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

RCSB PDB
PROTEIN DATA BANK

Reegina Tyagi

INTRODUCTION

Protein folding is a fundamental biological process that plays a crucial role in determining the structure and function of proteins. Proteins are essential molecules involved in virtually all cellular functions, including enzymatic activity, structural support, signaling, and immune response. The process of protein folding involves the transformation of a linear chain of amino acids into a specific three-dimensional structure, which is necessary for the protein to perform its biological function. Understanding and predicting the folding rates of proteins and peptides is vital for various scientific and medical applications, including drug design, understanding genetic disorders, and developing new therapeutic strategies.

Predicting protein folding rates has been a longstanding challenge in the field of molecular biology. Traditional experimental methods to study protein folding, such as X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy, provide detailed insights into protein structures but are often time-consuming, labor-intensive, and limited in their ability to capture the dynamic nature of folding processes. Moreover, these methods require significant resources and expertise, making it difficult to study a large number of proteins systematically.

The advent of computational methods has revolutionized the study of protein folding. Computational approaches offer a powerful and efficient means to analyze and predict protein folding dynamics by leveraging large datasets and sophisticated algorithms. Among these approaches, machine learning (ML) and deep learning (DL) techniques have shown great promise in extracting complex patterns and relationships from high-dimensional biological data. These techniques have the potential to significantly enhance our understanding of protein folding mechanisms and improve the accuracy of folding rate predictions.

Machine learning models, such as support vector machines, random forests, and gradient boosting regressors, can be trained on various data sources, including protein sequences, structural features, and molecular dynamics simulations. These models can capture intricate relationships between the input features and folding rates, enabling accurate predictions. Deep learning models, particularly neural networks, offer additional advantages by automatically learning feature representations from raw data, reducing the need for extensive feature engineering.

This thesis aims to develop advanced machine learning and deep learning models to predict the folding rates of proteins and peptides. The study employs a comprehensive computational methodology that integrates various bioinformatics tools and techniques to navigate the complexities of protein folding dynamics and enhance predictive capabilities.

One of the fundamental steps in this study is the use of Pfeature, a program designed to extract a wide range of features from protein sequences. Feature engineering is a critical aspect of machine learning, as it significantly improves the quality of input data and enhances model performance. Pfeature generates features related to amino acid composition, physicochemical properties, and structural characteristics, providing a rich set of descriptors for model training.

Additionally, the study employs Graph Signal Processing (GSP) techniques to represent protein structures as networks. This approach allows for a more in-depth examination of the connections between residues that influence folding kinetics. By representing protein structures as graphs, GSP techniques enable the extraction of low-frequency components of graph signals, which are informative for protein folding rates.

Molecular dynamics (MD) simulations play a crucial role in this study by modeling the atomic movements within proteins under varied conditions. Amber23, a widely used tool for MD simulations, facilitates the understanding of energetic and structural alterations that occur during the folding process. These simulations enrich the dataset with dynamic insights into protein behavior, providing crucial parameters for precise model training. MD simulations allow researchers to observe the folding process in silico, offering a detailed view of the intermediate states and energy landscapes that proteins navigate during folding.

The integration of multiple data sources and analytical methods ensures that the developed models not only accurately predict folding rates but also contribute to the theoretical understanding of protein biophysics. By combining data science, machine learning, and computational biology, this work aims to provide new insights into one of the most intricate biological processes and its applications in genetic and medication design research.

Materials and Methodology

Data Collection

Protein Data : The primary dataset for this study was sourced from the Protein Folding Database (PFDB) and various research publications. PFDB is a comprehensive resource that includes folding rates for 141 single-domain globular proteins, with 89 classified as two-state proteins and 52 as non-two-state proteins. The data in PFDB were standardized to a temperature of 25°C using the Eyring–Kramers equation, ensuring consistency across the dataset. This adjustment was verified by comparing the estimated and empirically observed logarithmic rate constants for 14 different proteins at 25°C, resulting in an enhanced quality database. The PFDB serves as a benchmark for the creation and evaluation of theoretical and predictive protein folding research.

Amino Acid Properties Data : In addition to protein folding rates, detailed data regarding 48 physicochemical properties of the 20 standard amino acids were collected. These properties include molecular mass, hydrophobicity, charge, and others, which are crucial for understanding protein folding dynamics. Data on these properties were generated using the Pfeature software, providing a comprehensive feature set for each amino acid in a protein sequence.

