MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
KYIV NATIONAL UNIVERSITY OF TECHNOLOGIES AND DESIGN
Faculty of Chemical and Biopharmaceutical Technologies
Department of Biotechnology, Leather and Fur

QUALIFICATION THESIS

on the topic **Leveraging biological experimental mutation and functional data to validate an AI-based protein design method**
First (Bachelor's) level of higher education
Specialty 162 "Biotechnology and Bioengineering"
Educational and professional program "Biotechnology"

Completed: student of group BEBT-21
Wang WEIYAN

Scientific supervisor
Olga ANDREYEVA, Dr. Sc., Prof.

Reviewer
Tetiana HALENOVA, Ph.D., As. Prof.

Kyiv 2025

KYIV NATIONAL UNIVERSITY OF TECHNOLOGIES AND DESIGN

Faculty: Chemical and Biopharmaceutical Technologies
Department: Biotechnology, Leather and Fur
First (Bachelor's) level of higher education
Specialty: 162 Biotechnology and Bioengineering
Educational and professional program Biotechnology

APPROVE
Head of Biotechnology, Leather and Fur
Department, Professor,
Dr. Sc., Prof.
___________Olena MOKROUSOVA
«___»_______________2025

ASSIGNMENTS
FOR THE QUALIFICATION THESIS
Wang Weiyan

1. Thesis topic **Leveraging biological experimental mutation and functional data to validate an AI-based protein design method**
Scientific supervisor Dr. Sc., Prof. Olga Andreyeva

approved by the order of KNUTD "05" March 2025, № 50-уч

2. Initial data for work: assignments for qualification thesis, scientific literature on the topic of qualification thesis, materials of Pre-graduation practice

3. Content of the thesis (list of questions to be developed): literature review; object, purpose, and methods of the study; experimental part; conclusions

4. Date of issuance of the assignments 05.03.2025

WORK CALENDAR

№	The name of the stages of the qualification thesis	Terms of performance of stage	Note on performance
1	Introduction	until 11 April 2025	
2	Chapter 1. Literature review	until 20 April 2025	
3	Chapter 2. Object, purpose, and methods of the study	until 30 April 2025	
4	Chapter 3. Experimental part	until 11 May 2025	
5	Conclusions	until 15 May 2025	
6	Draw up a bachelor's thesis (final version)	until 25 May 2025	
7	Submission of qualification work to the supervisor for feedback	until 27 May 2025	
8	Submission of bachelor's thesis to the department for review (14 days before the defense)	28 May 2025	
9	Checking the bachelor's thesis for signs of plagiarism (10 days before the defense)	01 June 2025	Similarity coefficient ____% Citation rate ____%
10	Submission of bachelor's thesis for approval by the head of the department (from 7 days before the defense)	04 June 2025	

I am familiar with the task:
Student ____________ Wang WEIYAN

Scientific supervisor ____________ Olga ANDREYEVA

SUMMARY

Wang WEIYAN. Leveraging biological experimental mutation and functional data to validate an AI-based protein design method. – Manuscript.

Qualification thesis on the specialty 162 «Biotechnology and Bioengineering». – Kyiv National University of Technologies and Design, Kyiv, 2025.

Protein mutation design is a pivotal technology for precise regulation of protein functions, with significant applications in biomedicine and industrial enzyme engineering. Traditional experimental methods face limitations such as lengthy cycles, high costs, and low mutation-site hit rates. The emergence of artificial intelligence (AI) offers innovative solutions to these challenges. This study systematically analyzes ProtSSN, an AI-based protein mutation design software developed by Professor Hong Liang's team at Shanghai Jiao Tong University. Using public datasets, the research verifies the algorithm's predictive accuracy and investigates its structure-sensing mechanism, elucidating its advantages and limitations in practical applications.

The study validates ProtSSN's performance through quantitative experiments and innovatively explores the correlation between protein secondary structures, Solvent Accessible Surface Area (SASA), and mutation effects. Findings reveal that ProtSSN integrates protein sequence semantics and 3D structural topology via a dual-modal pre-training framework. Leveraging Equivariant Graph Neural Networks (EGNN), it quantifies structural features (e.g., hydrophobic cores in α-helices, hydrogen bonds in β-sheets) to analyze mutation-induced perturbations. ProtSSN's lightweight architecture overcomes computational bottlenecks of traditional molecular simulations, enhancing wet-lab mutant screening efficiency for industrial enzyme optimization and antibody affinity maturation. However, the model's handling of dynamic irregular loops requires improvement, suggesting future integration of molecular dynamics or expanded training data for specialized proteins.

A multi-dimensional evaluation framework confirms ProtSSN's efficacy in structure-driven mutation design, establishes sequence-structure-function correlations, and provides a reusable methodology for AI protein tool development.

This work advances protein engineering from empirical trial-and-error toward a computational paradigm, with potential applications in enzyme catalyst design and therapeutic antibody development.

Key words: *ProtSSN; structure perception; decoupling attention; performance verification; artificial intelligence; protein mutation design*

TABLE OF CONTENTS

INTRODUCTION

This study focuses on the artificial intelligence-driven protein design software ProtSSN. By integrating deep mutation scan data, protein secondary structure prediction, and solvent accessible surface area (SASA) correlation analysis, it systematically reveals its technical advantages, mechanism of action, and engineering application potential in mutation effect prediction. Studies have shown that ProtSSN achieves quantitative coding of geometric features of the protein microenvironment (such as hydrogen bond networks and hydrophobic cores) through a dual-modal collaborative pre-training framework (fusion of sequence semantics and three-dimensional structure topology) and an equivariant graph neural network (EGNN), significantly improving the accuracy of mutation prediction. In the Protein Gym benchmark test, its prediction performance for single-point mutations (Spearman ρ=0.429) and multi-point mutations (ρ=0.550) is significantly better than that of traditional models, especially in the regions of regular secondary structures (α -helix, extended chain). Further research revealed that the solvent-accessible surface area (SASA) was significantly globally negatively correlated with the prediction error (weighted average ρ= -0.25). The prediction reliability was stronger in the high SASA region (>0.4), while the conformational synergy effect in the core region increased the complexity of the prediction.

The relevance of this topic lies in the analysis of the structural mechanism of artificial intelligence protein mutation software.

The main objectives include determining the performance advantages of ProtSSN in the prediction of mutation effects, establishing a quantitative connection between SASA values and prediction errors, and analyzing the relationship between the secondary structure and prediction accuracy.

The methods include conducting performance benchmark tests of single-point and multi-point mutations on the Protein Gym dataset, conducting statistical analysis of the secondary structure distribution using SOPMA and DSSP, and performing Spearman correlation modeling of the prediction error relationship of SASA.

The object of the study – ProtSSN.

The subject of the study – Testing an AI-based protein design method.

The scientific novelty lies in systematically studying the quantitative connection between AI-based mutation prediction and microenvironment characteristics, such as secondary structure rules and SASA.

The practical significance of this research lies in its open-source software development, which promotes a paradigm shift in synthetic biology and biomanufacturing – from empirical trial and error approaches to computation-first strategies.

This study demonstrates that ProtSSN can significantly enhance the screening efficiency of wet laboratory mutants, while reduce the cost of experimental trial and error, and enable the rapid design of therapeutic antibodies and industrial enzymes. The global negative correlation between SASA and the prediction error (weighted ρ=-0.25) provides an operational guideline for optimizing the stability of residue levels in protein engineering. Meanwhile, the analysis of the relationship between the secondary structure and the prediction accuracy also provides a direction for the future development of software.

Chapter I LITERATURE REVIEW

1.1 Background of the Intelligent Development of Protein Design Software

1.1.1 Limitations of Traditional Protein Design Approaches

The traditional protein design method is based on classical molecular mechanics theory. Its core principle is to precisely calculate the interaction forces between atoms, such as electrostatic attraction and van der Waals forces, to construct mathematical models and thereby simulate the protein folding process and stable state. This method follows a fundamental assumption that the structure of natural proteins corresponds to the lowest stable state in the entire energy system. Therefore, protein design can be understood as finding the optimal three-dimensional structural arrangement scheme that minimizes the overall energy in a high-dimensional space composed of tens of thousands of atomic coordinates. However, this theoretical model based on thermodynamic equilibrium states has gradually exposed deeper problems when dealing with the dynamics and complexity of actual biological systems.

