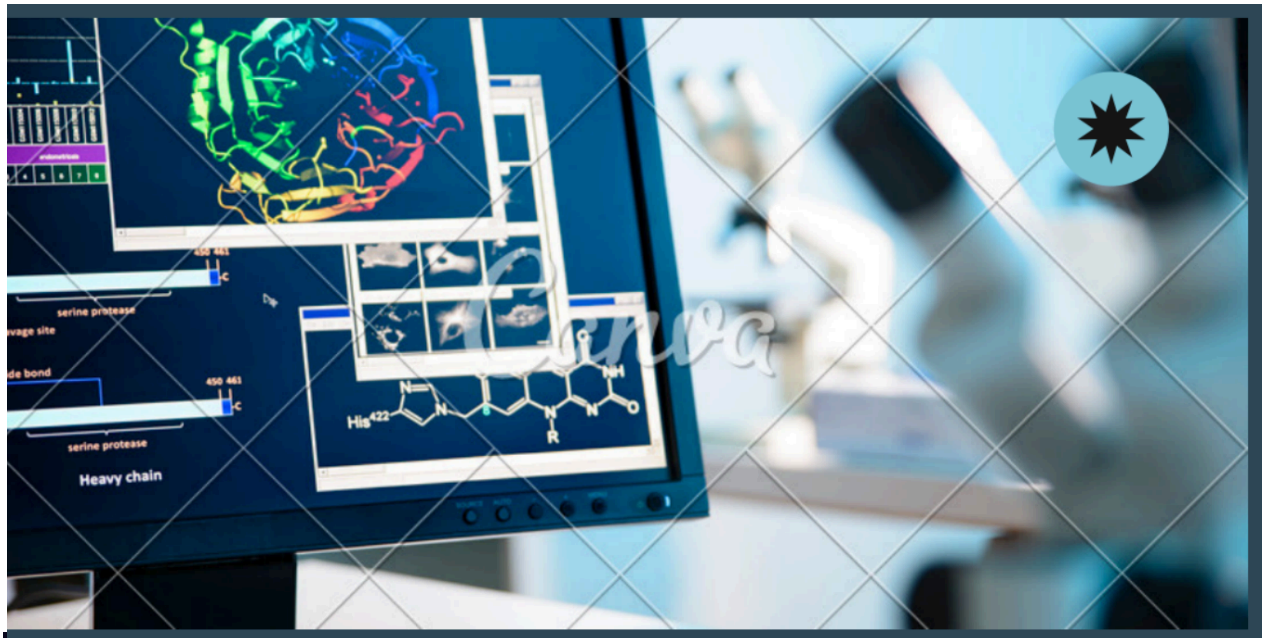
Developing Machine Learning Models For Predicting Fold-Rates of Proteins and Peptides

RCSB PDB
PROTEIN DATA BANK



Reegina Tyagi

INTRODUCTION

One basic biological mechanism that is essential to understanding the structure and function of proteins is protein folding. Practically every cellular process, including enzymatic activity, structural support, signaling, and immunological response, depends on proteins. A linear chain of amino acids is folded into a particular three-dimensional structure during the process of folding proteins, which is required for the protein to carry out its biological activity. For a variety of scientific and medical applications, such as medication design, the study of hereditary abnormalities, and the development of novel treatment approaches, it is essential to comprehend and anticipate the folding speeds of proteins and peptides.

A long-standing problem in molecular biology is predicting the speeds at which proteins fold. Although they offer detailed insights into protein structures, traditional experimental methods to study protein folding, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy, are frequently labor-intensive, time-consuming, and limited in their ability to capture the dynamic nature of folding processes. Moreover, systematic studies of a large number of proteins are challenging due to the enormous resources and expertise required for these methods.

The study of protein folding has changed dramatically since the development of computer techniques. Protein folding dynamics can be effectively analyzed and predicted by computational methods that utilize advanced algorithms and substantial datasets. Machine learning (ML) and deep learning (DL) techniques are two of these methods that have demonstrated significant promise in the extraction of intricate patterns and correlations from high-dimensional biological data. These methods could greatly advance our knowledge of protein folding mechanisms and increase the precision of folding rate estimates.

