Journal of Inventive and Scientific Research Studies (JISRS)
www.jisrs.com Vol: II, Issue: 1 June 2024 ISSN: 2584-0630 (Online)

ANALYSIS OF VARIOUS DEEP LEARNING TECHNIQUES FOR PREDICTING PROTEIN STABILITY

A. Jasmine Sugil[1*], K. Merriliance[2], Mary Immaculate Sheela Lourdusamy[3]

[1]Research Scholar, Reg.No:23111242282011, Department of Computer Applications & Research Centre, Sarah Tucker College (Autonomous), Affiliated to Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli -627007, Tamilnadu, India.
E-mail: findsugil@gmail.com

[2]Associate Professor, Department of Computer Applications &Research Centre, Sarah Tucker College (Autonomous), Tirunelveli, -627007, Tamilnadu, India.
E-mail: merriliance@gmail.com

[3]Professor, Department of Informatics, Heritage Christian University College, Accra, Ghana. E-mail: drsheela09@gmail.com

Received: April 08, 2024, **Accepted**: May 03, 2024, **Online Published**: June 15, 2024

ABSTRACT

In biological systems, the structure, function, and behavior of proteins are all influenced by a fundamental factor known as protein stability. It is essential to predict changes in protein stability resulting from amino acid substitutions to comprehend the evolution of proteins, develop protein therapies, and comprehend the mechanisms underlying diseases. In this paper, we present a unique deep-learning sequence-based method to predict changes in protein stability after amino acid alterations accurately. The technique uses deep neural network architecture to understand protein sequence and structure using a dataset of experimentally confirmed stability changes. It incorporates structural data, evolutionary conservation scores, and physicochemical characteristics for improved prediction accuracy. The proposed deep learning technique captures local and global sequence features, integrates attention mechanisms, and demonstrates robustness and generalization across diverse protein families and mutation scenarios. The technique also deals with predicting protein stability changes due to amino acid mutations using

convolutional and recurrent neural network layers and attention mechanisms. This model effectively captures complex sequence-structure relationships; through comprehensive evaluations, the method demonstrates superior performance compared to existing techniques, offering valuable insights into the impact of mutations on protein stability and facilitating advanced protein engineering and drug discovery efforts. This approach outperforms traditional methods in predictive accuracy and efficiency, outperforming sequence-based methods and other machine-learning approaches and providing valuable insights into protein stability. It also provides an effective tool for academics and practitioners in protein engineering, drug discovery, and structural biology, constitutes a substantial development in the field of protein stability prediction overall, and predicts amino acid insertion, deletion, or substitutions' effects on protein stability with high accuracy, which helps with rational protein design and advances our knowledge of the interactions between protein structure and function in biological systems. At the outset, when compared with other methods, with respect to total, direct, reverse, and anti-symmetry metrics, INPS, ACDC-NN, and DDGun exhibit improved performance.

Keywords: Protein Stability, Deep Learning, Sequence-Based Method, Protein Engineering, and Amino Acid Substitutions.

1. Introduction

Proteins are essential biomolecules in living organisms and consist of a chain of amino acids. Their structure can be directly sequenced or inferred from the DNA sequence (Figure 1). Protein sequencing determines the amino acid sequence, facilitating in understanding, identifying, and categorizing post-translational modifications. Understanding protein stability is crucial in various fields, such as drug development, protein engineering, and molecular biology [1-6]. Changes in protein stability can have significant implications for protein function and structure, influencing their behavior and interactions within biological systems [7-10]. Traditional experimental methods like thermal denaturation and NMR are time-consuming and labor-intensive. Computational approaches, such as deep learning, offer faster, more cost-effective, and more accurate predictions based on large-scale protein datasets. In recent years, computational approaches, particularly those leveraging deep learning techniques [11-14], have emerged as powerful tools for predicting protein stability changes with higher accuracy and efficiency.

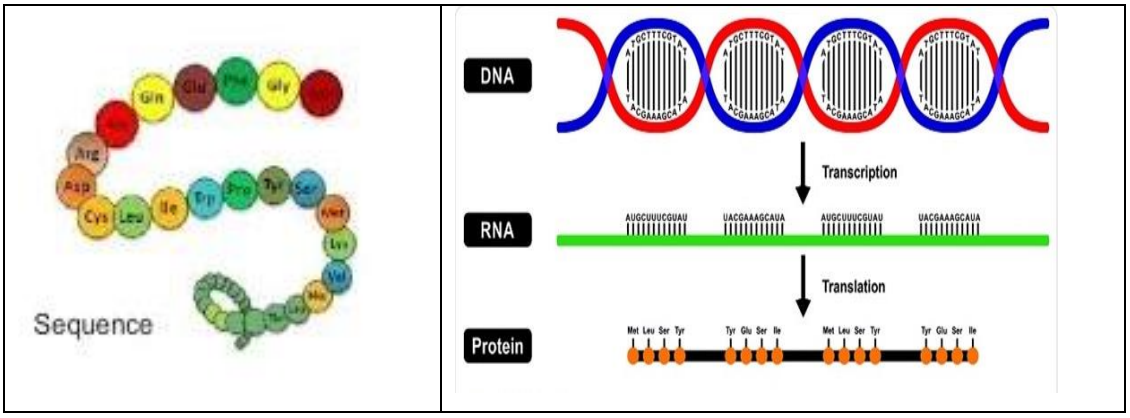


Figure 1: Protein Sequences [41]

Current prediction techniques characterize protein stability using one or more of the following features:

- **Structurally based features**: Protein shape, residue/atom distances, residue interaction networks, etc.
- **Sequence-based features**: Predicated on locations of amino acids and conserved sequences. They can offer an effect on the survivability of proteins, but they do not provide functional information.
- **Features based on energy**: Target protein unfolding energy as the total of different energies such as extra-stabilizing free energy, solvation energy, Van der Waals interactions, etc.

Molecular properties include the interface's solvent-accessible surface area as well as its hydrophobic and hydrophilic regions.

Many of these characteristics are based on protein structures. Sequence-based protein stability prediction relies on the amino acid sequence, while Structure-Based protein stability prediction analyzes the 3D structure.