From the perspective of physical modeling, the parametric system of the classical force field has essential simplification. Traditional tools break down protein-protein interactions into linear combinations of discrete energy terms such as van der Waals forces, electrostatic interactions, and hydrogen bonds. Although this "divide and conquer" strategy can reduce computational complexity, it sacrifices the quantum effects and dynamic synergy of many-body interactions. For example, the polarization effect of the hydrogen bond network shows dynamic fluctuation characteristics in real solutions, while the traditional model simplifies it as a static effect with fixed distances and angles. The instantaneous electron cloud distribution of the π-π packing interaction is approximated as a rigid geometric match, thereby ignoring the induced dipole interaction between the aromatic rings. These simplifications result in the energy function being unable to accurately describe the delicate balance of entropy-enthalpy compensation during the protein folding process, especially causing systematic deviations in the modeling of the flexible loop region and the solvent exposure interface[1].

At the level of computational complexity, protein design problems have been proved to belong to the mathematical category of NP-hard[2]. The Monte Carlo sampling and molecular dynamics simulation adopted by the traditional method essentially traverse the high-dimensional conformational space through random rows. With the increase in the number of mutation sites, the dimension of the conformational space shows an exponential expansion. The number of conformational combinations that need to be evaluated for a design task containing n mutation sites can reach 20^n (20 amino acid possibilities), which makes the search for the global optimal solution computationally infeasible. Even if the enhanced sampling technique is introduced, the local minimum trap will still cause the algorithm to converge prematurely to the suboptimal solution. More crucially, the rough approximation of the energy function amplifies the error of conformational sampling. When the model performs optimization on the wrong potential energy surface, the input of computing resources may instead lead to an intensification of the deviation between the prediction results and the real biological system[3].

From the perspective of functional design, there is a mismatch between the theoretical goals of traditional methods and engineering requirements. Although the principle of minimizing free energy focuses on the structural stability of proteins, it does not incorporate key functional indicators such as catalytic activity and substrate specificity into the unified optimization framework. This single-goal-oriented design logic often leads to the contradictory phenomenon of "excessive stability and functional loss", that is, although the mutation scheme can improve thermal stability, it inhibits catalytic efficiency due to the rigidity of the active pockets. In addition, the processing ability of the physical model for the synergistic mutation effect is also limited. When the mutation sites are distributed in different domains, the nonlinear coupling of their long-range interactions is difficult to accurately describe through the simple superposition of energy terms. This limitation is particularly prominent in application scenarios such as industrial enzyme modification that require simultaneous optimization of multiple characteristics, thereby forcing researchers to rely on empirical trial and error and experimental iterations, deviating from the original intention of rational design.

The deep-seated theoretical predicament actually stems from the dynamic nature that life systems possess. The traditional way regards proteins as isolated thermodynamic systems, but ignores that in the cellular environment, the conformation of proteins fluctuates continuously and intermolecular interactions are frequent. Biological principles such as chaperone protein-assisted folding and post-translational modification regulation have not been effectively simulated in existing physical models. This static thinking mode leaves traditional methods inadequate when it comes to the functional adaptation requirements of proteins in the real biological environment, exposing the inherent contradiction between the theoretical system and the complexity of life.

1.1.2 Opportunities for the Application of Artificial Intelligence Technology

The field of protein mutation design is at a critical period of transformation from relying on experience to relying on data. The integration of artificial intelligence technology has opened up a brand-new path to overcome the drawbacks of traditional methods. Deep learning, by deeply mining the potential rules in a large number of protein sequences and structures, has broken the limitations of physical modeling in terms of computational efficiency and can also model complex biological systems from multiple dimensions. The core of this technological leap is the redefinition of the mapping logic of "sequence – structure - function": traditional tools rely on clearly defined energy functions and conformational sampling operations, while artificial intelligence, with the help of implicit feature learning processes, transforms protein design problems into pattern recognition and construction tasks in high-dimensional spaces.

The development trajectory of existing protein design software clearly demonstrates the evolution of this technology. Early tools such as Rosetta Design constructed energy functions based on molecular mechanics principles and optimized mutation schemes through Monte Carlo sampling. However, their computational complexity increased exponentially with the number of mutations sites[4]. Take the design of industrial enzymes as an example. The traditional method requires thousands of CPU hours for the full combination analysis of five mutation sites, and the prediction accuracy of long-term synergistic effects is relatively low. With the breakthroughs of structure prediction tools such as AlphaFold[5], generative AI models have begun to emerge. ProteinMPNN captures sequence evolution patterns through pre-trained language models and can generate stable mutation schemes within minutes[6]. The ESM series models utilize self-supervised learning to parse sequence-functional associations and can achieve an accuracy rate of over 80% in the task of functional annotation. However, these tools still have significant limitations. The lack of structural constraints in sequence models can easily lead to physically unreasonable designs, while pure structural models have difficulty capturing evolutionary conservation characteristics.

Under this background, multimodal fusion has become a key direction for technological breakthroughs. The ProtSSN model developed by Shanghai Jiao Tong University[7] innovatively integrates the three-dimensional structure information of proteins with sequence data for collaborative modeling. Its core innovation lies in the construction of a "structure-guided mutation effect prediction" paradigm. This model encodes the atomically level features of the local conformation with the aid of the Geometric Vector perceptron (GVP) and employs the decoupled attention mechanism to effectively achieve the dynamic balance between the sequence evolution pattern and the spatial physical constraints[7]. Compared with the sequence-dependent model, ProtSSN has significant advantages in the identification of catalytic active sites and the design cycle. This technical advantage is attributed to ProtSSN's unified modeling of the multi-level characteristics of proteins. From the hydrophobic effect of atomic spacing to the allosteric communication between domains, the model achieves the fusion cognition of cross-scale rules through hierarchical feature extraction, providing a more efficient and precise solution for protein mutation design. Moreover, in terms of model training, ProtSSN has pre-trained with a large amount of data to create a model with stronger applicability. ProtSSN combines the protein structure data predicted by AlphaFold with the sequence information in UniProt[8] to build a system capable of analyzing protein characteristics across different scales. It can not only better cope with complex protein structures, but also, through special coding techniques, convert three-dimensional structural information into numbers that are convenient for calculation, greatly accelerating the calculation speed

In the future, the application of artificial intelligence in the field of protein design will revolve around three major directions: First, build a more interpretable modeling system, and use technologies such as Grad-CAM to visualize the decision-making logic of the model and enhance the credibility of the prediction results; The second is to build a closed-loop system that deeply integrates dry and wet experiments, and achieve bidirectional optimization of experimental data and computational models through active learning strategies. The third is to explore multi-objective collaborative optimization algorithms to seek the optimal solution among contradictory indicators such as protein stability, catalytic activity, and substrate specificity. The open-source nature of ProtSSN and its outstanding performance in the Protein Gym[9] benchmark test have provided significant support for the co-construction and sharing of the technical ecosystem within the field, accelerating the transformation of protein design from the traditional experience-driven model to a precise design paradigm dominated by data and algorithms. Inject new innovative impetus into cutting-edge fields such as synthetic biology and targeted drug development.

1.2 Introduction to ProtSSN Software and Its Core Technological Breakthroughs

1.2.1 ProtSSN Software Architecture and Function Positioning

ProtSSN is an open-source multimodal protein language model jointly developed by Professor Hong Liang's team from Shanghai Jiao Tong University and the Shanghai Artificial Intelligence Laboratory. ProtSSN adopts a unique dual-modal collaborative pre-training framework, deeply integrating protein sequence and structural information. This framework optimizes the current situation where protein sequence data is abundant but crystal structure data is relatively scarce, and constructs a funnel-shaped learning pipeline. In this architecture, the sequence encoder and the structure encoder complement each other. The sequence encoder inherits the parameters of pre-trained models such as ESM-2. Through learning from massive protein sequences, it can accurately analyze the coevolution laws in amino acid sequences, discover the hidden long-range dependencies in the sequences, and provide a solid sequence foundation for subsequent analysis.

The construction of the structural encoder relies on the Geometric vector perceptron (GVP), whose function is to extract the spatial feature information at the atomic level of proteins. This module not only analyzes geometric parameters such as the Cα skeleton topology and side chain orientation, but also constructs residue maps through the K-nearest neighbor (kNN) algorithm, encoding the local conformation of proteins as discretized structural elements. This encoding mechanism can precisely capture the subtle features of protein structures, providing a crucial structural information basis for the study of protein functions.

The adaptive fusion of the two modes relies on the dynamic routing algorithm, which can automatically optimize the weight ratio of the sequence mode and the structural mode according to the conformational rigidity characteristics of the target area. In the conformational flexible region, since the sequence information plays a dominant role in the function, the algorithm will enhance the influence of the sequence mode. In the conformational rigidity region, the weight of the structural mode will be prioritized for improvement. Through this dynamic adjustment, ProtSSN can fully leverage the advantages of each modality in various protein analysis scenarios, comprehensively and accurately understanding protein characteristics.