Protein sequences, structural characteristics, and molecular dynamics simulations are just a few examples of the data sources that can be used to train machine learning models, such as support vector machines, random forests, and gradient boosting regressors. Accurate predictions are made possible by these models' ability to capture complex correlations between folding rates and input characteristics. Additional benefits come from deep learning models, especially neural networks, which eliminate the need for intensive feature engineering by automatically learning feature representations from raw data.

The goal of this thesis is to create sophisticated deep learning and machine learning models to forecast the rates at which peptides and proteins fold. The research utilizes an all-encompassing computational approach that combines different bioinformatics instruments and methods to navigate the intricacies of protein folding dynamics and improve predictive powers.

Using Pfeature, a programme made to extract a variety of features from protein sequences, is one of the key components of this work. A crucial component of machine learning is feature engineering, which raises the caliber of input data and boosts model performance. Pfeature provides a rich set of descriptors for training models by generating features pertaining to the structure, physicochemical properties, and composition of amino acids.

Furthermore, the study represents protein structures as networks using Graph Signal Processing (GSP) approaches. This method makes it possible to investigate the relationships between residues that affect folding kinetics in greater detail. GSP approaches allow the extraction of low-frequency components of graph signals, which provide information about protein folding rates, by portraying protein structures as graphs.

The work heavily relies on molecular dynamics (MD) simulations, which describe the atomic movements within proteins under various situations. The popular MD simulation tool Amber23 makes it easier to comprehend the structural and energetic changes that take place during folding. These simulations add dynamic insights into protein behavior to the dataset and supply key parameters for accurate model training. By using MD simulations, scientists can watch the folding process in silico and get a close-up look at the intermediate states and energy landscapes that proteins travel through.

It is ensured that the produced models not only help to the theoretical understanding of protein biophysics but also accurately forecast folding rates through the integration of different data sources and analytical methodologies. This effort intends to bring new insights into one of the most complex biological processes and its applications in genetic and pharmaceutical design research by combining data science, machine learning, and computational biology.

Materials and Methodology

Data Collection

Protein Data: The Protein Folding Database (PFDB) and a number of academic articles provided the main dataset for this investigation. Folding rates for 141 single-domain globular proteins are available in the extensive PFDB database; 89 of these are categorized as two-state proteins and 52 as non-two-state proteins. To ensure uniformity throughout the dataset, the Eyring–Kramers equation was used to standardize the data in PFDB to a temperature of 25°C. An improved quality database was produced as a result of this change being confirmed by contrasting the estimated and empirically observed logarithmic rate constants for 14 distinct proteins at 25°C. For the purpose of developing and assessing theoretical and predictive protein folding studies, the PFDB acts as a standard.

Data on Amino Acid Properties: Comprehensive information on 48 physicochemical characteristics of the 20 standard amino acids was gathered, in addition to protein folding rates. These characteristics—among others, hydrophobicity, charge, and molecular mass—are essential for comprehending the dynamics of protein folding. The Pfeature programme was used to create the data on these characteristics, giving each amino acid in a protein sequence a complete feature set.

Computational Requirements

Software Tools

Computer programmes and applications used for certain tasks or purposes. This study made use of various bioinformatics tools and software:

- **Pfeature**: This software was utilized to extract features from protein sequences. The Pfeature tool provides a diverse set of features that greatly enhance the quality of input data for machine learning models. The tool offers descriptors pertaining to the amino acid content, physicochemical attributes, and structural traits.
- **Amber23**: This tool enables the execution of molecular dynamics (MD) simulations, which simulate the movement of atoms within proteins under different situations. Amber23 offers valuable and dynamic information about protein behavior, which is essential for comprehending the energetic and structural changes that take place during the folding process.

- **Visual Molecular Dynamics (VMD):** It is a software tool that was employed to visualize protein structures and the outcomes of molecular dynamics simulations. It aids in the interpretation of simulation results and comprehension of the structural alterations in proteins that occur during the folding process.

Data Preparation

Data Preprocessing

Data preparation is an essential step in guaranteeing the integrity and uniformity of the dataset. Various methodologies were utilized to preprocess the data:

- **Data standardization**: It was performed using techniques such as StandardScaler and MinMaxScaler. The StandardScaler function adjusts the data such that it has an average value of zero and a standard deviation of one. On the other hand, the MinMaxScaler function rescales the data so that it falls within the range of 0 to 1. The MinMaxScaler was determined to be the most efficient method for the 48 amino acid characteristics dataset.
- **Normalization**: Each feature was standardized to guarantee that all features have the same scale and contribute equally to the model training process. This phase is essential for algorithms that depend on distance measurements, such as k-nearest neighbors and support vector machines.

