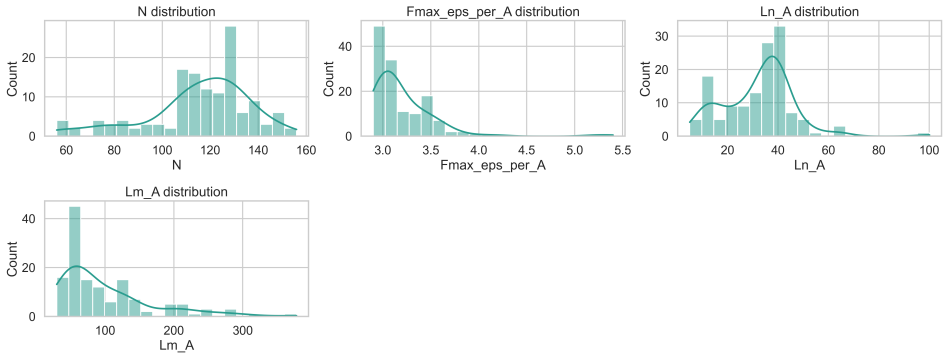

Figure 1: Distributions of numerical features within the dataset. The histograms reveal feature scale, skewness, and variance prior to model standardization.

The response variable $F_{\max}$ (ϵ Å^{-1}) shows mild right-skew due to a small number of high-force domains. Length-derived features span broader ranges.

Pairwise Analysis

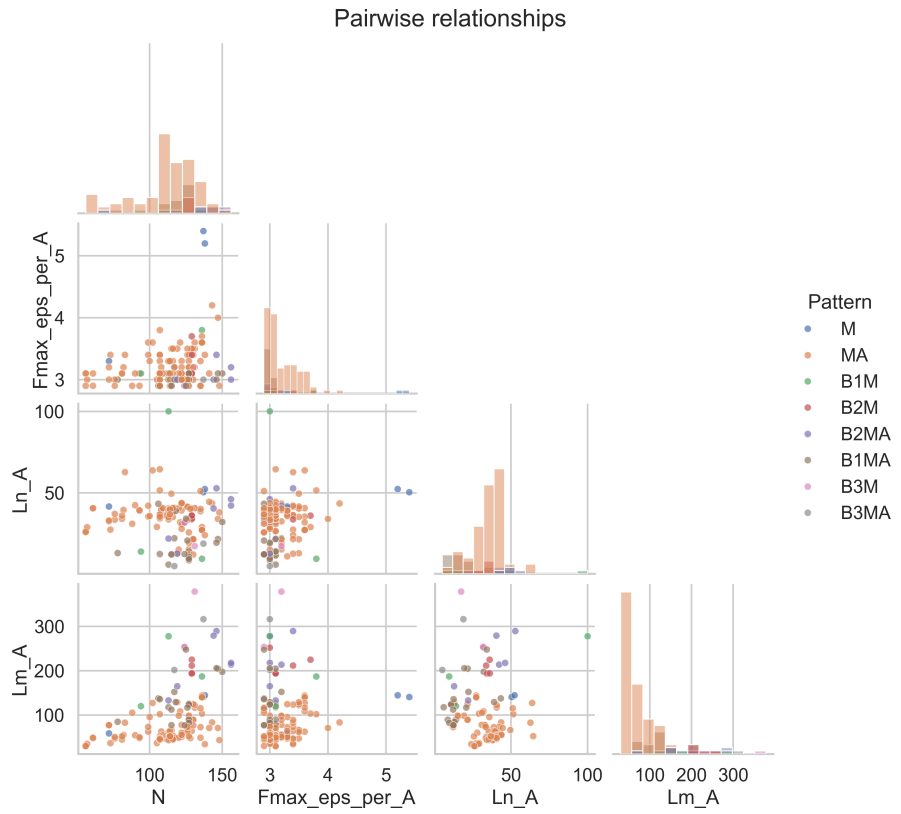

Figure 2: Predicted vs. true $F_{\max}$ values for the ridge regression model on the test set.

Pairwise plots (Figure 2) reveal strong redundancy among length-based variables and secondary-structure fractions, indicating substantial multicollinearity within the raw feature set. CATH class coloring highlights fold-dependent clustering, particularly among β-sandwich domains.