Computational Requirements

Software Tools

Several bioinformatics tools and software were utilized in this study:

- **Pfeature**: This software was used for feature extraction from protein sequences. Pfeature generates a wide range of features that significantly improve the input data quality for machine learning models. It provides descriptors related to amino acid composition, physicochemical properties, and structural characteristics.
- **Amber23**: This tool facilitates molecular dynamics (MD) simulations, modeling atomic movements within proteins under various conditions. Amber23 provides dynamic insights into protein behavior, which are crucial for understanding the energetic and structural alterations occurring during the folding process.
- **Visual Molecular Dynamics (VMD)**: VMD was used for visualizing protein structures and the results of molecular dynamics simulations. It helps in interpreting the simulation data and understanding the structural changes in proteins during folding.

Data Preparation

Data Preprocessing

Data preprocessing is a critical step to ensure the quality and consistency of the dataset. Several techniques were employed to preprocess the data:

- **Standardization**: Techniques such as StandardScaler and MinMaxScaler were applied to the data. StandardScaler transforms the data to have a mean of zero and a standard deviation of one, while MinMaxScaler scales the data to a range between 0 and 1. MinMaxScaler was found to be the most effective for the 48 amino acid properties dataset.
- **Normalization**: Each feature was normalized to ensure that all features contribute equally to the model training process. This step is crucial for algorithms that rely on distance measurements, such as k-nearest neighbors and support vector machines.

Feature Extraction and Selection

Feature extraction methods were employed to generate meaningful descriptors from the protein data. The extracted features were then subjected to feature selection techniques to identify the most relevant features for model training.

- **For Pfeature Data**: Pfeature was used to generate features related to amino acid composition, physicochemical properties, and structural characteristics. These features provide a comprehensive representation of the protein sequences.
- **For Graph Signal Processing (GSP) Data**: GSP techniques were used to represent protein structures as networks. This approach allows for the extraction of low-frequency components of graph signals, which are informative for protein folding rates. By representing protein structures as graphs, GSP techniques enable a more in-depth examination of the connections between residues that influence folding kinetics.
- **Amber23 Dataset**: Molecular dynamics (MD) simulations were performed using Amber23 to extract energy-based features. These simulations provided detailed information about the dynamic behavior of proteins, including total energy, kinetic energy, potential energy, angles, bonds, dihedrals, van der Waals interactions, electrostatic interactions, solvation energy (GB), non-bonded interactions, and surface energy. The average values of these parameters were extracted from the production phase of the MD simulations. Additionally, the radius of gyration was calculated from the coordinate files using cpptraj to provide insights into the compactness and folding state of the proteins and VMD was used for rmsd and saltbridges.

Model Development

Machine Learning Models

Several machine learning models were developed and evaluated for predicting protein folding rates:

- **Linear Regression**: Used as a baseline model to understand the linear relationship between features and folding rates. Linear regression helps in establishing a reference point for comparing the performance of more complex models.
- **Support Vector Machines (SVM)**: Applied for both regression and classification tasks. SVMs are effective for high-dimensional datasets and can handle non-linear relationships. The models were trained using support vector regression (SVR) to predict folding rates.
- **Random Forest**: An ensemble learning method that uses multiple decision trees to improve predictive performance. Random Forest models are robust to overfitting and can capture complex interactions between features. They were particularly useful in handling the diverse set of features generated by Pfeature.
- **Gradient Boosting Regressors**: Another ensemble method that builds models sequentially, with each new model correcting the errors of the previous ones. Gradient Boosting Regressors were employed to enhance predictive accuracy by focusing on difficult-to-predict instances.

Hyperparameter Tuning: Grid search and random search methods were used to optimize hyperparameters for both machine learning and deep learning models. Parameters such as learning rate, batch size, and the number of layers were tuned to achieve the best performance. Hyperparameter tuning was critical in maximizing the models' predictive capabilities and preventing overfitting.

Model Training and Validation

The models were trained and validated using cross-validation techniques to ensure robustness and generalizability:

- **K-Fold Cross-Validation**: The dataset was divided into k subsets, and the model was trained on k-1 subsets while being validated on the remaining subset. This process was repeated k times, and the results were averaged to provide a comprehensive evaluation. K-fold cross-validation helps in assessing the model's performance across different subsets of the data, ensuring that the model generalizes well to unseen data.
- **Learning Curves**: Learning curves were plotted to analyze the model's performance over different training set sizes. This helped identify overfitting and underfitting issues. By

examining the learning curves, the study could determine the optimal amount of training data required and the model's behavior as it learned from increasing amounts of data.