Feature Extraction and Selection

Feature extraction techniques were utilized to get significant descriptors from the protein data. The collected features were further analyzed using feature selection techniques to choose the most pertinent features for model training.

- **Pfeature Data:** Pfeature was utilized to create features pertaining to amino acid composition, physicochemical parameters, and structural traits. These characteristics offer a thorough depiction of the protein sequences.
- **Regarding Graph Signal Processing (GSP) data:** Protein structures were represented as networks using GSP techniques. This method enables the retrieval of graph signals' low-frequency components, which provide valuable information about protein folding processes. GSP approaches allow for a more comprehensive analysis of the interconnections between residues that impact the folding kinetics by portraying protein structures as graphs.
- **The Amber23 dataset:** utilized molecular dynamics (MD) simulations to extract energy-based characteristics. The simulations yielded comprehensive data on the dynamic properties of proteins, encompassing total energy, kinetic energy, potential energy, angles, bonds, dihedrals, van der Waals contacts, electrostatic interactions, solvation energy (GB), non-bonded interactions, and surface energy. The mean values

of these parameters were obtained from the production phase of the MD simulations. In addition, the radius of gyration was determined from the coordinate files using cpptraj to get an understanding of the compactness and folding status of the proteins. VMD was employed for calculating rmsd and saltbridges.

Model Development

Machine Learning Models

Multiple machine learning models were created and assessed to forecast protein folding rates:

- **Linear regression:** It is employed as a fundamental model to comprehend the linear correlation between characteristics and folding rates. Linear regression is useful for setting a baseline for evaluating the performance of more intricate models.
- **Support Vector Machines (SVM):** These are a type of machine learning algorithm that is used for classification and regression tasks. SVMs are particularly effective in cases where the data is not linearly separable, as they can map the data into a higher-dimensional space to find separating Submitted applications for both regression and classification assignments. Support Vector Machines (SVMs) are highly efficient at analyzing datasets with a large number of dimensions and are capable of handling complex non-linear relationships. The models underwent training using support vector regression (SVR) in order to forecast folding rates.
- **Random Forest:** It is an ensemble learning technique that utilizes several decision trees to enhance forecast accuracy. Random Forest models exhibit resilience against overfitting and have the ability to capture intricate relationships between features. They were especially effective in managing the wide range of characteristics produced by Pfeature.
- **Gradient Boosting Regressors:** These are an ensemble method that constructs models in a sequential manner, where each subsequent model aims to rectify the mistakes made by the preceding models. Gradient Boosting Regressors were utilized to improve the accuracy of predictions by specifically targeting occurrences that are challenging to anticipate.

Hyperparameter Tuning: The hyperparameters for both machine learning and deep learning models were optimized using grid search and random search approaches for hyperparameter tuning. The parameters, including the learning rate, batch size, and number of layers, were adjusted in order to optimize performance. Optimizing hyperparameters was crucial for maximizing the prediction capabilities of the models and avoiding overfitting.

Model Training and Validation

The models underwent training and validation using cross-validation techniques to guarantee their robustness and capacity to generalize.

- **K-Fold Cross-validation:** The dataset was partitioned into k subsets, with the model being trained on k-1 subsets and verified on the remaining subset. The method was iterated k times, and the outcomes were averaged to yield a full assessment. K-fold cross-validation aids in evaluating the model's performance on various subsets of the data, guaranteeing that the model can effectively generalize to unknown data.
- **Learning Curves**: Learning curves were generated to assess the model's performance across various training set sizes. This aided in the identification of problems related to overfitting and underfitting. Through the analysis of learning curves, the study was able to ascertain the ideal quantity of training data necessary and the behavior of the model as it acquired larger amounts of data.