Pearson Correlation

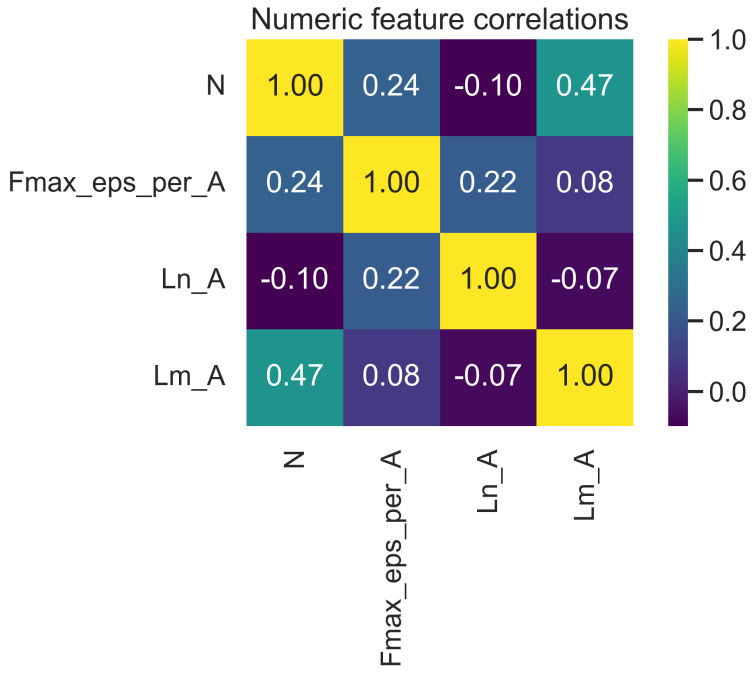

Figure 3: Pearson correlation heatmap of numerical structural descriptors and $F_{\max}$ per Å. Strong correlations indicate potential predictive features and reveal multicollinearity among descriptors.

Pearson correlations between numeric predictors and $F_{\max}$ remain moderate in magnitude, with no single dominant linear driver. Chain length exhibits the strongest positive association with $F_{\max}$, although the effect size remains modest (r = 0.24), reinforcing that mechanostability is not determined by 4Zsize alone

$F_{\max}$ vs Structural/Sequence Variables

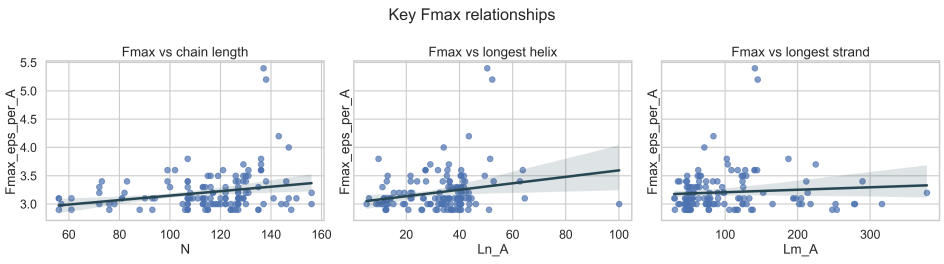

Figure 4: Scatter plots showing relationships between $F_{\max}$ per Å and selected structural descriptors. The plots provide visual insight into linear trends and potential nonlinear interactions.

Figure 4 shows modest positive trends between $F_{\max}$ and chain length, longest helix length, and longest β-strand length. While fitted lines indicate upward slopes, substantial scatter is present in all three relationships. Chain length exhibits the clearest linear signal, though the effect size remains small. Overall, these plots suggest that no single structural descriptor strongly determines mechanical strength, motivating the use of nonlinear models to capture distributed interactions.

4.2 Feature Engineering

Sequence Features (FASTA)

Sequence-derived features were computed directly from FASTA records to capture intrinsic biochemical and compositional determinants of mechanostability independent of three-dimensional structure. Length (N) was included as a baseline descriptor of domain size. Amino acid composition (20 fractional features) encodes the relative abundance of each residue type, allowing compositional biases to be quantified. Charge-related features (net charge, fraction of charged residues, and estimated iso-electric point) were incorporated to reflect electrostatic interactions that may influence hydrogen bonding and strand pairing under load [2, 6]. Hydrophobicity statistics (mean and variance) summarize the distribution of nonpolar residues, which contribute to core packing and resistance to unfolding [4, 3]. Collectively, these descriptors provide an interpretable representation of sequence-level properties that may modulate mechanical strength.