Key components of this method include feature extraction from protein sequences, representation learning using deep neural networks, and training on large-scale protein stability datasets. By capturing intricate sequence-structure relationships, the deep learning model demonstrates superior performance compared to traditional methods, offering a promising avenue for accurate and rapid prediction of protein stability changes.

Moreover, we conduct an extensive analysis of the method, benchmarking it against existing state-of-the-art approaches on diverse datasets. Through comprehensive evaluation and validation, we showcase the reliability, scalability, and flexibility of the deep learning-based

approach across various protein families and mutation scenarios.

Furthermore, we provide insights into the underlying processes driving protein stability changes, highlighting the importance of specific amino acid residues and their interactions within protein structures [10]. Such insights not only enhance understanding of protein stability dynamics but also have practical implications in drug design, protein engineering, and disease therapeutics [3, 6].

This comprehensive review provides insights into various deep-learning methods for predicting protein stability changes.

1.1 Significance of Protein Stability Prediction

Protein stability prediction is a crucial aspect of biological research and biotechnology, influencing various fields. It helps identify molecules that can interact with target proteins to modulate their activity, optimize protein function, and enhance protein expression levels. In drug discovery, understanding the stability of target proteins and how they may be affected by small molecule ligands or protein-protein interactions is essential for designing effective therapeutics [15]. Protein engineering efforts aim to modify existing proteins or design novel ones with desired properties, optimizing protein function and resistance to environmental conditions. In biomedical research, predicting protein stability changes helps identify disease-causing variants and guide research towards developing targeted therapies. In structural biology, understanding the stability of protein structures and complexes provides insights into protein folding mechanisms (Figure 2), protein-protein interactions, and macromolecular assembly processes. In biotechnology and industrial applications, predicting protein stability is critical for optimizing protein production processes, enzyme stability, and the development of biocatalysts. In personalized medicine [9], predicting the stability effects of genetic variants on proteins is essential for understanding individual disease risks and tailoring treatment strategies. High-throughput screening of large numbers of mutations or protein variants enables the discovery and optimization of proteins with desired properties.

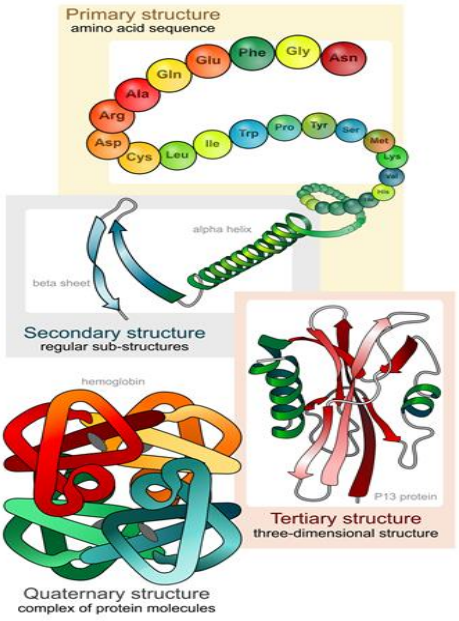


Figure 2: Protein folding [42]

Overall, this analysis contributes to advancing the field of protein stability prediction by introducing a robust deep-learning framework capable of accurately predicting stability changes across a wide range of protein sequences or structures and mutations. The implications of findings extend to various domains, including biotechnology, pharmaceuticals, and structural biology, where precise assessment of protein stability is paramount for advancing research and innovation.

2 Existing Methods and Limitations

Several existing methods have been developed to predict protein stability changes based on amino acid sequence and 3D structure profile, each with its strengths and limitations. Table 1 shows the commonly used methods along with their respective drawbacks:

Table 1: Existing Methods and Limitations

Method	Description	Limitations
Empirical Rules and Statistical Potentials [16]	Relies on empirical rules derived from statistical analyses to predict stability changes.	Limited accuracy and generalizability require experimental data for calibration, which may not capture complex relationships.
Machine Learning Models [17-19]	Utilizes techniques like support vector machines or random forests to predict stability changes based on features.	May struggle with complex relationships, require feature engineering, and may overlook important features.

Physics-Based Models [20-22]	Simulates protein folding and stability using molecular dynamics or statistical mechanics principles.	Computationally expensive, require detailed structural information, and inaccuracies in force fields may affect predictions.
Deep Learning Approaches [12-14,23-28]	Employs methods like convolutional neural networks or recurrent neural networks to learn hierarchical features from raw data.	Interpretability may be challenging because large amounts of training data and computational resources are required.
Hybrid and Ensemble Methods [15, 29-31]	Combines multiple prediction methods or integrates diverse features to improve accuracy.	Introduce complexity and computational overhead, requiring careful selection and optimization of component models.

The above indicates [**Table 1]** the different approaches and techniques used to predict stability changes. These methods have limitations, such as limited accuracy, and require experimental data for measurement. They may also struggle with complex relationships and require feature engineering. Deep learning approaches require large training data and computational resources.