Performance Metrics

In order to assess the effectiveness of the created models, multiple metrics were employed:

- **Mean Squared Error (MSE)**: The Mean Squared Error (MSE) is a metric that quantifies the average of the squared differences between the expected and actual values. Mean Squared Error (MSE) is a widely utilized metric in regression tasks, which quantifies the accuracy of the model.
- **R-Squared (R^2)**: R-Squared (R^2) is a statistical measure that quantifies the proportion of the variability in the dependent variable that can be explained by the independent variables. The coefficient of determination, R^2, serves as a measure of how well the model fits the data.
- **Root Mean Squared Error (RMSE)**: Base The Root Mean Squared Error (RMSE) is the square root of the average of the squared differences between the predicted values and the actual values. It quantifies the prediction error of the model in the same units as the target variable.
- **Mean Absolute Error (MAE)**: The Mean Absolute Error (MAE) is a metric that calculates the average absolute difference between the predicted and actual values. It provides a measure of the model's accuracy that is less affected by outliers compared to the Mean Squared Error (MSE).

The performance indicators were utilized to compare various models and choose the most effective ones for forecasting protein folding rates.

RESULT AND DISCUSSION

Sequence-Based Descriptor Analysis

The protein sequence-based descriptors exhibited substantial associations with folding rates. Different machine learning models, such as support vector machines, gradient boosting, and random forest regressors, were trained using these descriptors. The findings demonstrated that the models well captured the correlation between sequence-based characteristics and the speed at which folding occurs.

All Protein 48 Properties Dataset:

- **Support Vector Machines Regressor**: A support vector regression (SVR) model utilizing a radial basis function (RBF) kernel demonstrated enhanced performance, yielding a mean squared error **(MSE) of 2.89 and an R-squared (R^2) value of 0.39.** The utilization of a non-linear kernel facilitated the capture of intricate correlations within the data.
- **Random Forest Regressor**: The random forest model has a mean squared error **(MSE) of 8.97 and a coefficient of determination (R^2) value of 0.34.** The ensemble method proved to be efficacious in managing the heterogeneous array of characteristics and capturing intricate interplays among them.

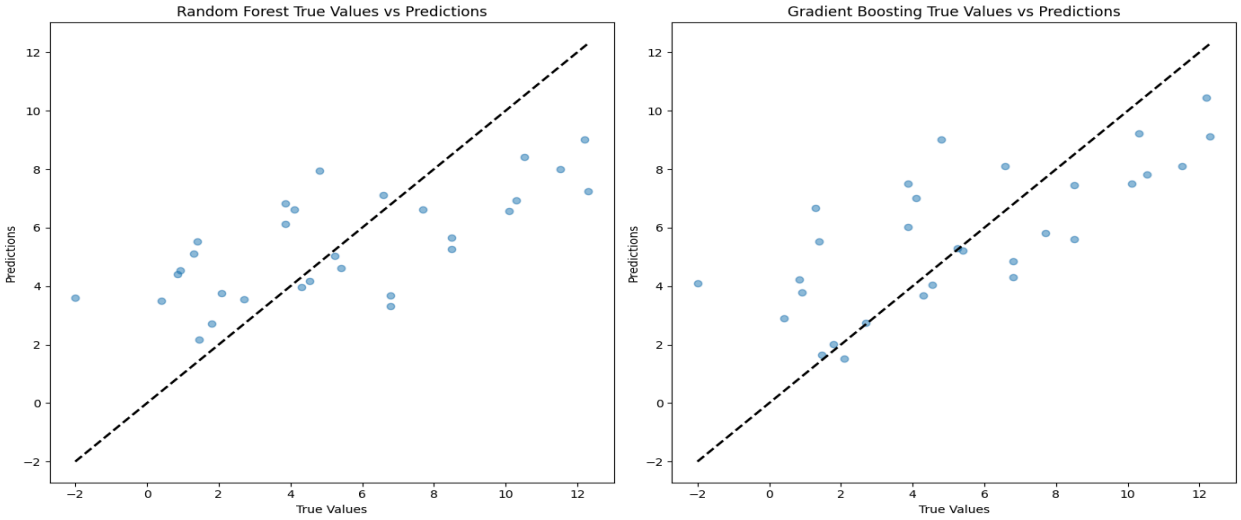
Dataset containing 48 properties of the 2s protein.

Multiple machine learning models, such as support vector machines and random forest regressors, were trained using these descriptors. The results demonstrated that the models well captured the correlation between sequence-based characteristics and folding speeds.