Performance Metrics

To evaluate the performance of the developed models, several metrics were used:

- **Mean Squared Error (MSE)**: Measures the average squared difference between the predicted and actual values. MSE is a commonly used metric for regression tasks, providing a measure of the model's accuracy.
- **R-Squared (R^2)**: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. R^2 provides an indication of the goodness of fit of the model.
- **Root Mean Squared Error (RMSE)**: The square root of the mean squared error, providing a measure of the model's prediction error in the same units as the target variable.
- **Mean Absolute Error (MAE)**: Measures the average absolute difference between the predicted and actual values, providing a measure of the model's accuracy that is less sensitive to outliers than MSE.

These performance metrics were used to compare different models and select the best-performing ones for predicting protein folding rates.

Sequence-Based Descriptor Analysis

The Sequence based descriptors derived from protein sequences showed significant correlations with folding rates.Various machine learning models, including support vector machines, gradient boosting and random forest regressors, were trained on these descriptors. The results indicated that the models were able to capture the relationship between sequence-based features and folding rates effectively.

All Protein 48 Properties Dataset:

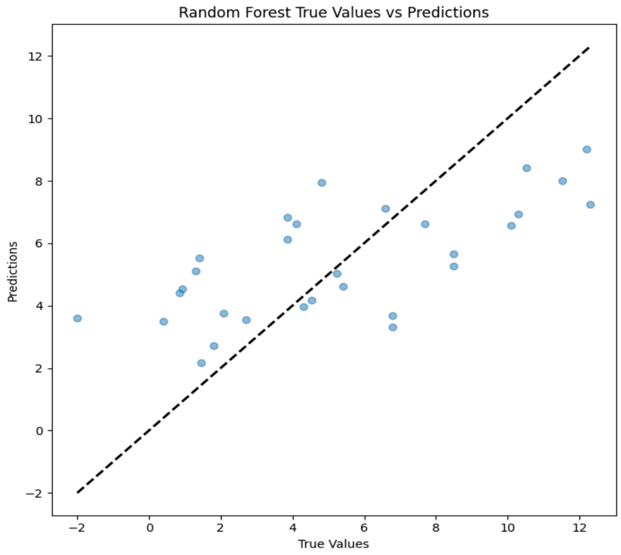
- **Support Vector Machines Regressor**: The SVR model with a radial basis function (RBF) kernel improved the performance, achieving an **MSE of 2.89 and an R^2 value of 0.39**. The non-linear kernel helped in capturing more complex relationships in the data.
- **Random Forest Regressor**: The random forest model with an **MSE of 8.97 and an R^2 value of 0.34**. The ensemble method was effective in handling the diverse set of features and capturing complex interactions between them.

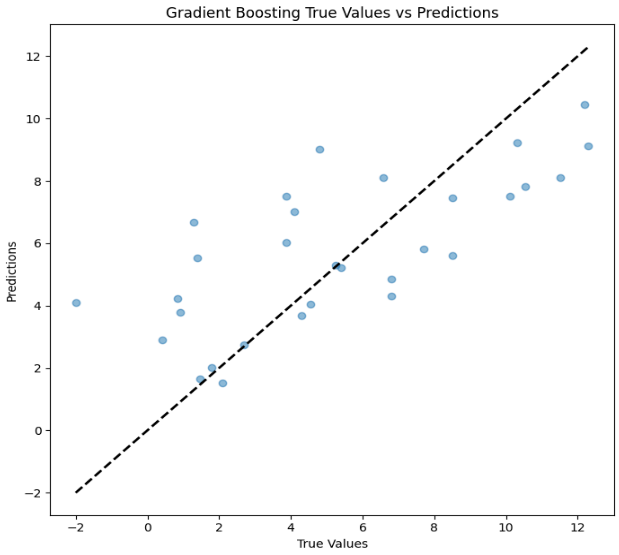
All 2s Protein 48 Properties Dataset: Various machine learning models, including support vector machines, and random forest regressors, were trained on these descriptors. The results indicated that the models were able to capture the relationship between sequence-based features and folding rates effectively.

- **Support Vector Machines Regressor**: The SVR model with a radial basis function (RBF) kernel improved the performance, achieving an **MSE of 3.09 and an R^2 value of 0.41**. The non-linear kernel helped in capturing more complex relationships in the data.
- **Random Forest Regressor**: The random forest model with an **MSE of 9.58 and an R^2 value of 0.407**. The ensemble method was effective in handling the diverse set of features and capturing complex interactions between them.