The core function of ProtSSN is to accurately determine the possible impact of mutations on protein function by analyzing the overall distribution pattern of amino acids in protein sequences and fully considering the synergistic effects among amino acid sites. It can directly predict the effects of unknown protein variants in zero-shot scenarios without the need for pre-training specific protein data[7].

What is more worthy of attention is that ProtSSN, as an open-source software, not only discloses all models but also comes with detailed user manuals. This feature enables researchers to easily integrate it into various research processes, significantly lowering the technical threshold for protein data analysis while helping researchers more efficiently mine the value of data. It plays a positive role in promoting the overall development of the field of protein research.

1.2.2 Core Technology Breakthroughs and Theoretical Innovations

The core of the technological breakthrough of ProtSSN software lies in the comprehensive innovation formed through the creative integration of multimodal information, the construction of an advanced model architecture system, and the introduction of cutting-edge algorithm support.

Traditional protein characterization methods mostly focus on a single dimension of amino acid sequences. ProtSSN, however, innovatively integrates cross-modal information of protein sequences and three-dimensional structures. With the progress made in the field of structure prediction by deep learning tools such as AlphaFold, it is possible to obtain high-precision three-dimensional protein models on a large scale and construct a brand-new multimodal analysis system. Given the close correlation between protein functions and their spatial configurations, microscopic structural features such as the connection mode of Cα atoms and the three-dimensional arrangement of side chains directly reflect the energy distribution of molecular folding and potential action sites. This model overcomes the limitations of traditional single-modal research by designing a sequence-conformation joint analysis module, achieving a comprehensive characterization of protein properties from multiple dimensions. It has significantly improved the accuracy of functional prediction and provided a multi-dimensional structural basis for in-depth research on the impact of protein mutations.

ProtSSN adopts an innovative dual-channel feature extraction framework. Its sequence feature extraction module draws on the parameter initialization methods of advanced models such as ESM-2 to analyze the co-evolution law among amino acid residues in polypeptide chains, that is, biological evolution causes the positions of various residues in the primary structure of proteins to form mutually restrictive variation associations. These coevolutionary characteristics provide an important basis for functional prediction. The spatial feature extraction component adopts the geometric vector sensing technology to capture the three-dimensional conformational features at the atomic scale and transform the continuous spatial distribution into the representation of discrete structural units. This complementary architecture gives full play to the advantages of different data modalities. The introduction of the dynamic routing mechanism enables the system to autonomously adjust the fusion weights of the two features based on the stability of the local conformation. This adaptive strategy allows the model to intelligently adjust the utilization ratio of sequence and structural information when facing different structural characteristics and prediction tasks, in order to obtain more reliable prediction results.

The technological breakthrough of ProtSSN is also inseparable from the application of cutting-edge algorithms. In the process of structural quantization, the software converts the complex three-dimensional structural information into computable discrete symbols through geometrically driven structural quantization technology, reducing the computational complexity while retaining the key characteristics of the structure.

The application of the decoupled attention mechanism enables the model to dynamically balance the relationship between sequence and structural information when processing them, pay more flexible attention to the features of different regions, and effectively prevent the prediction deviation caused by uneven information processing. In addition, ProtSSN adopts a self-supervised learning strategy and conducts pre-training on a large-scale protein sequence database, thereby learning the general characteristics of protein sequences and structures, enhancing the generalization ability of the model, reducing the reliance on a large amount of labeled data, and further improving its adaptability and accuracy in different protein mutation prediction tasks.

1.3 Significance and Research Contents of This Project

1.3.1 Research Significance

The transformation of protein design tools towards intelligence is the core challenge faced in the intersection of synthetic biology and computational biology. Because traditional protein design methods mainly rely on physical simulation and experimental trial and error, there are prominent problems such as low computational efficiency and difficulty in predicting the synergistic effect of multi-point mutations. This has largely restricted the research and development progress of practical application scenarios such as industrial enzyme modification and antibody drug research and development. This study takes the ProtSSN software developed by Professor Hong Liang's team from Shanghai Jiao Tong

University as the research object. Through systematic testing and analysis methods, it verifies its technical advantages as a new generation of artificial intelligence protein design tool, which has important theoretical research value and practical application significance.

This study innovatively proposes a systematic evaluation framework for protein multimodal characterization models in terms of theoretical methods. In view of the limitation that existing studies mostly focus on single-dimensional indicators such as prediction accuracy or operation speed, which are difficult to comprehensively reflect the practical application value of the model, a comprehensive evaluation system including key dimensions such as structural restoration degree, functional inference ability, and resource consumption efficiency is designed. The verification experiments carried out based on the Protein GYM standard dataset and combined with visualization analysis technology have confirmed the effectiveness of this evaluation paradigm. This theoretical innovation not only provides a reliable method for the performance verification of the ProtSSN model but also offers new ideas for the formulation of evaluation standards in the field of protein artificial intelligence research.

This study verified in the technical implementation that the modeling strategy integrating the spatial conformation and sequence evolution characteristics of proteins has significant advantages. ProtSSN, which integrates three-dimensional structural parameters and conservation analysis results, effectively improves the limitations of traditional methods in predicting local conformational dynamic changes. Experimental evaluation shows that this model performs excellently in residue variation prediction, improving the reliability of single-point and compound mutation prediction while significantly compressing the design cycle. This progress has brought about significant technological innovations in the field of biomedicine. It has certain application values in aspects such as shortening the R&D cycle in industrial enzyme design and precisely regulating molecular recognition characteristics and conformational stability in the field of antibody drug optimization.

1.3.2 Research Content

This study focuses on the ProtSSN protein intelligent design system developed by Professor Hong Liang's team from Shanghai Jiao Tong University, and constructs a multi-dimensional evaluation framework to mainly examine its performance in residue mutation detection and conformational stability prediction. The study selected biocatalcatalysts with industrial application value and clinically relevant therapeutic proteins as analytical samples, and established a standardized testing environment based on an internationally recognized protein database. The aim was to systematically compare the performance differences between this artificial intelligence platform and traditional computing tools and evaluate its transformation value in practical biotechnology applications.

The research first obtained the protein structure and mutation information from the Protein Database (PDB) and Protein Gym. During the data preparation stage, the PDB file of the protein and the mutation information in the dataset were organized into standardized inputs. Submit design tasks in batches and obtain prediction results through the correct input position of the ProtSSN software.

For the prediction results, after verifying the mutation prediction accuracy of the ProtSSN software through Spearman correlation analysis using statistical methods, this study analyzed the correlations between the secondary structure of the protein and the SASA aspect and the software prediction results respectively, thereby clarifying the influence of the secondary structure and SASA on the software prediction and the internal reasons.

The core significance of this study lies in establishing an economically efficient protein design evaluation system. By using open computing tools, it breaks through the limitation of high investment in traditional experiments and provides a feasible solution for the research environment with tight funds. Through the development of a unified data processing template and an intuitive result presentation method, a repeatable research framework suitable for the evaluation of intelligent protein design systems is constructed. The lightweight designed ProtSSN platform can complete operations under conventional computer configuration and the processing time of a single prediction task is controlled within 120 minutes, effectively lowering the technical threshold of protein modification research. This work not only confirms the practical value of artificial intelligence technology in the field of protein engineering, To more deeply reveal the intrinsic connection between amino acid sequences, spatial conformations and biological functions, and provide new technical support for accelerating the development process of industrial biocatalysts and therapeutic proteins.

Summary of the chapter I

1. Traditional protein design methods face significant technical bottlenecks. They rely on physical and chemical principles and experimental trial and error, and have core problems such as high computational complexity, long design cycle, and low success rate. Especially in complex structures (such as dynamic interaction networks) and functional design scenarios, they perform poorly and are difficult to meet the requirements of precise engineering.

2. Artificial intelligence technology has brought revolutionary breakthroughs to protein design. Through machine learning and deep learning algorithms, the hidden correlation rules among sequences, structures and functions can be efficiently analyzed, significantly improving the accuracy of mutation prediction (such as site synergy effect modeling) and computational efficiency (such as high-throughput design), promoting the transformation of protein engineering from empirical trial and error to a data-driven paradigm.

3. The ProtSSN software achieves a multi-dimensional performance leap through technological innovation. Based on a dual-modal collaborative pre-training framework (sequence semantics + three-dimensional structure topology) and a lightweight architecture (110 million parameters), the Spearman correlation coefficients for single-point and multi-point mutation prediction in the Protein Gym benchmark test reach 0.429 and 0.550 respectively. Go beyond the traditional model. This software innovatively integrates geometric coding with equi variable graph neural networks (EGNN), precisely capturing structural disturbances such as hydrogen bond breakage and hydrophobic core destruction, providing efficient solutions for complex scenarios (such as enzyme active site optimization).