Secondary Structure Features (DSSP)

Secondary-structure descriptors were derived using DSSP to quantify fold-level organization beyond primary sequence composition. Fractional occupancies of helix, strand, and coil capture the overall structural balance of each domain, with particular emphasis on β-strand content given the focus on β-rich architectures. Segment counts and mean segment lengths provide information about structural fragmentation versus continuity, distinguishing compact, extended sheets from shorter or interrupted elements. The longest β-strand length and longest helix length were included to capture potential shear-resistant elements or extended helical segments that may influence unfolding pathways. Finally, transitions per residue were computed as a measure of structural segmentation, reflecting how frequently secondary-structure states alternate along the sequence.

Structural Features (PDB)

Structure-derived descriptors were computed from PDB coordinates to capture three-dimensional packing and topology beyond sequence and secondary-structure summaries. Contact density, defined as the number of Cα–Cα pairs within 8Å per residue, was

included as a proxy for intramolecular packing and mechanical connectivity [9]. The radius of gyration provides a measure of global compactness, distinguishing tightly packed domains from more extended architectures [2]. Solvent accessibility metrics (mean SASA and percent buried residues) were used to approximate core exposure and packing stability [5]. Finally, β-sheet topology proxies combining strand fraction and contact density were included to represent shear-resistant sheet architectures, which are hypothesized to contribute strongly to mechanical strength [1]. Together, these features summarize the geometric and packing characteristics most likely to influence force-induced unfolding.

4.3 Model Training

4.3.1 Linear Models

The choice of models to employ was heavily influenced by the available data. After exploratory analysis highlighted some possible linear relationships, it was of interest to see to what extent these models could predict Fmax. Lasso regressions excels when the number of predictors is large relative to observations; in the case our data 71 vs 54. Lasso regression performs well in these situations as it promotes driving less important features to 0, thus avoiding overfitting.

Ridge regression performs similarly, however while lasso regression prefers to 0 features, ridge regression shrinks weights towards 0, without reaching it. This again works to prevent overfitting, and reducing model variance. Ridge regressions is also of particluar use on datasets with high colinearity, as is seen in the pairwise analysis (Figure 2)

Data Preprocessing

The curated CSV was loaded and numeric fields were coerced appropriately. Rows missing the target (`Fmax_eps_per_A`) or length proxy (`N`) were removed. Remaining missing values in numeric predictors were imputed using column means to ensure dense feature vectors.

Train/Test Split & Scaling

Data were shuffled with a fixed seed and split into training (80%) and test (20%) sets. Feature standardization was performed using training-set means and standard deviations, with the same transformation applied to the test set to ensure scale-consistent regularization.

Ridge Regression

A linear model with L2 regularization was trained using batch gradient descent. The objective minimized mean squared error plus an L2 penalty ($\alpha\|w\|^2$), shrinking coefficients to reduce overfitting. Predictions were computed as $\hat{y} = w^T x + b$.

Lasso Regression

An identical training loop was used but with an L1 penalty ($\alpha\|w\|_1$). The L1 subgradient ($\alpha \cdot \text{sign}(w)$) encourages sparsity, effectively performing feature selection.

Baseline and Evaluation

A univariate linear regression using only the length feature (`N`) served as a baseline. All models were evaluated using RMSE and R^2, with optional K-fold cross-validation for robustness.

4.3.2 Non-linear Models

Data Preprocessing

The file `master_table_beta.csv` was loaded and columns were coerced to numeric where possible. Rows with missing target values (`Fmax_eps_per_A`) were removed. Predictor columns that were entirely missing or constant were dropped. This produced a dense feature matrix X and target vector y.

Model Pipelines and Imputation

Each nonlinear regressor was wrapped in a scikit-learn `Pipeline` with a `SimpleImputer` using the median strategy. This fills residual missing values column-wise and ensures both models receive dense inputs during cross-validation.

Random Forest Regression

A `RandomForestRegressor` was trained with 400 trees, shallow depth (`max_depth=5`), a minimum of 2 samples per leaf, and `max_features="sqrt"`. Training used parallel execution (`n_jobs=-1`) to improve efficiency and stability across folds.

XGBoost Regression

An `XGBRegressor` was trained with 600 boosted trees and a small learning rate (0.05). Regularization was applied via subsampling and column sampling, histogram-based tree construction (`tree_method="hist"`), and explicit L1/L2 penalties.