PFeature composition dataset for 2s proteins:

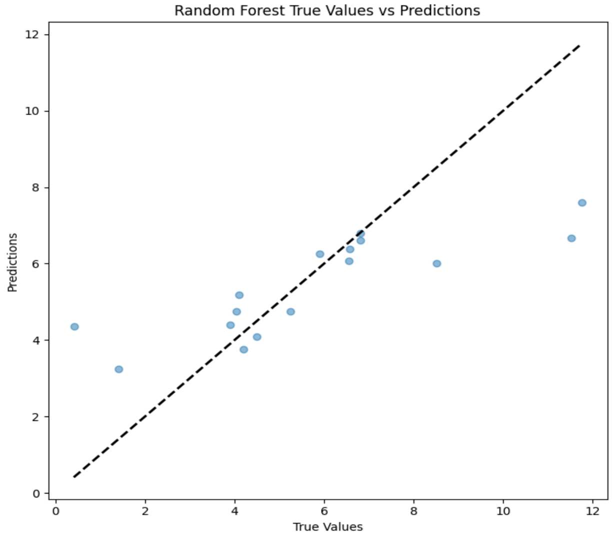
- **Gradient Boosting Regressors**: Also performed well, with an **MSE of 7.33 and an R^2 value of 0.50**. This model was effective in capturing difficult-to-predict instances.
- **Random Forest Regressor**: The random forest model outperformed the linear regression and SVM models, with an **MSE of 8.417 and an R^2 value of 0.428**. The ensemble method was effective in handling the diverse set of features and capturing complex interactions between them.

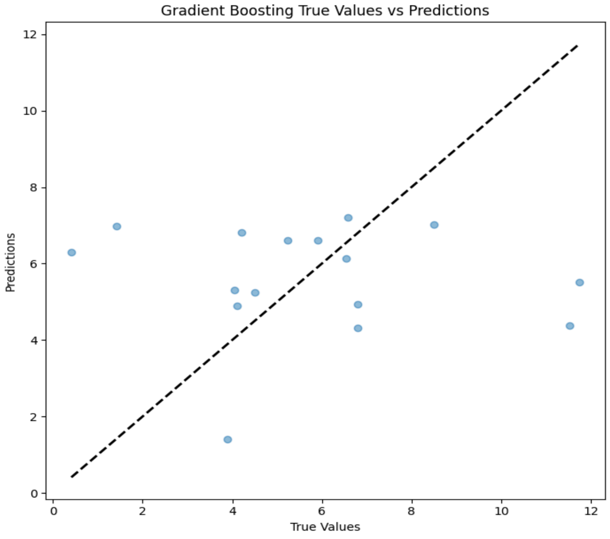




PFeature composition dataset for 2s proteins with single domain:

- **Gradient Boosting Regressors**: Also performed well, with an **MSE of 9.01 and an R² value of 0.37**. This model was effective in capturing difficult-to-predict instances.
- **Random Forest Regressor**: The random forest model outperformed the linear regression and SVM models, with an **MSE of 4.30 and an R² value of 0.51**. The ensemble method was effective in handling the diverse set of features and capturing complex interactions between them.