4. This research has laid the theoretical and application foundation for the intelligent design of proteins. This study will analyze the negative correlation of SASA- prediction error, providing a quantitative basis for the optimization of residue stability, and the analysis of the relationship between secondary structure and prediction accuracy provides a direction for software optimization. The subsequent chapters will conduct an in-depth analysis of the structure-function correlation mechanism.

Chapter II

OBJECT, PURPOSE, AND METHODS OF THE STUDY

2.1 Research basis and theoretical framework

A very important part of protein function prediction lies in establishing a high-precision association model between amino acid sequence variations and biological functions. Traditional methods are often limited to single-dimensional features. Sequence models rely on evolutionary conservation but ignore the dynamics of three-dimensional structures. Although structural models can analyze local geometric constraints, they have difficulty capturing long-range synergy effects. This study conducted systematic tests based on the ProtSSN software. This software realizes the joint modeling of multi-scale features of proteins by integrating semantic encoding and geometric topological representation, providing a new theoretical paradigm for the prediction of mutation effects.

The theoretical innovation of ProtSSN stems from the systematic integration of the sequence-structure-function relationship of proteins. Its semantic encoding module is based on a large-scale protein sequence library and uses the evolutional-Scale Language Model (ESM-2) to extract the co-evolutionary patterns among residues. This module captures long-range dependencies across sequences through Masked Language Modeling (MLM), such as the co-conservation characteristics of catalytic sites or allosteric regulatory regions. The geometric coding module takes the Equivariant Graph Neural Network (EGNN) as the core. By constructing the residue space proximity graph to dynamically update the node coordinates and features, it accurately quantifies the perturbation effects of the three-dimensional microenvironment such as hydrogen bond networks and hydrophobic accumulation. The synergistic effect of the two is achieved through the gated attention mechanism, dynamically balancing the feature weights in different functional scenarios. For example, it enhances the geometric sensitivity of core hydrophobic interactions in thermal stability prediction and strengthens the semantic constraints of active pockets in catalytic activity evaluation.

This study first replicated some of the test experiments involved in the research on ProtSSN software by Tan Yang et al.[7] in 2023. The optimized k20_h512 model configuration was adopted strictly in accordance with the literature content. The evaluation dataset was selected from 217 proteins in the replacement module of the Protein GYM database. It covers a wide range of functional categories such as enzyme catalysis and molecular interaction. The predictive performance was quantified by Spearman's rank correlation coefficient (Spearman's ρ), which evaluated the accuracy of mutation ranking by analyzing the monotonicity association between the predicted values of the model and the experimentally determined DMS score. The above-mentioned experimental reproduction is used to verify the prediction accuracy of the ProtSSN software, and the results should be consistent with the content of the literature.

In order to further analyze the adaptability of the model to different mutation patterns, this study innovated on the basis of the original experiment and divided the test set into two categories according to the nature of mutations: the single-point mutation system (148 proteins) and the compound mutation system containing multi-point mutations (69 proteins).

The former focuses on the independent effect of local residue replacement, while the latter involves a synergistic effect. ProtSSN explicitly models the long-range associations among residues through the global message passing mechanism of graph neural networks. Theoretically, it can more accurately capture the synergistic perturbations of protein conformations caused by multi-point mutations. In addition, by integrating DSSP (Dictionary of Secondary Structure of Proteins) with statistics for combined analysis of secondary structures, SOMPA algorithm for predicting secondary structures (such as α-helix and β-fold), and solvent-accessible surface area (SASA) calculated by DSSP,

This study further explores the relationship between the secondary structure and the accuracy of software prediction, as well as the correlation between the software prediction error and the SASA value. The construction of this multi-level analysis framework provides systematic theoretical support for revealing the prediction mechanism of ProtSSN and its application in directed evolution.

2.2 Experimental Materials and Methods

2.2.1 Data Sources and Processing

Protein Gym is a standardized evaluation platform in the field of protein variation prediction. It integrates deep mutation scanning experimental data and clinical variation annotation information, providing a unified performance evaluation benchmark for different prediction algorithms. Its dataset adopts a multi-dimensional classification system including mutation forms (point mutation/frameshift mutation), data sources (DMS continuous values/clinical binary labels), and training modes (unsupervised/supervised learning). It covers key functional indicators such as enzyme activity, molecular recognition, and structural stability, and involves various biological source proteins such as humans, microorganisms, and viruses. Each data unit records in detail the mutation sites, experimental measurement values, and related meta-information. This platform supports two main evaluation modes: zero-shot prediction for assessing the generalization performance of models and supervised learning suitable for parameter optimization. It includes the performance comparison results of over 70 benchmark models to form an objective algorithm evaluation system. Its open-source feature is manifested in providing complete data acquisition guidance and automated analysis tools, and making up for the insufficiency of standardized evaluation in this field through a standardized evaluation framework. It has built an important bridge for the development of computational models, the design of experimental schemes and interdisciplinary research.

This experiment uses the Protein Gym dataset and selects the dataset of mutation effects of the replacement mutations as the benchmark. The dataset contains multiple mutation samples, and each sample provides data information such as mutation points, DMSscore, and DMSscore_bin.

To verify the prediction accuracy of ProtSSN, the ProtSSN software is run using Tencent Cloud’s high-performance GPU and six-core processor. Input the three-dimensional structure file of the protein (in PDB format), the FASTA file of amino acid sequence information, and the CSV file covering the mutation point information of the protein. Output the prediction results of the corresponding mutators and compare the results generated by running under different models.

The data processing steps are as follows:

(1) According to the prompts in the database, screen and retain the valid data for this study;

(2) Uniformly adjust the PDB file format to ensure that the atomic coordinates of all structure files are consistent with the residue numbers;

(3) The predicted values output by ProtSSN and the deep mutation scan experimental data were classified and organized by protein type to form a one-to-one corresponding data table for subsequent statistical analysis.

2.2.2 Experimental Verification Methods

1. Total processing of experimental data:

(1.1) Through the Protein Gym database, the mutation and structural information of a total of 217 proteins with their replacement mutation types were collected and the data were organized one by one according to the original file names.

(1.2) Input the organized files, including the PDB file and the CSV file containing the mutation information, into the corresponding folder of the ProtSSN software, run the shell script, obtain the prediction result, and output the result in CSV file form.

By calculating the Spearman's correlation coefficient (Spearman's ρ), the consistency between the predicted data of ProtSSN and the experimental data of deep mutation scanning was analyzed.

Correlation analysis: Analyze Spearman's ρ, and the formula is:

$$\rho = 1 - [(6\,\Sigma d^2) / [n \bullet (n^2 - 1)]] \quad (2.1)$$

Among them, di is the difference between the predicted ranking and the experimental ranking of the i-th mutant, and n is the total number of samples.

2. Data analysis and processing methods for the experimental reproduction part:

(2.1) The calculation process was constructed based on Python 3.9. The target variables DMSscore and ProtSSN_k20_h512 (predicted value of protein mutation) were extracted from 217 CSV files using the Pandas (v1.3.5) library. The remaining fields were excluded because they were irrelevant to the research objectives. After reading the data through pd. read_csv, the dropna method is used to clear the rows containing missing values (NaN), and it is verified that the amount of valid data in each file is ≥3 to ensure statistical reliability (all files meet the conditions and no data is skipped).

(2.2) The Spearman's correlation coefficient (Spearman's ρ) is calculated through the Spearman R function of SciPy (v1.7.3) (Formula 2-1). The histograms and kernel density estimation (KDE) curves were plotted using Seaborn (v0.11.2) and Matplotlib (v3.5.0). The mean line of the red dotted line and the $\mu \pm \sigma$ interval (0.304-0.630) were marked, and the PDF vector diagram was output at a resolution of 300 dpi.

3. Partial data analysis and processing methods for single-point and multi-point mutations:

(3.1) The analysis process is constructed based on Python 3.9. Firstly, two columns of data, DMSscore and ProtSSN_k20_h512, are extracted from the CSV file. The 'pd. read_csv' of the Pandas library is used for reading, and the rows containing missing values are cleared through the 'dropna' method. Meanwhile, filter the datasets with an effective sample size less than 3. Subsequently, the Spearman rank correlation coefficients of single-point mutations (148 datasets) and mixed mutations (69 datasets) were calculated using the 'spearman' function of the SciPy library to reduce outlier interference by a non-parametric method, and the mean (μ), median (M), and standard deviation (σ) of the correlation coefficients were calculated through NumPy.