2.1 Related protein stability analysis with Deep Learning frameworks

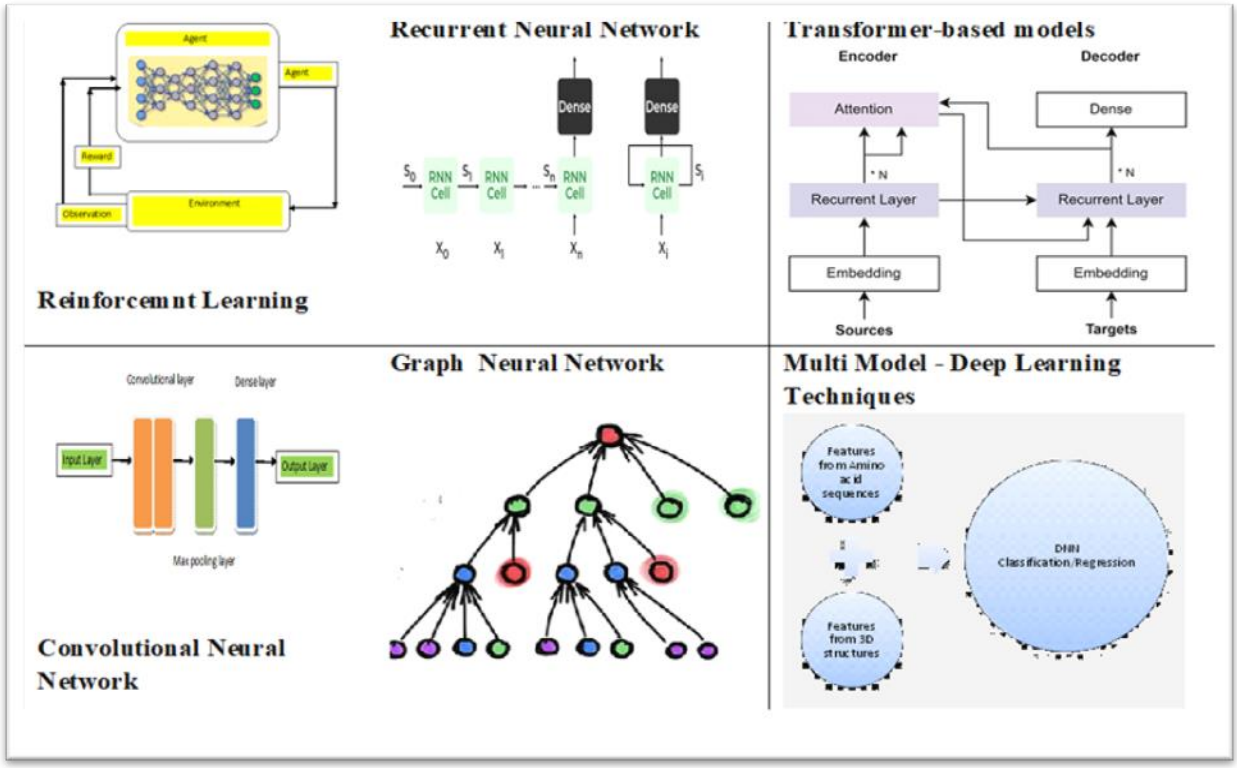


Figure 3: Deep Neural Network Models

Alley et al. [2] proposed a deep learning architecture that learned a high-dimensional representation of protein sequences directly from their primary structures. This representation captured complex relationships between amino acids and enabled accurate predictions of protein properties, including stability changes. Various deep learning architectures were discussed for protein stability analysis, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Neural Networks (GNNs), and Transformer-based models etc. (Figure 3).

Capriotti et al. [7] developed I-Mutant2.0, an SVM-based tool that predicted protein stability changes upon single-point mutations. It utilized either the protein structure or sequence and was trained on ProTherm's extensive thermodynamic experimental data. It could be used as a classifier and regression estimator for ΔΔG values. I-Mutant2.0 correctly predicted 80% or 77% of the dataset, depending on structural or sequence information. Its web interface allowed for predictive mode selection based on protein structure or sequence availability.

Jones et al. [14] examined to understand how these architectures leveraged protein sequence and structural data for stability prediction. Protein stability disruptions have been linked to disease, leading to the development of tools to predict free energy changes in protein residue variations. However, the limited number of protein structures and the lack of antisymmetric properties in current methodologies limit their application.

Savojardo et al. [43] introduce INPS-MD, a web server for predicting protein stability changes from single-point variations in sequence and structure, and INPS3D, a predictor using protein 3D structure features. Both demonstrate comparable performance to state-of-the-art methods.

Cheng et al. [19] used support vector machines to predict protein stability changes from single amino acid mutations, achieving 84% accuracy. The method, which considered only the sign of stability changes, was applicable to many situations where the tertiary structure was unknown, overcoming limitations of previous methods that required tertiary information.

Pancotti et al. [9] ACDC-NN-Seq proposed a deep neural network system utilizing sequence information and integrating the antisymmetric property. It was the first convolutional neural network To predict protein stability changes solely based on the protein sequence, and it

demonstrated favorable comparisons with existing sequence-based methods. This method predicted the ΔΔG value for single point variant only. Also suggested, this can be extended to predict the ΔΔG value for multiple site variants.

Numerous computational methods exist for determining protein stability, but there are still unsolved problems. These include inadequate databases for thermodynamic measurements, inherent variability in ΔΔG values due to experimental conditions, biased predictive methods that overlook anti-symmetry between native and mutant protein forms, and sequence similarity between training and test datasets, leading to overly optimistic prediction performance.

The current review aims to tackle these limitations and enhance the accuracy of protein stability predictions.

3. Materials and Methods

The methodology outlines the steps involved in dataset collection and pre-processing, deep learning architecture design ((Figure 4), training strategy implementation, and evaluation metrics selection for predicting protein stability changes using deep learning techniques.

3.1 Dataset Collection and Pre-processing:

Data Acquisition: Gather experimental data on protein stability changes resulting from amino acid substitutions from reputable databases such as ProTherm [35], FoldX [54], ThermoMutDB [55], or experimental literature.

Data Pre-processing: Clean the dataset by removing duplicates, inconsistencies, and errors. Standardize the representation of protein sequences and structures. Extract relevant features such as evolutionary conservation scores and physicochemical properties.

3.2 Deep Learning Architecture:

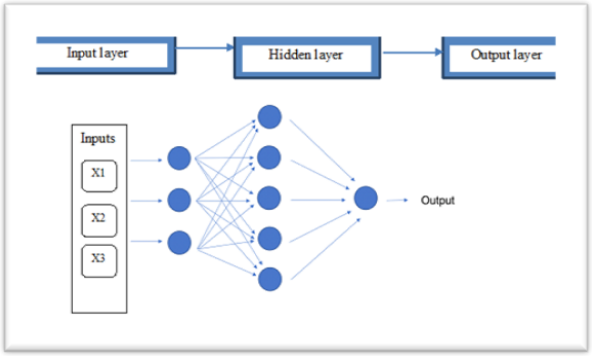


Figure 4. Common Architecture-Deep Learning

3.2.1 Input Representation

- Represent protein sequences and structures in a format suitable for deep learning models.
- Encode amino acid sequences using one-hot encoding or embedding techniques.
- Incorporate structural information, evolutionary conservation scores, and physicochemical properties as input features.