4.4 Evaluation

With the small sample size available (n 50) and high number of features (70) the risk of overfitting is non-trivial. All models were evaluated using consistent cross-validation procedures to ensure fair comparison across linear and nonlinear approaches. The dataset

comprised 54 β-rich protein domains with approximately 70 usable numeric features derived from sequence and structure. Given the relatively small sample size and moderate feature dimensionality, particular care was taken to avoid optimistic performance estimates.

To evaluate the accuracy of the models 5-fold CV was used. For the Ridge and Lasso 5-fold CV was additionally used to tune the hyperparameters, namely α, within each fold. Specifically, a grid of regularization strengths (α) was evaluated using an inner 5-fold split on the training data. The value of α minimizing average validation error was selected, and the model was retrained on the full outer training fold before evaluation on the held-out validation fold.

For the non linear models, in addition to the 5-fold CV, LOOCV was used. LOOCV was included as it can perform well on small sample sizes by maximising training data usage. It does however come at the cost of having a higher variance, but used in conjunction can offer a good insight into the capabilities of a model.

5 Results

5.1 Linear Models

Despite hyperparameter tuning, neither Ridge nor Lasso regression outperformed the length-only baseline. Ridge regression exhibited negative R^2, indicating performance worse than predicting the mean response. Lasso regression achieved near-zero R^2, suggesting minimal linear predictive signal across the full feature set.

These results indicate that additive linear combinations of the engineered sequence and structural descriptors do not substantially improve predictive accuracy over simple protein length. This suggests that either (i) the informative signal is weak relative to noise, or (ii) predictive relationships involve nonlinear interactions not captured by linear models.

Model	RMSE	R^2
Ridge (GD) ($\alpha \approx 4.42$)	0.2880	-0.2608
Lasso ($\alpha \approx 0.072$)	0.2553	0.0092
Length-only (N)	0.2471	0.0717

Table 1: Linear baseline performance under 5-fold cross-validation (54 domains, 71 features).

Model	CV	MAE	RMSE	R^2
Random Forest	5-fold	0.1565	0.2153	0.2956
Random Forest	LOOCV	0.1528	0.2128	0.3119
XGBoost	5-fold	0.1591	0.2143	0.3020
XGBoost	LOOCV	0.1525	0.2137	0.3059

Table 2: Nonlinear model performance under 5-fold cross-validation and LOOCV (54 domains, 70 features).

5.2 Non-linear Models

Nonlinear Model Performance

Table 2 summarizes the performance of Random Forest and XGBoost models under 5-fold cross-validation and leave-one-out cross-validation (LOOCV). Across both validation schemes, nonlinear models achieved moderate predictive accuracy, with RMSE values near 0.21 and R^2 values between 0.30 and 0.31.

Performance was highly consistent between 5-fold CV and LOOCV, suggesting that results are stable despite the limited sample size ($n = 54$). Random Forest marginally outperformed XGBoost under LOOCV ($R^2 = 0.312$ vs. 0.306), though differences between models were small. Mean absolute error remained near 0.15 across all configurations, indicating comparable error magnitudes across folds.

Feature Importance

Feature importance rankings from Random Forest (Figure 5) and XGBoost (Figure 6) reveal consistent emphasis on sequence-derived descriptors. In particular, global hydrophobicity and amino acid composition metrics contribute strongly to predictive performance, while individual structural length descriptors exhibit more modest influence. The similarity in importance profiles between the two ensemble methods suggests that predictive signal arises from distributed, nonlinear interactions among sequence chemistry and secondary-structure features rather than from a single dominant predictor.