(3.2) To present the distribution characteristics, the histogram and kernel density estimation (KDE) curve were plotted using Seaborn and Matplotlib. The mean line was marked with a red dotted line in the figure, and the $\mu \pm \sigma$ interval range was marked. The output was a PDF vector diagram of 300 dpi.

2.2.3 Methods for processing experimental results

This chapter of the study first reproduced some experiments related to ProtSSN literature. The k20_h512 model of ProtSSN was used to comprehensively analyze the 217-substitution mutation information contained in the Protein Gym database, and the average Spearman correlation coefficient was obtained as 0.46726940061418487. Then, on the basis of the previous work, innovations were made and analyses were carried out according to the mutation types (single-point mutation, multi-point mutation). Figure 2.1 shows the Spearman correlation analysis graph of the predicted values obtained by the software after analyzing the mutation information of 217 proteins and the data from the deep mutation scanning experiment.

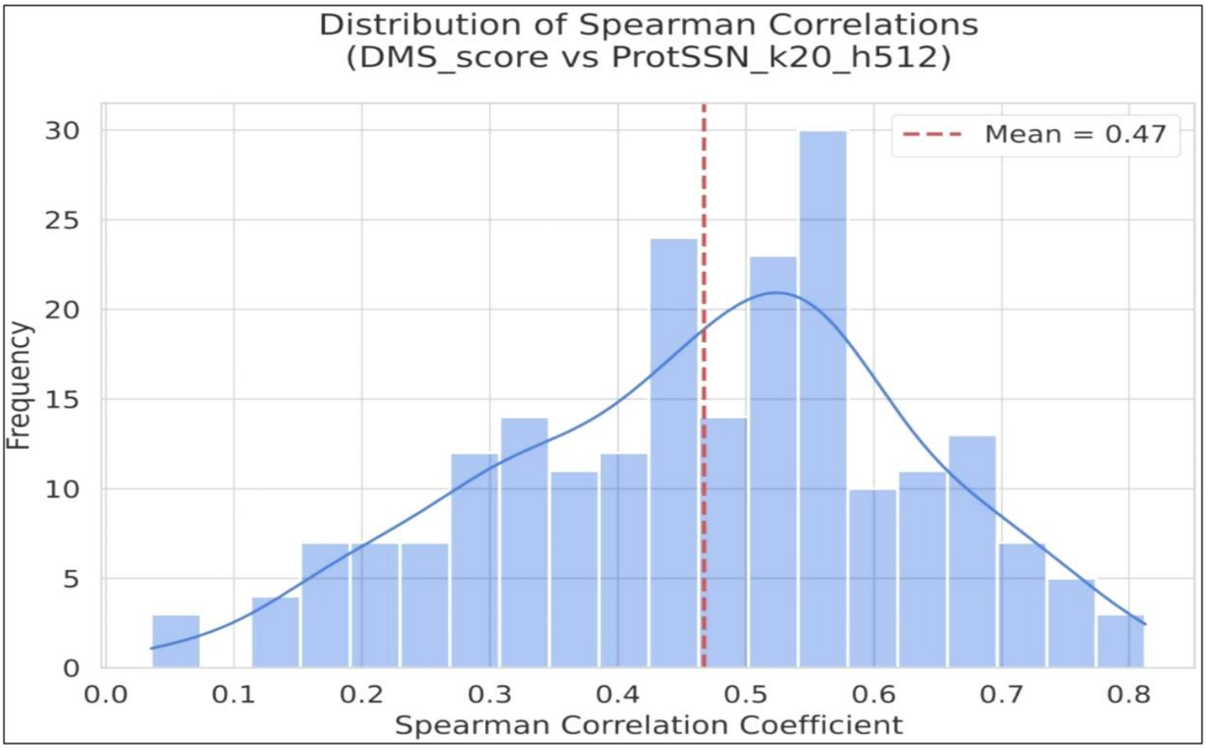

Figure 2.1 Spearman correlation histogram of predicted values with deep mutation scan experimental data

In this study, the protein information of the Protein Gym database was grouped and tested according to mutation types (single-point mutations and multi-point mutations). A total of 148 proteins containing only single-point mutations in the mutation file were selected, and a total of 69 proteins containing both single-point mutations and multi-point mutations were selected. The k20_h512 model under the ProtSSN framework was adopted to predict the mutation effects of the two types of proteins respectively. The consistency between the prediction results and the experimental data was quantified through the Spearman correlation coefficient, and then the performance of the model in different mutation scenarios was analyzed.

Figure 2.2 shows the Spearman correlation analysis graph between the predicted values and the actual values of 148 file software with only single-point mutations.

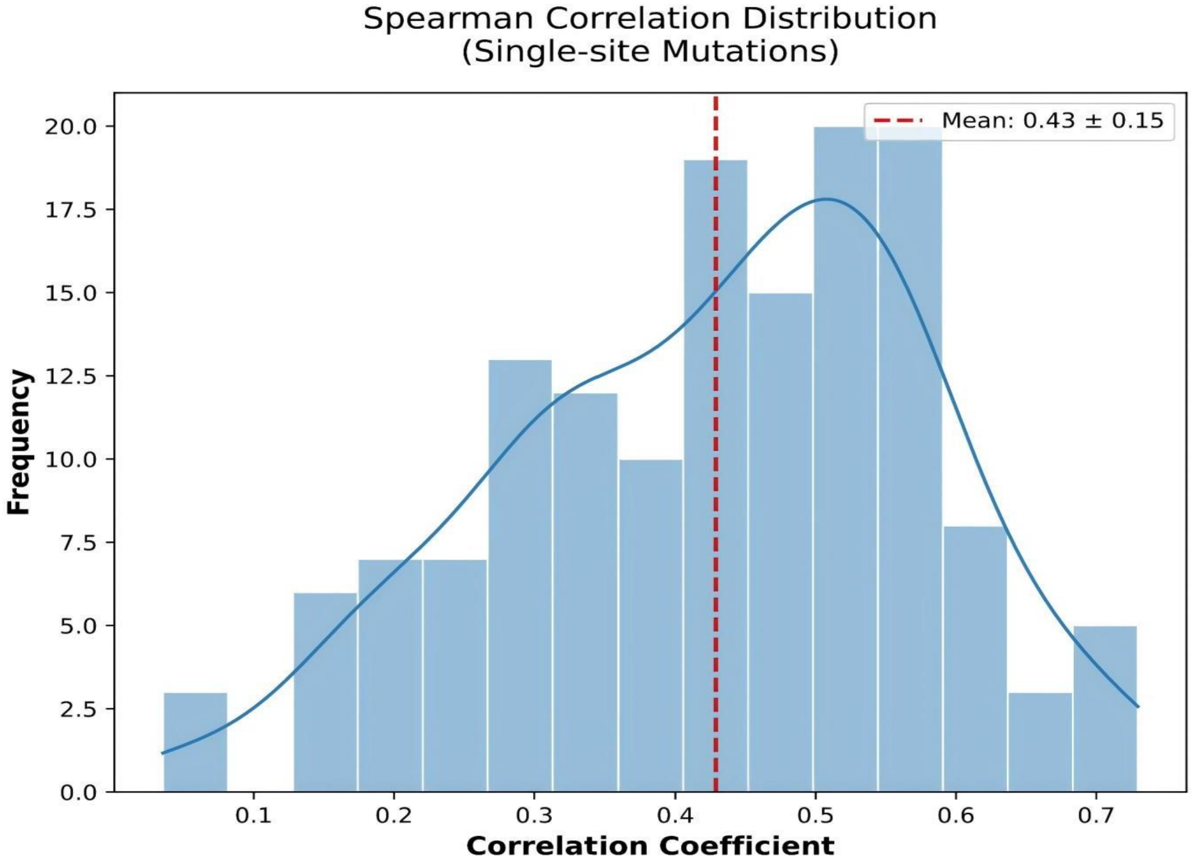

Figure 2.2 Spearman correlation histogram of single-point mutation

It can be seen from the graph that the average Spearman correlation coefficient of the model is 0.429, and the correlation coefficient is concentrated in the range of 0.28-0.58, with the peak located in the range of 0.4-0.5. It indicates that it has moderate accuracy in predicting the mutation effect of a single amino acid site and can effectively capture the monotonic association between the mutation site and the function or stability of the protein. This single-point performance stems from the model's deep encoding of the semantic information of protein sequences - extracting the long-range dependencies of amino acid sequences through pre-trained language models, combined with the geometric characteristics of the local microenvironment, to analyze the perturbation effect of single-point mutations on the interaction between active centers and domains.

Figure 2.3 shows the Spearman correlation analysis of 69 files that contain both single-point and multi-point mutations. It can be seen from the figure that the Spearman correlation coefficient is 0.550, and the mixed mutations are distributed from 0.39 to 0.60, showing a right-biased trend, which reflects the superiority of ProtSSN in dealing with complex mutations.