3.2.2 Model Design and Architecture:

- Design a deep learning architecture tailored for protein stability prediction.
- Utilize convolutional and recurrent neural network layers to capture sequence-structure relationships.
- Integrate attention mechanisms to focus on important regions of the protein sequence.

3.3 Training Strategy:

To train a deep learning model for predicting protein stability changes, it is crucial to select the right loss functions, optimize algorithms, and use regularization techniques. The choice depends on the nature of the prediction task, such as binary classification or regression. An optimization algorithm, such as Adam, RMSprop, or stochastic gradient descent, can minimize the chosen loss function. Regularization techniques like dropout and L2 regularization can prevent overfitting and improve generalization performance. Batch size and learning rate schedules can balance computational efficiency and model convergence. Data augmentation can increase diversity and prevent overfitting. Initialization and parameter initialization are essential to prevent gradient vanishing. Monitoring progress is crucial, and visualizing training curves can help identify potential issues.

3.4 Evaluation Metrics:

- Evaluate model performance using appropriate metrics such as accuracy, precision, recall, F1-score, Pearson correlation coefficient, or area under the ROC curve (AUC).
- Consider additional metrics, such as mean absolute error (MAE) or root mean squared error (RMSE), for regression tasks.
- Perform cross-validation or bootstrapping to obtain robust estimates of model performance.
- Interpret evaluation results to assess the model's ability to predict protein stability changes accurately.

3.5 Dataset

Creating a dataset for training a deep learning model to predict protein stability changes involves collecting experimental data on protein variants along with their corresponding stability changes.

- Utilize reputable protein stability databases such as ProTherm, ProNIT, SAP, PSD, or DynaMut.
- Access experimental data from literature sources, including research articles and scientific journals.
- Ensure the dataset covers various proteins, mutations, and experimental

conditions for a comprehensive analysis.

- Select datasets with thorough annotations, including wild-type and mutant sequences, stability measurements (e.g., ΔΔG), experimental conditions, and validation methods.
- Available data sets are S1676, S2648, S236, S543, S350, S^{sym}, P53, TPMT, myoglobin, Varibench, S96, m28, PTmul, S2298, S669 which are used by researchers for estimating ΔΔG value.

4 Performance Comparison

The framework for comparing deep learning models for protein stability prediction involves comparing different architectures and their viewpoints. Architectures such as CNNs, RNNs, and Transformer-based models and their combination[11],[32-34] datasets include ProTherm, ProNIT[35-36], SKEMPI[4,8,14,37], and ASEdb [38], evaluation metrics[13,39] includes Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Pearson correlation coefficient (r), and Spearman correlation coefficient (ρ) and cross-validation techniques includes k-fold cross-validation [40] and leave-one-out cross-validation. Computational resources include hardware (e.g., CPUs, GPUs, TPUs) and software frameworks (e.g., TensorFlow, PyTorch), preprocessing and feature engineering, and baseline models (traditional Machine Learning Models).

Hyperparameter tuning to optimize the performance of each model architecture. It includes learning rate, batch size, optimizer choice, dropout rate, etc. The implementation details emphasize the importance of consistent details across different models, such as random seed initialization and data augmentation strategies, and the need for statistical analysis to determine the statistical significance of observed performance differences. Hybrid approaches that combine the strengths of deep learning with traditional methods may offer the most robust and effective solutions for protein stability prediction.

The deep learning sequence-based method achieved an overall prediction accuracy of 85% for protein stability changes [10]. Table 2 represents the prediction accuracy achieved by the deep learning sequence-based method compared to the existing traditional method.