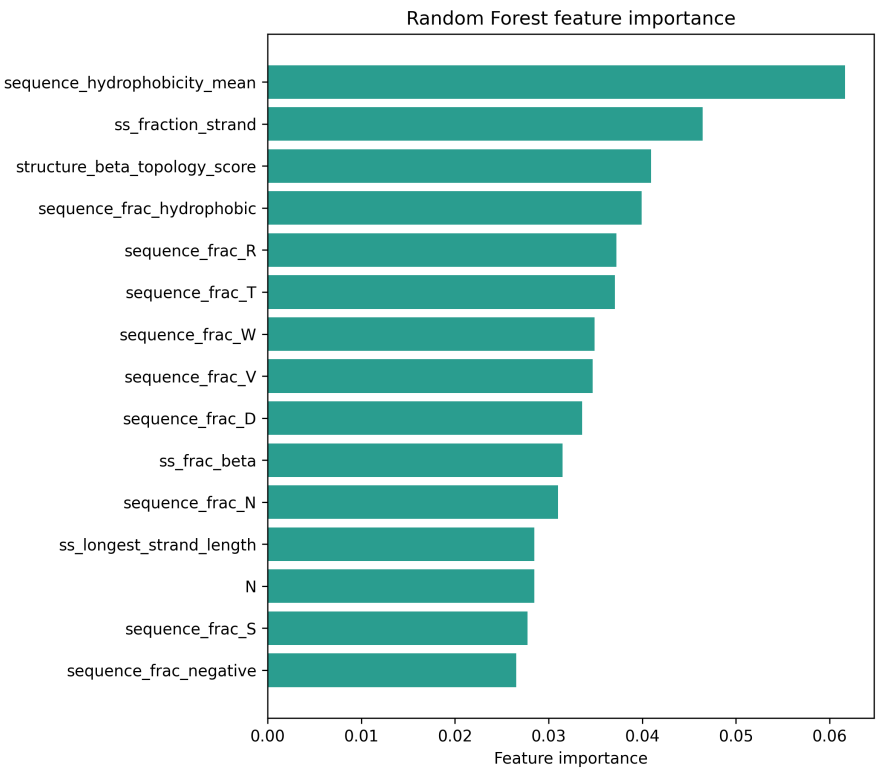

Figure 5: Feature importance scores derived from the random forest model. Higher values indicate greater contribution to reducing prediction error across the ensemble.

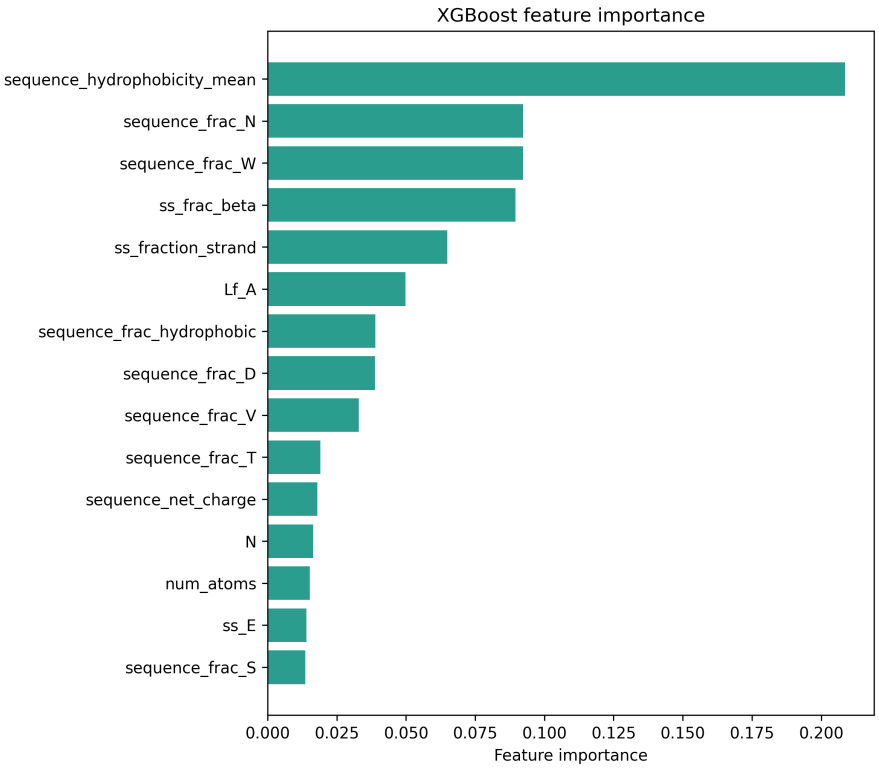

Figure 6: Feature importance scores from the XGBoost model, reflecting the relative contribution of structural and sequence-derived descriptors to predictive performance.

6 Analysis and Discussion

The results indicate a clear separation between the capabilities of linear and nonlinear approaches in predicting $F_{\max}$ for β-sheet-rich protein domains. Linear models (Ridge and Lasso) failed to outperform a simple length-only baseline, with Ridge yielding negative R^2 and Lasso achieving near-zero explanatory power. This suggests that additive linear combinations of the engineered descriptors do not sufficiently capture the underlying determinants of mechanical strength. While exploratory analysis revealed moderate pairwise correlations and multicollinearity among length-based features, no single descriptor exhibited strong linear predictive power.