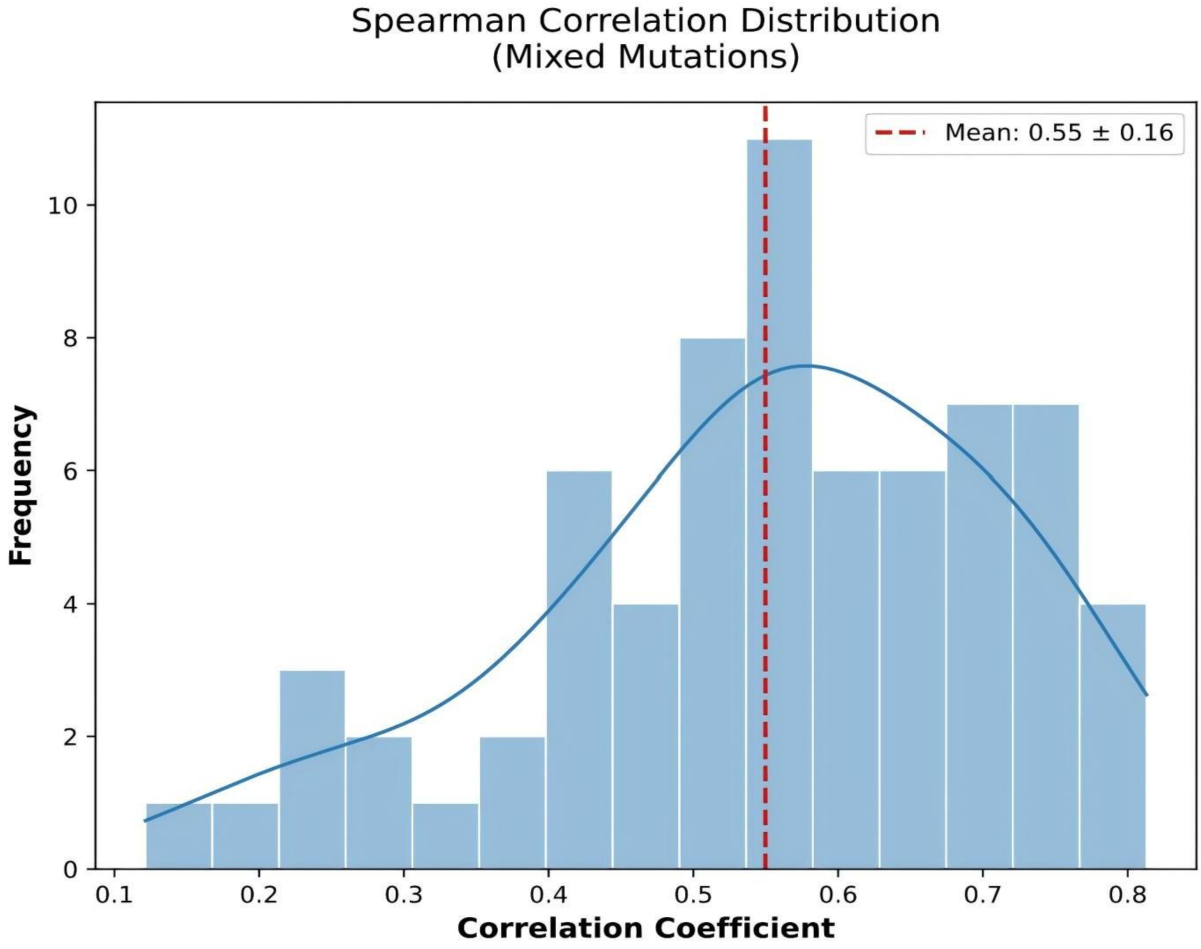

Figure 2.3 Contains the Spearman correlation histograms of single-point and multi-point mutations

This phenomenon reflects the model's ability to capture the synergistic effects of multiple mutation sites. That is, by modeling the spatial dependence between residues through the rotational translation isotropic graph neural network (EGNN), the nonlinear interactions between amino acid sites, such as domain cooperative folding and hydrogen bond network reconstruction, and other higher-order effects, can be effectively captured. For example, in the prediction of thermal stability-related mutations, multi-point mutations often enhance stability by optimizing the hydrophobic interactions or salt bridge networks within proteins, and the model's analytical ability for such synergistic effects directly improves the prediction accuracy in complex mutation scenarios.

The dual-modal coding framework of ProtSSN integrates sequence semantic embedding and three-dimensional structure topological coding, which is the core mechanism of its performance advantage. At the sequence level, the deep semantic representations extracted by pre-trained language models (such as ESM-2) can capture the grammatical rules and functional associations of amino acid sequences. At the structural level, graph neural networks supplement spatial position information for the geometric modeling of the protein microenvironment. Especially in complex mutations, structural information is crucial for predicting stability (such as Δ Tm, Δ G) and functional synergies. Furthermore, the model's zero-shot prediction ability enables it to demonstrate generalization in unseen proteins and mutation types, allowing it to handle common "cold start" issues in wet experiments, such as the mutation design of novel enzymes or viral proteins, without the need for additional labeled data.

From the perspective of practical application value, the stable performance of the model in single-point mutations and its outstanding performance in complex mutations provide efficient computational guidance for directed evolution experiments. The Gaussian Pearson coefficient indicates that it can effectively narrow the screening range of mutants and reduce the cost of experimental trial and error, especially suitable for scenarios such as optimizing binding affinity and improving the thermal stability of industrial enzymes in antibody engineering.

The performance of ProtSSN in single-point and multi-point mutations further validates the necessity of sequence-structure co-coding. Its ability to capture the synergistic effects of complex mutations provides key technical support for artificial intelligence-driven protein engineering. Future research can further optimize the model's ability to characterize the dynamic interaction of long-distance domains, and expand the training data by combining a wider range of species and structural types, thereby enhancing its application potential in special scenarios such as the prediction of viral protein variations.

Summary of chapter II

1. The performance verification of ProtSSN is based on the Protein Gym dataset and structural feature analysis. Its Spearman coefficient (0.429/0.550) for single-point/multi-point mutation prediction is significantly better than that of the traditional model. The EGNN network accurately models structural disturbances such as hydrogen bond breakage, and the computational efficiency is significantly improved.

2. The experimental methods and verification system adopt strict data processing techniques and verification methods to ensure the reliability and repeatability of performance evaluation.

3. The experimental results show that ProtSSN performs significantly better than traditional methods in protein structure prediction and functional design, and quantitative analysis clarifies its technical advantages.

4. The computational efficiency and accuracy of the software have been verified through experiments, supporting its practical application in protein engineering, such as reducing the cost of experimental trial and error and accelerating the mutant screening process.

Chapter III EXPERIMENTAL PART

3.1 Overview of Protein Secondary Structure

The secondary structure of a protein is a local spatial conformation formed by the backbone of the polypeptide chain through hydrogen bonding. It is a key level connecting the primary sequence with the tertiary three-dimensional structure, and its conformational characteristics directly affect the stability, functional activity, and response pattern to mutations of the protein. The structure at this level mainly includes α -helix, extended strand (the basic unit that constitutes β-folding) and random coil. Each conformation is formed through specific hydrogen bond patterns and amino acid sequence tendencies. And it plays a core role in the folding, functional realization and evolutionary adaptation of proteins[10].

The α -helix is one of the most common regular secondary structures, formed by the right-handed helix of the main chain atoms, and a hydrogen bond is formed between the carbonyl oxygen (C=O) of the i-th residue and the amino hydrogen (N-H) of the i-4th residue, creating a tight helical conformation. Its stability depends on the accumulation of hydrophobic residues (such as leucine and isoleucine) in the core region, as well as the spatial compatibility of side chain groups, often forming the domain core or transmembrane region of proteins, providing rigid support for functional sites. The extended chains form β-folded sheets through hydrogen bonds of the main chains of multiple peptide chains, which are divided into two types: parallel and antiparallel. Their stability stems from the alternating arrangement of hydrophobic groups on the side chains and van der Waals interactions between the sheets, constituting an important component of the rigid framework of proteins and commonly found in regions such as immunoglobulin domains and catalytic cores of enzymes. Irregular coiling is an irregular conformation lacking periodic hydrogen bonds, which endowing proteins with flexibility. It is mostly distributed in domain connection regions, ligand binding pockets, or active centers of enzymes, mediating protein-protein interactions or functional regulation through dynamic conformational changes.