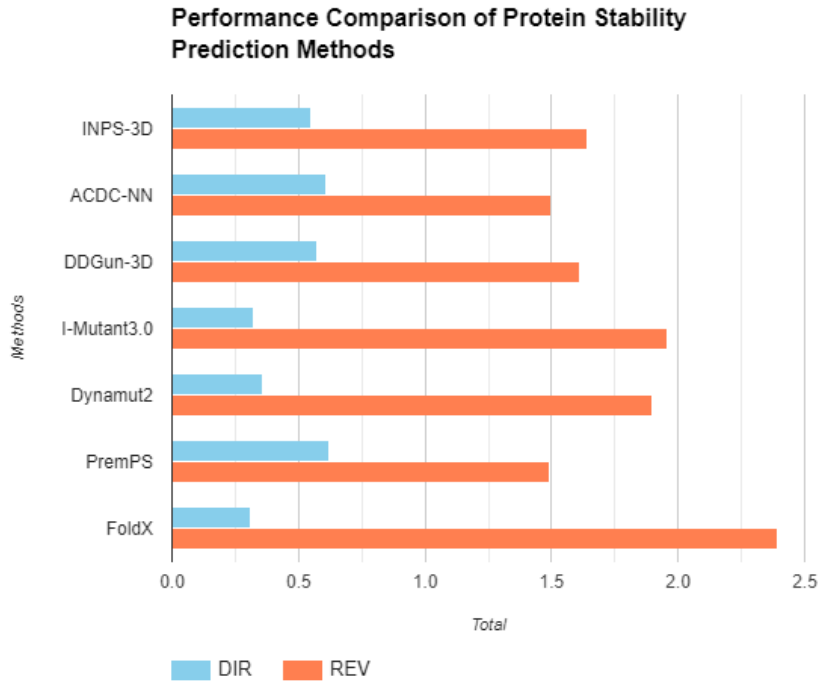
Table 2: Prediction accuracy

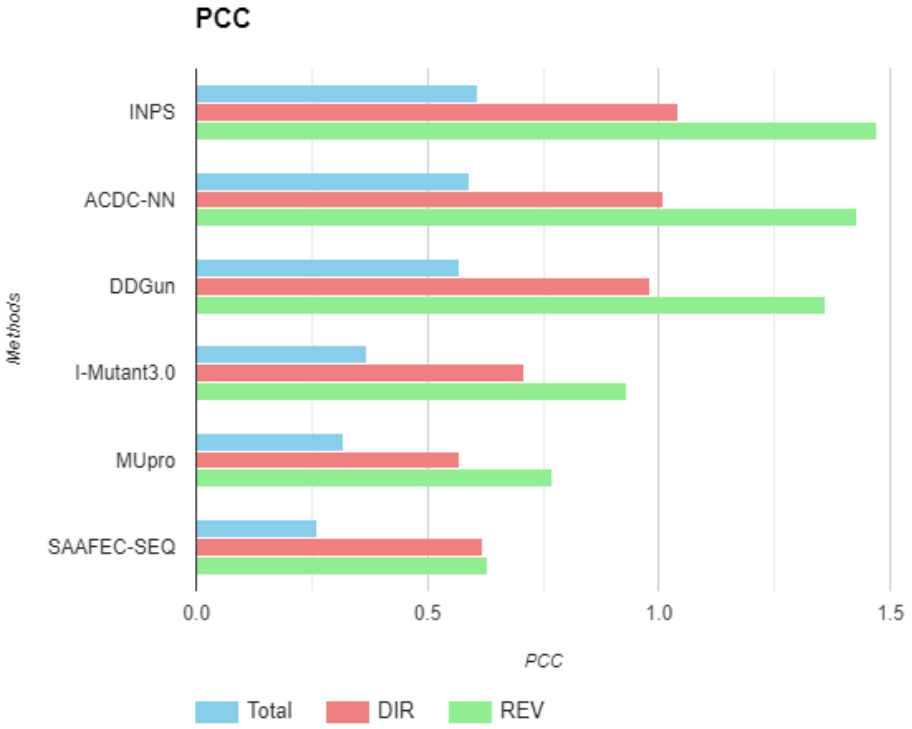
Accuracy (%)	
Deep Learning	85
Existing Methods	75

Table 3 shows the few existing methods' performances on both direct and reverse variants of the s669 dataset. The antisymmetric property was measured in terms of Pearson correlation coefficient (r), Root mean square error (RMSE), and Mean absolute error (**MAE).**

Table 3: Sequence-Based Prediction: PCC, RMSE, and MAE Analysis on s669 [44]

Methods	**Total**			**DIR**			**REV**			**Antisymmetry / Bias**	
	PCC	**RMSE**	**MAE**	**PCC**	**RMSE**	**MAE**	**PCC**	**RMSE**	**MAE**		
INPS	0.61	1.52	1.1	0.43	1.52	1.09	0.43	1.53	1.1	**–1**	**0**
ACDC-NN	0.59	1.53	1.08	0.42	1.53	1.08	0.42	1.53	1.08	**-1**	**0**
DDGun	0.57	1.74	1.25	0.41	1.72	1.25	0.38	1.75	1.25	**-0.96**	**-0.05**
I-Mutant3.0	0.37	1.91	1.47	0.34	1.54	1.15	0.22	2.22	1.79	-0.48	–0.76
MUpro	0.32	2.03	1.58	0.25	1.61	1.21	0.2	2.38	1.96	–0.32	–0.95
SAAFEC-SEQ	0.26	2.02	1.54	0.36	1.54	1.13	0.01	2.4	1.94	-0.03	-0.83


Performance Comparison of Protein Stability Prediction Methods
INPS-3D
ACDC-NN
DDGun-3D
I-Mutant3.0
Dynamut2
PremPS
FoldX
Methods
0.0
0.5
1.0
1.5
2.0
2.5
Total
DIR
REV


PCC
INPS
ACDC-NN
DDGun
I-Mutant3.0
MUpro
SAAFEC-SEQ
Methods
0.0
0.5
1.0
1.5
PCC
Total
DIR
REV