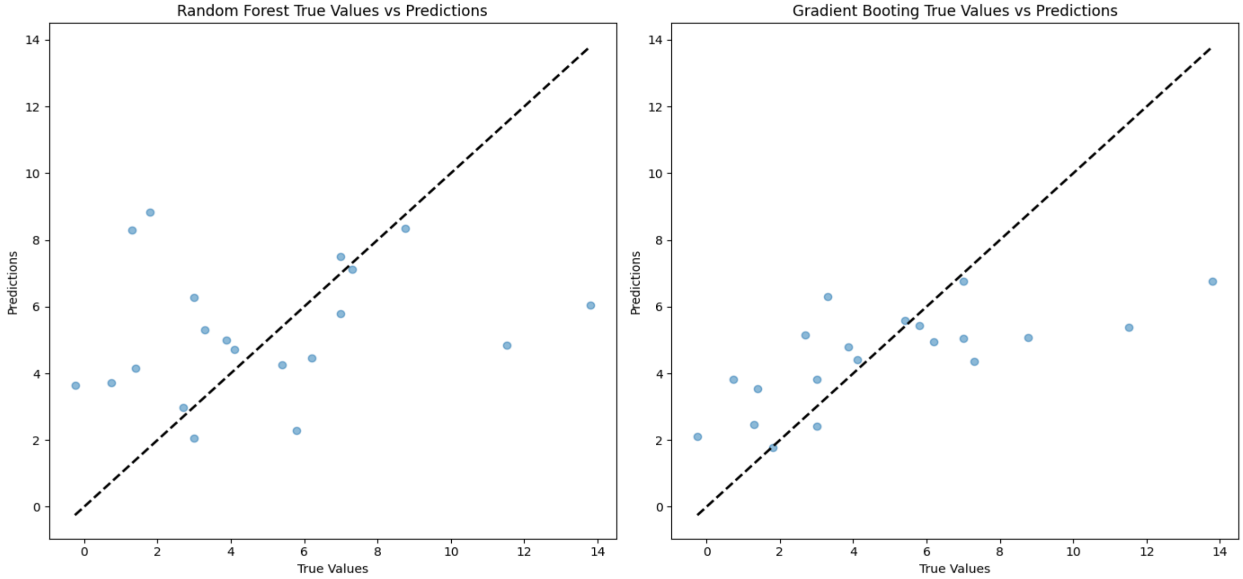


GSP-Based Descriptor Analysis

Graph Signal Processing (GSP) techniques were employed to generate protein structures as networks. The modified Residue Interaction Graph (RIG) model was used to capture long-range interactions crucial for protein folding.

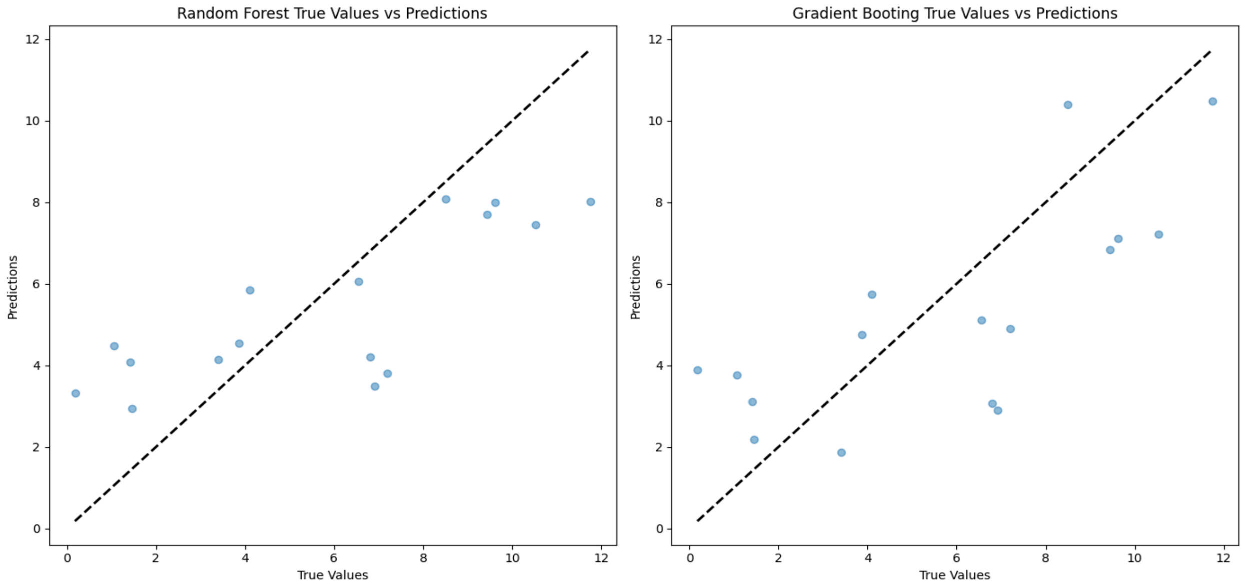
2s proteins with both single and multidomain GSP dataset:

- **Random Forest**: The GSP-based descriptors were analyzed using various machine learning models. The random forest model achieved an **MSE of 1.995 and an R² value of 0.53**, showing a significant improvement over the sequence descriptors.
- **Gradient Boosting Regressors**: Also performed well, with an **MSE of 2.769 and an R² value of 0.387**. This model was effective in capturing difficult-to-predict instances.



2s proteins with both single domain GSP dataset:

- **Random Forest**: The GSP-based descriptors were analyzed using various machine learning models. The random forest model achieved an **MSE of 5.908 and an R² value of 0.53**, showing a significant improvement over the sequence descriptors.
- **Gradient Boosting Regressors**: Also performed well, with an **MSE of 6.06 and an R² value of 0.52**. . This model was effective in capturing difficult-to-predict instances.