The biological functions of secondary structures are closely related to their conformational characteristics. Regular structures (α -helical, extended chains) form the structural basis for proteins to resist thermal denaturation and chemical disturbances through stable hydrogen bond networks and hydrophobic interactions. Their integrity directly affects the melting point temperature (Tm) and folding free energy (Δ G) of proteins. For example, the hydrophobic packing of the α-helical core and the β-folded sheets of the extended chains provide structural rigidity for proteins, while the flexibility of irregular curling allows conformational changes at functional sites, such as induced fit when enzymes bind to substrates or allosteric activation of signal proteins.

This hierarchical characteristic of structure-function makes the secondary structure a key sensitive area for mutation effects. Mutations occurring in regular structures are prone to disrupt stable hydrogen bonds or hydrophobic interactions, resulting in a decrease in thermodynamic stability or functional abnormalities. However, mutation effects in irregular curling are more related to local flexible changes.

In protein engineering and the prediction of mutation effects, the precise analysis of secondary structures is an important prerequisite for understanding the influence of mutations. The ordered conformations of α-helical and extended chains provide quantifiable structural features for the computational model, such as hydrogen bond density and hydrophobic contact area, while the dynamic characteristics of irregular curling pose challenges to the flexible conformation modeling of the model.

This chapter focuses on the basic concepts and conformational characteristics of secondary structures. On this basis, combined with the prediction data of the ProtSSN model in the previous study, it deeply analyzes the correlation between the distribution of secondary structures and the accuracy of mutation prediction, reveals the differential mechanism between regular structures and irregular curling in mutation responses, and provides theoretical support for the design of proteins targeting secondary structures.

3.2 Research Methods

1. Research method of secondary structure correlation analysis based on:

(1.1) Collect protein data from the Protein GYM database. Based on the CSV file results output in Chapter One and the Secondary structures of Proteins analyzed through DSSP (Dictionary of Secondary Structure of Proteins), correspond them one by one according to the location information. Integrated into one dataset (combined_data.csv file), this file only retains single-point mutations among 217 proteins, and multi-point mutations are excluded. Table 3.1 shows the secondary structure identifiers and structural characteristics contained in DSSP, which are used for the subsequent result analysis.

Table 3.1 **List of DSSP Secondary Structure Type Identifiers**

Secondary structure type	Identifier	Structural characteristics
α -helix	H	Common helical structures maintain stable regular hydrogen bonds between atoms in the main chain
3 ▪ σ−helix	G	Similar to the α-helix, every three amino acid residues form one helix, and ten atoms participate in the hydrogen bond ring
π-helix	I	Relatively rare, every five amino acid residues form one helix
Extension chain	E	For a– part of the β-folded structure, the peptide chain extends in a sheet-like form, and the hydrogen bonds between the chains are maintained
β-bridge	B	The individual bridge structure in the β-fold sheet represents the hydrogen bond in the β-fold
Turn the corner	T	Connect different secondary structural units to change the direction of the peptide chain
Bend	S	High curvature, not dependent on hydrogen bond classification, reflecting the local conformation of the peptide chain
Random curling	-	Loose peptide chain regions with irregular secondary structures

(1.2) The experiment was based on the integrated dataset (combined_data.csv), including key fields such as mutation sites, secondary structure types, predicted scores, and experimental measurement values. First, perform data cleaning to eliminate the samples with missing Secondary_Structure, Prediction_Score, and DMS_Score to ensure data integrity. Calculate the absolute prediction error (Prediction_Error = |Prediction_Score - DMS_Score|) for valid samples, and set are a sonable range threshold for DMS_Score ∈ [0,1] to eliminate the interference of experimental measurement outliers.

(1.3) Based on the Secondary_Structure field, the data is divided into structural subsets such as H (α -helix), E (β -collapse), T (rotation Angle), etc. Non-parametric statistical tests are performed on each subset, that is, the monotony correlation between the score and the experimental value is quantitatively predicted through Spearman's rank correlation coefficient (ρ). The *p*-value was obtained by using the two-sided hypothesis test to evaluate the statistical significance. To ensure the reliability of the analysis, only the secondary structure categories with a sample size of ≥5 is retained.

(1.4) Reveal the correlation pattern between structural features and prediction performance by comparing the charts: Figure 3.1 is a bar chart arranged in descending order of ρvalues. The viridis color level mapping correlation coefficient intensity is used, and the positive and negative correlations are distinguished by the dotted line with y=0. Figure 3.2 is a box plot of synchronous sorting. The Set2 color system is used to display the distribution of prediction errors, and outliers are hidden to highlight the characteristics of the main data. Both graphs have standardized coordinate axis ranges (ρ ∈ [-1,1], Prediction_Error ∈ [0,1]), and visual rendering is achieved through Matplotlib 3.6 to ensure the rigor and interpretability of the result presentation.

2. Research method for Secondary Structure correlation Analysis Based on SOMPA:

(2.1) Mutation data of 217 proteins collected from the Protein GYM database. According to the results of Chapter Two, proteins with extremely high and extremely low correlation coefficients between the predicted values of ProtSSN and the Spearman coefficient of DMSscore were selected, and the mutation sites and protein structures were recorded.

(2.2) Use the SOPMA tool to predict the secondary structure of the protein amino acid sequence, and determine the distribution range and proportion of -helices, extended chains and random coiling.

(2.3) Compare and analyze the SOPMA and ProtSSN results (the prediction results required for this chapter have been obtained in the experiments of Chapter Two), and analyze the correlation between the prediction accuracy and the secondary structure through the proportion of different types of secondary structures.

(2.4) Verification and Discussion: Discuss the correlation mechanism in combination with structural biology knowledge and draw conclusions.

3.3 Results and Analysis

3.3.1 Correlation Analysis of Secondary Structure Based on DSSP

This study adopts the above-mentioned methods Obtain the "Prediction-actual Correlation by Secondary Structure" (Figure 3.1) and the "Prediction Error Distribution by Secondary The two charts "Structure" (Figure 3.2) analyzed the predictive performance of the model under different types of protein secondary structures.

In Figure 3.1, we calculated the Spearman correlation coefficient (Spearman ρ) between the predicted values and the actual values under different protein secondary structure types (B, H, -, G, T, S, I, E). The data show that the correlation coefficients of each secondary structure type are generally close to 0. Compared with the ideal strong correlation situation, the degree of correlation between the predicted values of the model and the actual values in this study is relatively weak. It is speculated that this might be due to the retention of only single-point mutations, insufficient data volume, and the possible existence of special mutations.

Among various secondary structures, the correlation coefficients of β-bridge (B), α-helix (H), and random curl (-) are relatively high, indicating that in these two structures, the consistency between model predictions and actual situations is relatively better than in other structures. To a certain extent, this reflects that these two structures are relatively easier to be captured by the model in terms of mutation laws.

The structures of 3_{10}-helix (G), rotation Angle (T), curvature (S), π-helix (I), and extension chain (E) have lower correlation coefficients and approach 0, indicating that under these secondary structure types, the accuracy of predicting mutations by the model has an extremely low correlation with the actual situation. It might be due to the complexity of these structures or the characteristics of the data that the model has a relatively high difficulty in learning such mutation patterns.

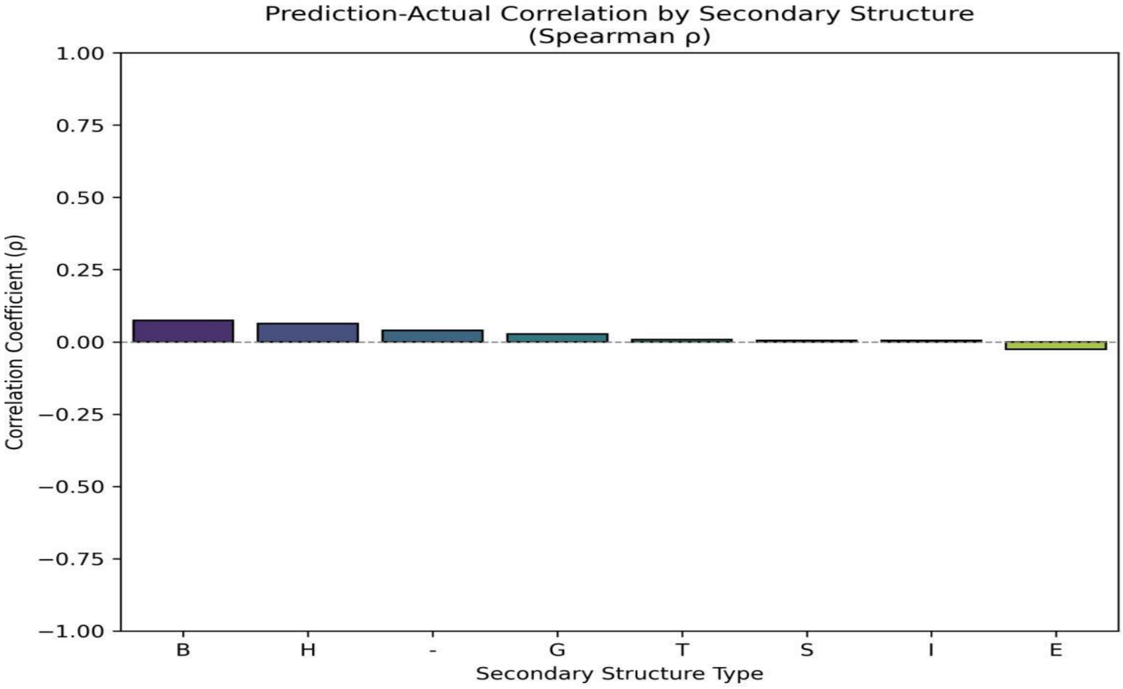

Figure 3.1 Correlation analysis of predicted and actual values under the secondary structure