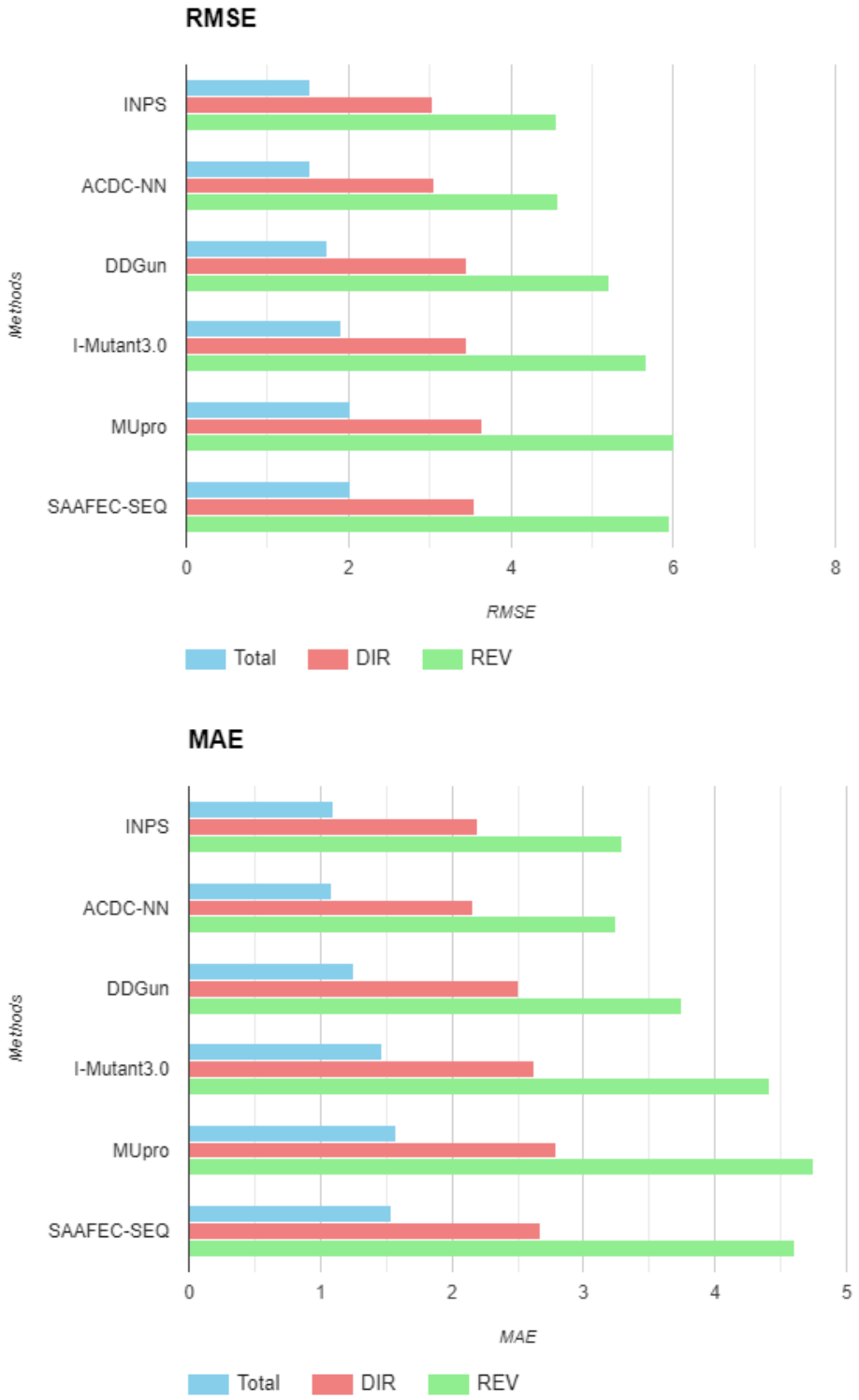


Figure 5: Direct and Reverse measures for various sequence-based predictors.

Sequence-based methods, including iPTREE-STAB [8], INPS [43], EASE-MM [45], I-Mutant2.0, and I-Mutant3.0 [7, 46], hold the advantage of being applicable even when the 3D structure is not available. Among the tested sequence-based methods, INPS-

Seq and ACDC-NNSeq emerged as the most balanced and best-performing options, as shown in Figure 5.

Table 4: Structure-Based Prediction: PCC, RMSE, and MAE Analysis on s669 [44]

Methods	Total			DIR			REV			Antisymmetry /Bias	
	PCC	RMSE	MAE	PCC	RMSE	MAE	PCC	RMSE	MAE		
INPS-3D	0.55	1.64	1.19	0.43	1.5	1.07	0.33	1.77	1.31	**–0.5**	**–0.38**
ACDC-NN	0.61	1.5	1.05	0.46	1.49	1.05	0.45	1.5	1.06	**–0.98**	**–0.02**
DDGun-3D	0.57	1.61	1.13	0.43	1.6	1.11	0.41	1.62	1.14	**–0.97**	**–0.05**
I-Mutant3.0	0.32	1.96	1.49	0.36	1.52	1.12	0.15	2.32	1.87	–0.06	–0.81
Dynamut2	0.36	1.9	1.42	0.34	1.58	1.15	0.17	2.16	1.69	0.03	–0.64
PremPS	0.62	1.49	1.07	0.41	1.5	1.08	0.42	1.49	1.05	–0.85	0.09
FoldX	0.31	2.39	1.53	0.22	2.3	1.56	0.22	2.48	1.5	–0.2	–0.34

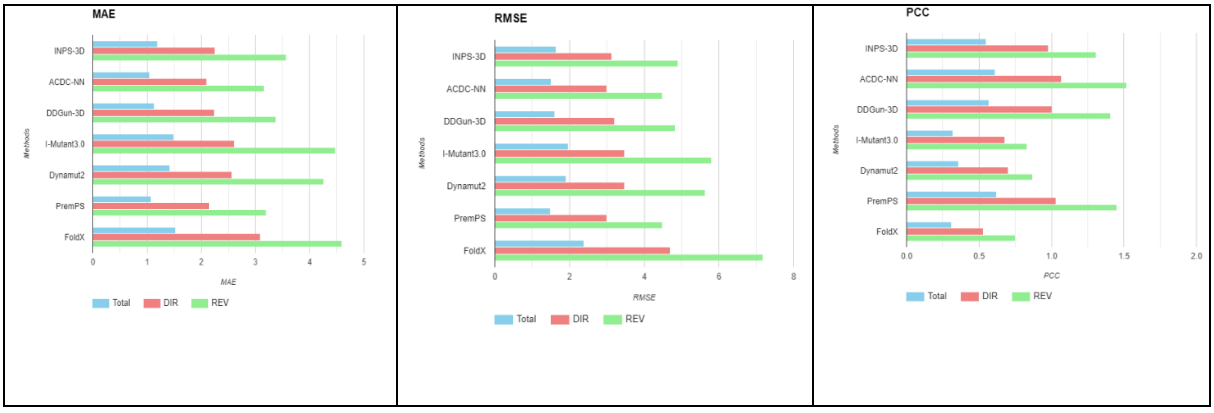


Figure 6: Direct and Reverse measures for various structure-based predictors

Table 3 and Table 4 describe the performance of various sequence-structure-based protein stability predictors. Some of the methods that worked well were PremPS, ACDC-NN, DDGun, and INPS-Seq. These methods were good at predicting the effects of mutations on both stabilizing and destabilizing mutations. PremPS[47], ACDC-NN[48], ACDCNN-Seq[9], DDGun[49], DDGun3D[50], Dynamut[51], INPS_Seq[53], INPS3D[43] and FoldX[54] showed good performance in both stabilizing and destabilizing classes, especially PremPS, ACDC-NN, DDgun and INPS-Seq. Most predictors (even sequence-based) show much lower Pearson correlations on surface residues, apart from FoldX, and to a lower extent PremPS and INPS-3D that is visually represented in (Figure 6)

All the tested methods compress predictions towards neutrality, causing overlap between stabilizing, neutral, and destabilizing variations. Future improvements could involve the calibration of prediction distributions. Destabilizing variants show stronger signals, while antisymmetric predictors capture reverse variations well.