The results suggest that representing protein structures as networks and analyzing the interactions between residues can significantly enhance the predictive power of machine learning models.

Energy-based Descriptor Analysis

Molecular dynamics simulations using Amber23 were employed to extract energy-based descriptors. These descriptors included total energy, kinetic energy, potential energy, and various interaction energies (e.g., van der Waals, electrostatic). The energy-based descriptors were found to be crucial for accurate folding rate prediction.

- **Molecular Dynamics (MD) Simulations**: The MD simulations provided detailed insights into the energetic and structural alterations occurring during protein folding. The average values of total energy, kinetic energy, potential energy, and other interaction energies were extracted and used as features for model training.
- **Integration with Machine Learning Models**: The random forest model trained on energy-based descriptors achieved an **MSE of 7.20 and an R² value of 0.47**, indicating the highest predictive accuracy among all the models tested. The inclusion of energy-based descriptors significantly improved the model's ability to predict folding rates accurately.

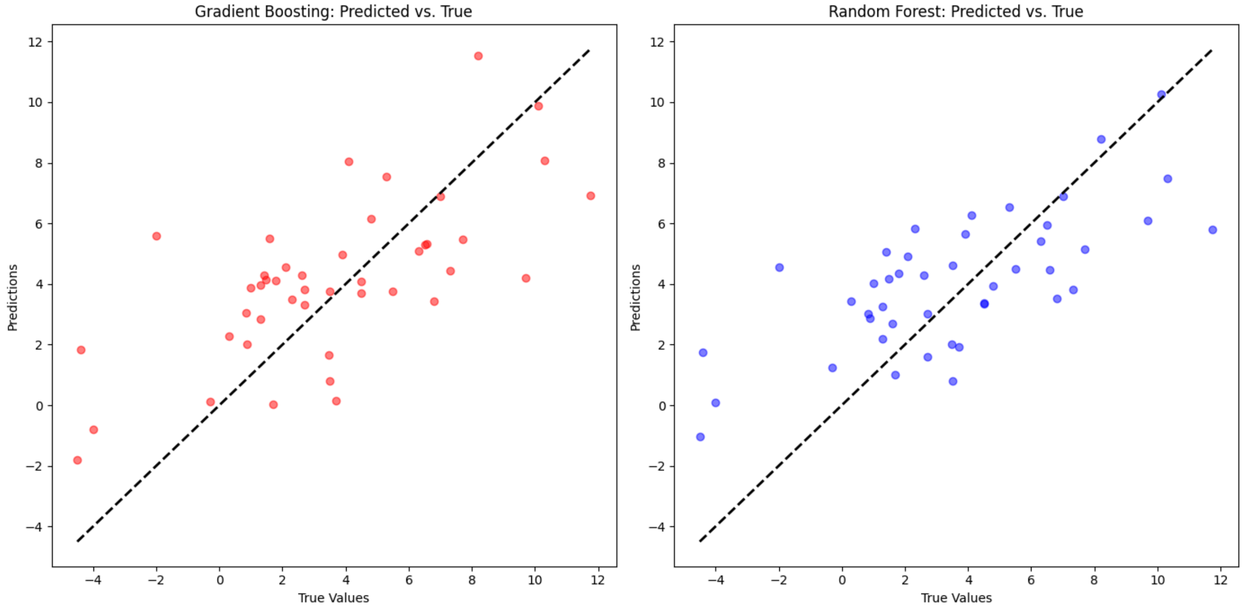
The results from the energy-based descriptor analysis underscore the importance of dynamic insights into protein behavior for accurate folding rate prediction. The integration of molecular dynamics simulations with machine learning models provides a powerful approach to understanding protein folding mechanisms.

Model Performance Comparison

The performance of different models was compared based on various metrics, including mean squared error (MSE), R-squared (R^2), and root mean squared error (RMSE). The results indicated that ensemble models, such as random forest and gradient boosting regressors, provided the best performance. Deep learning models, particularly neural networks, also showed promising results but required careful tuning to avoid overfitting.

All proteins energy Data:

- **Random Forest**: Achieved the best overall performance with an **MSE of 7.20 and an R^2 value of 0.47** when trained on energy-based descriptors.
- **Gradient Boosting Regressors**: Also performed well, with an **MSE of 7.78 and an R^2 value of 0.43**. This model was effective in capturing difficult-to-predict instances.



The comparison of model performance highlights the effectiveness of ensemble methods for predicting protein folding rates. The results suggest that combining multiple data sources and leveraging advanced modeling techniques can significantly enhance predictive accuracy.