- **Support Vector Machines Regressor**: The support vector regression (SVR) model, with a radial basis function (RBF) kernel, demonstrated enhanced performance by reaching a mean squared error **(MSE) of 3.09 and a coefficient of determination (R^2) value of 0.41.** The utilization of a non-linear kernel facilitated the capture of intricate correlations within the data.
- **Random Forest Regressor**: The support vector regression (SVR) model, with a radial basis function (RBF) kernel, demonstrated enhanced performance by reaching a mean squared error **(MSE) of 3.09 and a coefficient of determination (R^2) value of 0.41.** The implementation of a non-linear kernel facilitated the capture of intricate linkages within the dataset.

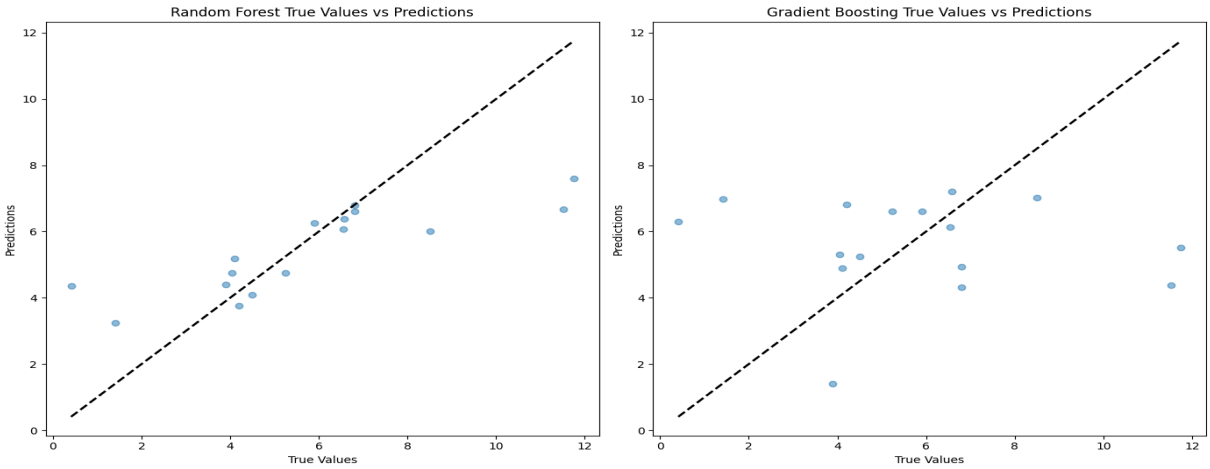
PFeature composition dataset for 2s proteins:

- **Gradient Boosting Regressors**: Additionally, it exhibited strong performance, achieving a mean squared error **(MSE) of 7.33 and an R-squared (R^2) value of 0.50.** This model successfully captured situations that were challenging to predict.
- **Random Forest Regressor**: The random forest model demonstrated superior performance compared to the linear regression and SVM models, with a mean squared error **(MSE) of 8.417 and a R^2 value of 0.428**. The ensemble method proved to be efficacious in managing the heterogeneous array of characteristics and capturing intricate interplays among them.



PFeature composition dataset for 2s proteins with single domain:

- **Gradient Boosting Regressors**: Additionally, it exhibited strong performance, achieving a Mean Squared Error **(MSE) of 9.01 and an R-squared (R^2) value of 0.37.** This model successfully captured situations that were challenging to prediction.
- **Random Forest Regressor**: The random forest model demonstrated superior performance compared to the linear regression and SVM models, achieving a mean squared error **(MSE) of 4.30 and an R-squared (R^2) value of 0.51.** The ensemble technique proved to be efficacious in managing the heterogeneous array of characteristics and capturing intricate interconnections among them.