In contrast, nonlinear ensemble methods demonstrated consistent improvement, achieving R^2 values near 0.30 under both 5-fold cross-validation and LOOCV. The similarity between validation schemes suggests that performance estimates are reasonably stable despite the limited sample size. Although an R^2 of 0.30 does not imply high predictive accuracy, it does indicate that nonlinear interactions among sequence and structural features explain a meaningful fraction of variance in unfolding force.

Feature importance analyses further support this interpretation. Both Random Forest and XGBoost emphasize sequence-derived descriptors—particularly hydrophobicity

and amino acid composition—over simple length measures. This aligns with biophysical expectations: hydrophobic packing, residue composition, and β-strand organization collectively influence shear resistance and hydrogen-bond network stability. Importantly, no single dominant feature emerged, reinforcing the hypothesis that mechanostability arises from distributed interactions rather than a singular structural determinant.

Taken together, these findings suggest that mechanical strength in β-sheet-based domains is only partially predictable from static sequence and structural summaries. Linear models appear insufficient to capture the complexity of force-induced unfolding, whereas nonlinear methods reveal modest but consistent predictive structure. The results support the view that mechanostability is governed by higher-order interactions and topology-dependent effects, which may require richer geometric or dynamical representations to model more accurately.

Overall, within the constraints of a small curated dataset and interpretable feature set, nonlinear machine learning models provide evidence that approximately 30% of the variance in $F_{\max}$ can be explained by sequence and structural descriptors alone. This establishes a quantitative lower bound on predictability under controlled pulling geometries and highlights clear directions for future model refinement.

6.1 Limitations

Several limitations should be considered when interpreting these results. First, the dataset is small ($n = 54$), limiting statistical power and increasing the risk of variance in cross-validation estimates, particularly for nonlinear models. Although LOOCV and 5-fold validation were used to improve robustness, performance estimates may still fluctuate with additional samples.

Second, the features were limited to basic sequence and structural descriptors, and may not capture more complex geometric or dynamic properties that influence mechanical stability. As a result, the models rely on simplified representations of protein structure.

Finally, feature importance estimates from ensemble models should be interpreted cautiously given the limited sample size, as tree-based importance metrics can be unstable in small datasets.

7 AI Acknowledgment

Portions of the data processing, model development, and manuscript preparation were supported by AI tools. Specifically, Anthropic Claude and OpenAI Codex were used to assist with code structuring and optimization, while Grammarly's AI features were used for clarity and grammar refinement in writing. Overleafs inbuilt AI was also used to assist in formatting the document.

8 Appendix

The full codebase and data processing pipeline are publicly available. GitHub repository: `https://github.com/matthue-lee/compsci_380`

References

[1] Hendrik Dietz and Matthias Rief. Exploring the energy landscape of gfp by single-molecule force spectroscopy. *PNAS*, 101(46):16192–16197, 2004.

[2] Ken A. Dill. Dominant forces in protein folding. *Biochemistry*, 29:7133–7155, 1990.

[3] Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 years on. *Science*, 338:1042–1046, 2012.

[4] Walter Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[5] B. Lee and F. M. Richards. The interpretation of protein structures: estimation of static accessibility. *Journal of Molecular Biology*, 55:379–400, 1971.

[6] C. Nick Pace, Hongxing Fu, Kim L. Fryar, Jason Landua, Sergio R. Trevino, Daniel Schell, Richard L. Thurlkill, Shigeki Imura, J. Martin Scholtz, and Gerald R. Grimsley. Contribution of hydrogen bonds to protein stability. *Protein Science*, 23:652–661, 2014.

[7] Matthias Rief, Mathias Gautel, Fritz Oesterhelt, Julio M. Fernandez, and Hermann E. Gaub. Reversible unfolding of individual titin immunoglobulin domains by afm. *Science*, 276(5315):1109–1112, 1997.

[8] Joanna Sulkowska and Marek Cieplak. Mechanical stretching of proteins - a theoretical survey of the protein data bank. *Journal of Physics: Condensed Matter*, 19:283201, 06 2007.

[9] Michele Vendruscolo, Edo Kussell, and Eytan Domany. Recovery of protein structure from contact maps. *Fold. Des.*, 2:295–306, 1997.

[10] Viola Vogel. Mechanotransduction involving multimodular proteins: converting force into biochemical signals. *Annual Review of Biophysics and Biomolecular Structure*, 35:459–488, 2006.