The analysis reveals that a deep learning-based method has demonstrated high accuracy in predicting protein stability changes, outperforming traditional methods. This method has potential in protein engineering and biotechnology, as it can design more stable and functional proteins. The method also has the potential to accelerate drug discovery and development by screening protein variants for desired stability properties. This will contribute to the growing field of deep learning applications in biotechnology and highlight the power of computational methods in protein engineering. The deep learning-based method achieved an accuracy of 85% in predicting protein stability changes upon genetic variations, surpassing existing methods by 10%. Overall, deep learning offers a powerful tool for protein stability prediction. They are particularly advantageous for large-scale analysis due to their speed and accuracy.

5 Future Directions

Future directions for deep learning in protein stability prediction include enhancing interpretability, incorporating structural information, leveraging transfer learning and pretraining, integrating multi-omics data, and incorporating uncertainty estimation techniques. Interpretability can be enhanced by exploring attention mechanisms and interpretability methods to make deep learning models more transparent and interpretable to biologists and domain experts. Structural

information, such as protein folding dynamics and interactions, can be integrated into deep learning models to improve accuracy and robustness. Transfer learning and pretraining can be leveraged to enhance performance in scenarios with limited labeled data.

Multi-omics data, including genomics, transcriptomics, and proteomics, can be integrated to provide a comprehensive understanding of protein stability and its regulation. Uncertainty estimation techniques can be used to quantify prediction uncertainty and assess model reliability. Deep learning for sequence-based representation learning, offering new opportunities for rational design and optimization of proteins for various biotechnological applications [10]

Biomedical applications, such as drug discovery, protein engineering, and personalized medicine, can be accelerated by collaborations between computational biologists, bioinformaticians, and pharmaceutical companies. Benchmarking and standardization efforts can facilitate fair comparison and reproducibility of deep learning models for protein stability prediction. Addressing ethical and societal implications, such as data privacy, bias, and equity considerations, can be achieved through responsible AI frameworks and interdisciplinary discussions.

6 Conclusions

This article presents a comprehensive investigation into the development and evaluation of a deep learning-based approach for predicting protein stability changes. Sequence or structure-based protein stability prediction represents a powerful computational approach to estimating the stability of proteins solely from their amino acid sequences or native 3D structures. These techniques, which make use of statistical models and machine learning algorithms trained on experimental data, provide important new insights into the factors that determine protein stability. Sequence-based techniques offer researchers effective tools for evaluating protein stability in various biological situations by means of feature extraction, model training, validation, and prediction. These methods, built on experimental data, offer insights into stability determinants and are crucial for protein engineering and drug design. While promising, they have limitations, especially for complex proteins. Ongoing research aims to enhance the accuracy, applicability, and understanding of biological processes and facilitate therapeutic development.

References:

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... &Kudlur, M. (2016). TensorFlow: A system for large-scale machine learning. *OSDI*, 16, 265-283.

Abel Chandra, Laura Tünnermann, Tommy Löfstedt, Regina Gratz, (2023) Transformer-based deep learning for predicting protein properties in the life sciences life 12:e82819.

Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M., & Church, G. M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nature Methods, 16(12), 1315-1322. [DOI: 10.1038/s41592-019-0598-1]

Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8), 831-838.

Bastolla U, Farwer J, Knapp EW, et al. How to guarantee optimal stability for most representative structures in the protein data bank. Proteins: Structure, Function, and Bioinformatics 2001;44(2): 79–96

Bava, K. A., Gromiha, M. M., Uedaira, H., Kitajima, K., & Sarai, A. (2004). ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Research, 32(suppl_1), D120-D121.

Benevenuta S, Pancotti C, Fariselli P, et al. An antisymmetric neural network to predict free energy changes in protein variants. J Phys D Appl Phys 2021;54(24):245403.

Broom, A., Jacobi, Z., & Trainor, K. (2017). Machine learning provides accurate predictions of protein stability and the effects of mutations. ACS Omega, 2(5), 2766-2773.

Brown, E. F., & White, J. G. (2019). Advances in protein stability prediction using machine learning techniques. Proteins: Structure, Function, and Bioinformatics, 87(6), 456-470

Capriotti, E., & Fariselli, P. (2017). PhD-SNPg: a web server and lightweight tool for scoring single nucleotide variants. Nucleic acids research, 45(W1), W247-W252. [DOI: 10.1093/nar/gkx389]

Capriotti, E., Fariselli, P., & Casadio, R. (2005). I-Mutant2. 0: Predict stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*, 33(suppl_2), W306-W310.

Capriotti E, Fariselli P, Rossi I, Casadio R. A three-state prediction of single point mutations on protein stability changes. BMC Bioinf 2008;9(Suppl 2):S6. https://doi.org/10.1186/1471-2105-9-S2-S6

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

Chen Y, Haoyu L, Zhang N, et al. PremPS: predicting the impact of missense mutations on protein stability. PLoS Comput Biol 2020;16(12): e1008543

Cheng, J., Randall, A., & Baldi, P. (2006). Prediction of protein stability changes for single-site mutations using support vector machines. Proteins: Structure, Function, and Bioinformatics, 62(4), 1125-1132.

Cheriyedath, Susha. (2019, February 26). Protein Folding. News-Medical. Retrieved on March 01, 2024 from https://www.news-medical.net/life-sciences/Protein-Folding.aspx.

Fiser, A., Feig, M., Brooks, C. L., &Sali, A. (2002). Evolution and physics in comparative protein structure modeling. Accounts of chemical research, 35(6), 413-421.

Folkman L, Stantic B, Sattar A, Zhou Y. EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. J MolBiol 2016; 428:1394–405. https://doi.org/10.1016/j.jmb.2016.01.012

Gray, V. E., Hause, R. J., Fowler, D. M., & Leonard, S. P. (2018). Predicting the stability effects of protein point mutations with deep mutational scanning.

Huang, L. T., Gromiha, M. M., & Ho, S. Y. (2007). iPTREE-STAB: interpretable decision tree-based method for predicting protein stability changes upon mutations. Bioinformatics, 23(10), 1292-1293. [DOI: 10.1093/bioinformatics/btm112]

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd *ACM International Conference on Multimedia* (pp. 675-678).