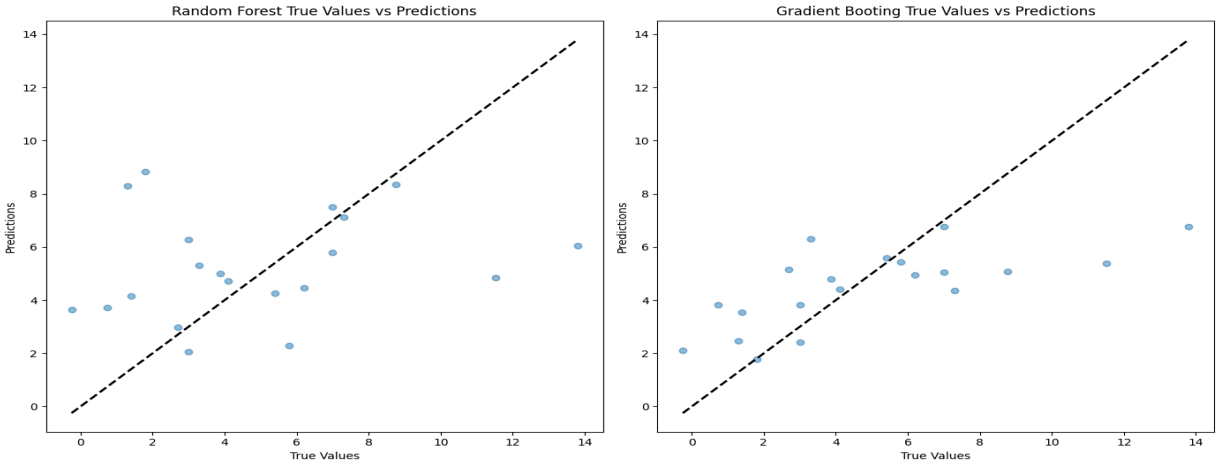


GSP-Based Descriptor Analysis

Protein structures were generated as networks using Graph Signal Processing (GSP) techniques. The revised Residue Interaction Graph (RIG) model was employed to capture essential long-range interactions involved in protein folding.

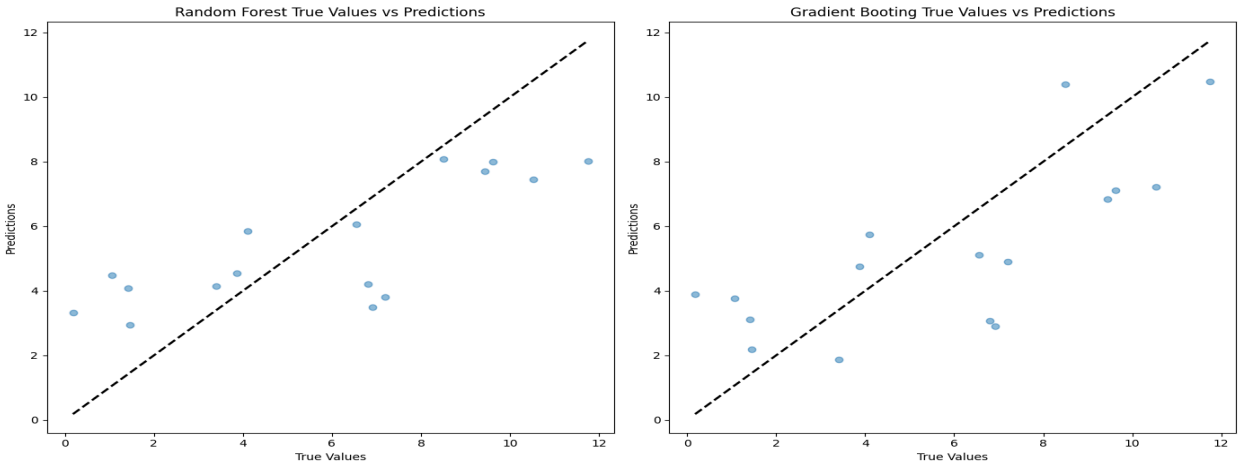
2s proteins with both single and multidomain GSP dataset:

- **Random Forest**: The descriptors based on the Generalized Sequential Pattern (GSP) were examined using several machine learning techniques. The random forest model obtained a mean squared error **(MSE) of 1.995 and a R^2 value of 0.53**, indicating a substantial enhancement compared to the sequence descriptors.
- **Gradient Boosting Regressors**: Additionally, it exhibited strong performance, achieving a Mean Squared Error **(MSE) of 2.769 and an R-squared (R^2) value of 0.387**. This model successfully captured situations that were challenging to predict.



2s proteins with both single domain GSP dataset:

- **Random Forest**: The descriptors based on the Generalized Sequential Pattern (GSP) were examined using several machine learning techniques. The random forest model obtained a mean squared error **(MSE) of 5.908 and a R^2 value of 0.53**, indicating a substantial enhancement compared to the sequence descriptors.
- **Gradient Boosting Regressors**: demonstrated strong performance, achieving a Mean Squared Error **(MSE) of 6.06 and an R-squared (R^2) value of 0.52.** This model successfully captured situations that were challenging to forecast.