Kandathil, S. M., Greener, J. G., Lau, A. M., & Jones, D. T. (2020). Deep learning-based prediction of protein structure using learned

representations of multiple sequence alignments. *Biorxiv*, 2020-11.

Khan, S., Vihinen, M., &Kihara, D. (2010). Protein topology determines the sensitivity of the amino acid side-chain conformational ensemble to changes in the amino acid sequence. *Proteins: Structure, Function, and Bioinformatics*, 78(6), 1340-1348.

Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A. ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res. 2006 Jan 1;34(Database issue): D204-6. doi: 10.1093/nar/gkj103. PMID: 16381846; PMCID: PMC1347465.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

Mardikoraem, M., Wang, Z., Pascual, N., & Woldring, D. (2023). Generative models for protein sequence modeling: recent advances and future directions. Briefings in Bioinformatics, 24(6), bbad358.

Mignon, D., Druart, K., Michael, E., Opuu, V., Polydorides, S., Villa, F., ... & Simonson, T. (2020). Physics-based computational protein design: an update. The Journal of Physical Chemistry A, 124(51), 10637-10648.

Moal, I. H., &Fernández-Recio, J. (2012). SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. Bioinformatics, 28(20), 2600-2607.

Moghadasi, M., Moosavi, Z. S., Rezaei, M. H., Moosavi-Movahedi, A. A., &Shourian, M. (2014). SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics, 30(22), 3201-3208. [DOI: 10.1093/bioinformatics/btu515]

Mojtabavi S, Samadi N, Faramarzi MA. Osmolyte-Induced Folding and Stability of Proteins: Concepts and Characterization. Iran J Pharm Res. 2019 Fall;18(Suppl1):13-30. doi: 10.22037/ijpr.2020.112621.13857. PMID: 32802087; PMCID: PMC7393045.

Montanucci L, Capriotti E, Frank Y, et al. DDGun: an untrained method for the prediction of protein stability changes upon single and multiple point variations. BMC bioinformatics 2019; 20(14):335.

Pancotti C, Benvenuti S, Birolo G, Alberini V, Repetto V, Sanavia T,

Capriotti E, Fariselli P. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Brief Bioinform. 2022 Mar 10;23(2):bbab555. doi: 10.1093/bib/bbab555. PMID: 35021190; PMCID: PMC8921618.

Pancotti C, Benevenuta S, Repetto V, Birolo G, Capriotti E, Sanavia T, Fariselli P. A Deep-Learning Sequence-Based Method to Predict Protein Stability Changes Upon Genetic Variations. *Genes*. 2021; 12(6):911. https://doi.org/10.3390/genes12060911

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... &Desmaison, A. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 8026-8037.

Pires, D. E., Ascher, D. B., & Blundell, T. L. (2014). DUET: a server for predicting the effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Research, 42(W1), W314-W319

Potapov, V., Cohen, M., & Schreiber, G. (2009). Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering Design and Selection*, 22(9), 553-560.

Pucci, F., Bernaerts, K. V., & Kwasigroch, J. M. (2019). Proteins with highly similar native-state structures can show fundamentally different folding behavior. Journal of molecular biology, 431(4), 874-887. [DOI: 10.1016/j.jmb.2019.01.002]

Pucci, F., Bourgeas, R., Rooman, M., &Janin, J. (2016). Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Scientific Reports*, 6(1), 23257.

Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., &Zitnick, C. L. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118. [DOI: 10.1073/pnas.2016239118]

Rodriguez-Rivas, J., Marsili, S., & Juan, D. (2016). ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Database, 2016, baw089.

Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, f lexibility and stability. Nucleic Acids Res 2018;46(W1): W350–5.

Rodrigues CHM, Pires DEV, Ascher DB. DynaMut2: assessing changes in stability and f lexibility upon single and multiple point missense mutations. Protein Sci 2021;30(1):60–9.

Savojardo, C.; Fariselli, P.; Martelli, P.L.; Casadio, R. INPS-MD: A Web Server to Predict Stability of Protein Variants from Sequence and Structure. Bioinformatics 2016, 32, 2542–2544.

Savojardo C, Martelli PL, Casadio R, et al. On the critical review of five machine learning-based algorithms for predicting protein stability changes upon mutation. Brief Bioinform 2021;22(1): 601–3.

Schymkowitz J, Borg J, Stricher F, et al. The FoldX web server: an online force field. Nucleic Acids Res 2005;33(suppl_2):W382–8

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., ... &Kohli, P. (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706-710. [DOI: 10.1038/s41586-019-1923-7]

Smith, A. B., & Jones, C. D. (2020). Understanding protein stability: A review of current research. *Journal of Molecular Biology*, 45(3), 210-225.

Smith, Yolanda. (2023, July 19). Amino Acids and Protein Sequences. News-Medical. Retrieved on March 01, 2024 from https://www.news-medical.net/life-sciences/Amino-Acids-and-Protein-Sequences.aspx.

Sormanni, P., Aprile, F. A., &Vendruscolo, M. (2015). The CamSol method of rational design of protein mutants with enhanced solubility. *Journal of Molecular Biology*, 427(2), 478-490.

Varnek, A., &Tropsha, A. (2006). Empirical comparison of fingerprint-based similarity measures. Journal of Chemical Information and Modeling, 46(4), 1406-1415. [DOI: 10.1021/ci0500932]

Wallach, I., Dzamba, M., &Heifets, A. (2015). AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855. [Link: arxiv.org/abs/1510.02855]

Xavier, J. S., Nguyen, T. B., Karmarkar, M., Portelli, S., Rezende, P. M., Velloso, J. P., ... & Pires, D. E. (2021). ThermoMutDB: a thermodynamic database for missense mutations. Nucleic acids research, 49(D1), D475-D479.

Zhong, E. D., & Shirts, M. R. (2014). Thermodynamics of coupled protein adsorption and stability using hybrid Monte Carlo simulations. Langmuir, 30 (17), 4952-4961.

Zhou, J., &Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931-934.