The findings indicate that by representing protein structures as networks and examining the interactions between residues, the prediction capability of machine learning models can be greatly improved.

Energy-based Descriptor Analysis

Energy-based descriptors were extracted using Amber23 in molecular dynamics simulations. The descriptors encompassed total energy, kinetic energy, potential energy, and different interaction energies such as van der Waals and electrostatic energies. The energy-based descriptors were determined to be essential for precise prediction of folding rate.

- **Molecular Dynamics (MD) Simulations**: The molecular dynamics simulations yielded comprehensive understanding of the energetic and structural changes that take place during the process of protein folding. The mean values of total energy, kinetic energy,

potential energy, and other interaction energies were collected and utilized as features for model training.

- **Integration with Machine Learning Models**: The random forest model, trained using energy-based descriptors, achieved a mean squared error **(MSE) of 7.20 and a R^2 value of 0.47.** These results indicate that it had the highest level of predictive accuracy compared to all other examined models. The incorporation of energy-based descriptors greatly enhanced the model's capacity to reliably forecast folding rates.

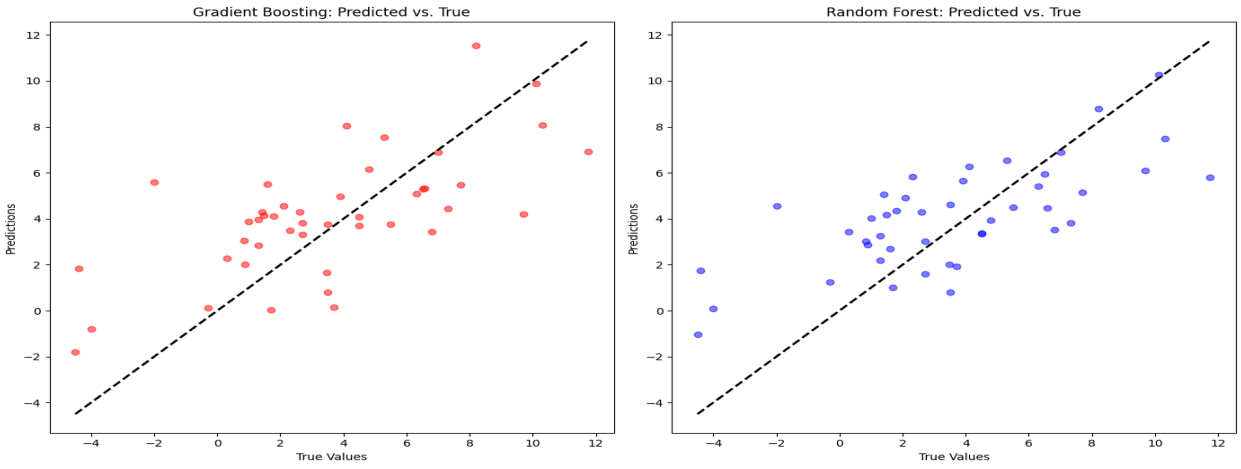
The findings from the examination of energy-based descriptors highlight the significance of dynamic understanding of protein function in order to make precise predictions about folding rates. Combining molecular dynamics simulations with machine learning models offers a robust method for comprehending the mechanics of protein folding.

Model Performance Comparison

A comparison was made between several models using various metrics, such as mean squared error (MSE), R-squared (R^2), and root mean squared error (RMSE). The findings demonstrated that ensemble models, such as random forest and gradient boosting regressors, exhibited superior performance. Deep learning models, specifically neural networks, have demonstrated encouraging outcomes but necessitated meticulous optimization to prevent overfitting.

All proteins energy Data:

- **Random Forest**: demonstrated superior performance when trained on energy-based descriptors, achieving an **MSE (Mean Squared Error) of 7.20 and a R^2 (R-squared) value of 0.47.**
- **Gradient Boosting Regressors**: demonstrated strong performance, achieving a Mean Squared Error **(MSE) of 7.78 and a R^2 value of 0.43.** This model demonstrated efficacy in catching occurrences that were challenging to forecast.



The evaluation of model performance underscores the efficacy of ensemble approaches in predicting protein folding rates. The findings indicate that the utilization of many data sources and the application of sophisticated modeling approaches can greatly improve the accuracy of predictions.