Figure 3.2 presents the distribution of prediction errors under each secondary structure type in the form of a box plot. The median prediction errors of different secondary structures are mostly concentrated between 10 and 15. From the box length and whisk length of the box plot, there are differences in the degree of data dispersion among different secondary structure types. The boxes of β-bridge (B), α-helix (H), and random curl (-) are relatively short, and the whisk lengths are also relatively short, indicating that the prediction error data under this structure type is relatively concentrated, and the stability of the model's prediction error is better. The boxes and whiskers of other structures are relatively long, indicating that the prediction error data is more dispersed. The fluctuation of the prediction error of the model under this structural type is large, and the stability is poor. Although there are certain differences in the degree of data dispersion among different structural types, the overall fluctuation range is relatively limited. Compared with the more ideal error distribution in other related studies, there is still room for optimization in the error level in this study.

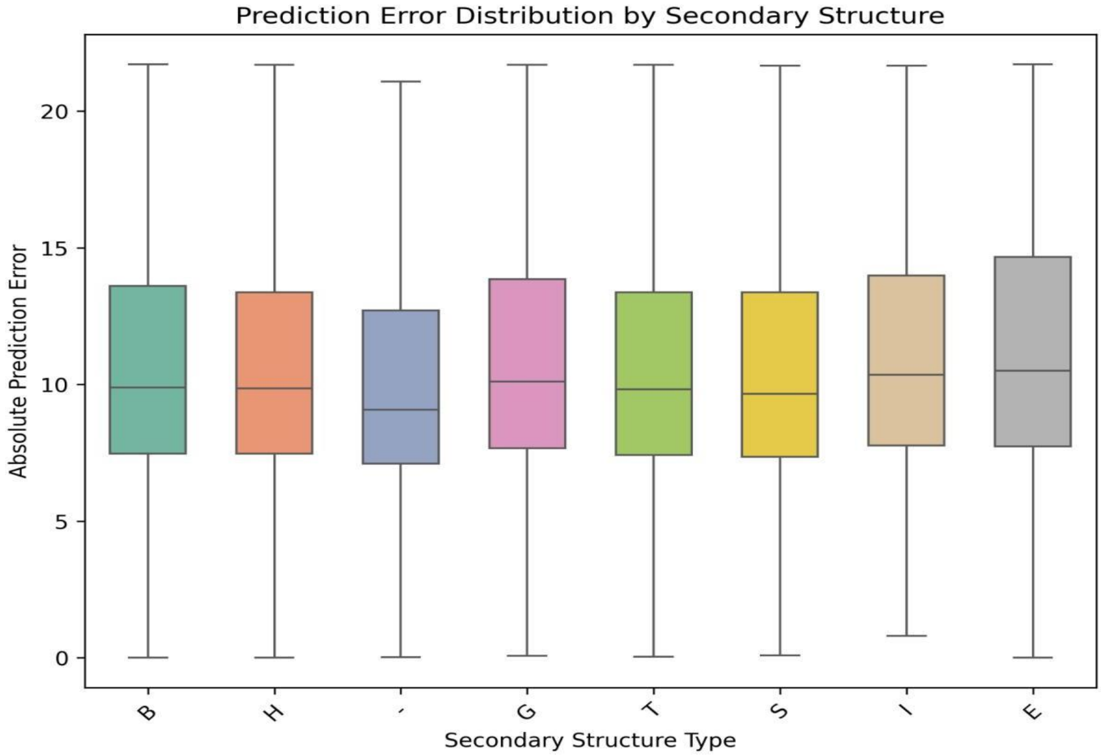

Figure 3.2 Distribution of prediction errors under -2 secondary structure

3.3.2 Correlation Analysis of Secondary Structure Based on SOMPA

To make up for the deficiencies of the above-mentioned experiments, this part of the study focused on the secondary structure distribution analysis of the top 4 high-accuracy predicted proteins and the bottom 4 low-accuracy predicted proteins in the Spearman correlation coefficient of the ProtSSN model in the Protein Gym benchmark through the SOPMA tool.

Focus on three conformations: α-helix, extension chain (β-folded basic conformational unit), and random curl, and analyze the quantitative relationship between the predictive performance of ProtSSN and the characteristics of the secondary structure. The proportion of secondary structures of each protein in the following text is directly analyzed by SOMPA software. The Spearman correlation coefficient reflected in the following study is from the relevant experimental results in Chapter II.

(1) *Protein part with low Spearman correlation coefficient*

In the secondary structure of A0A1I9GEU1_NEIME_Kennouche_2019 (Figure 3.3), the α -helix accounts for 28.57% (46 amino acids), the extended chain accounts for 11.80% (19 amino acids), and the irregular coiling accounts for 59.63% (96 amino acids).

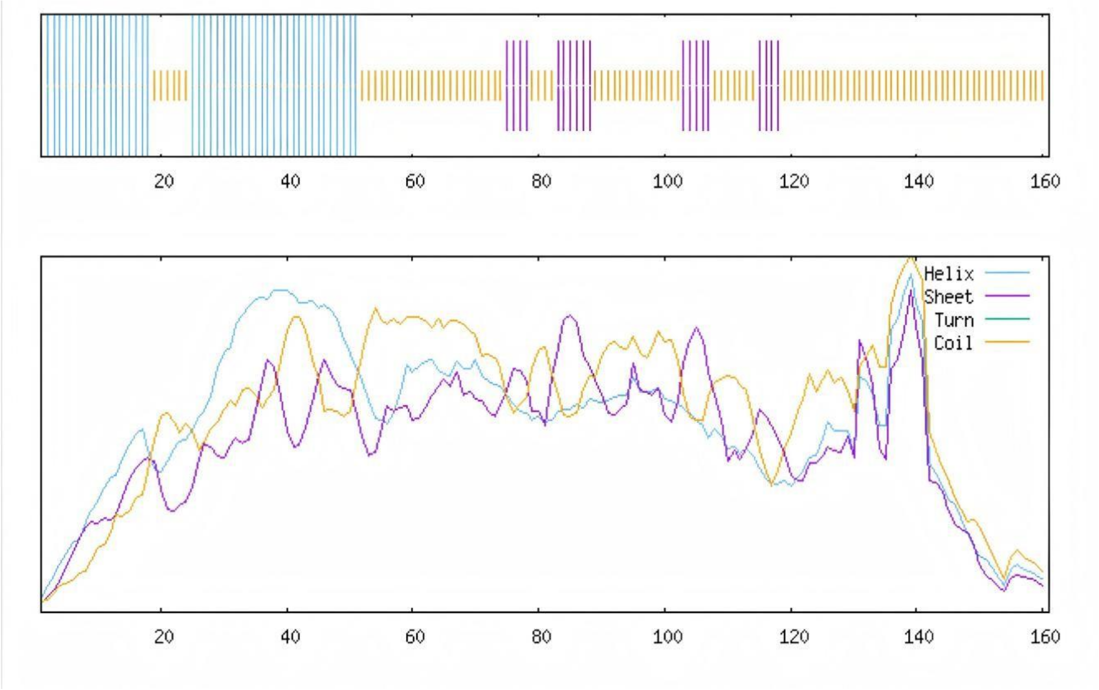

Figure 3.3 A0A1I9GEU1_NEIME_Kennouche_2019 ProtSSN mutation predicted Spearman rank correlation coefficient 0.03539592125272095

In the secondary structure of I6TAH8_I68A0_Doud_2015 (Figure 3.4), the α-helix accounts for 44.78% (223 amino acids), the extended chain accounts for 9.04% (45 amino acids), and the irregular curl accounts for 46.18% (230 amino acids).

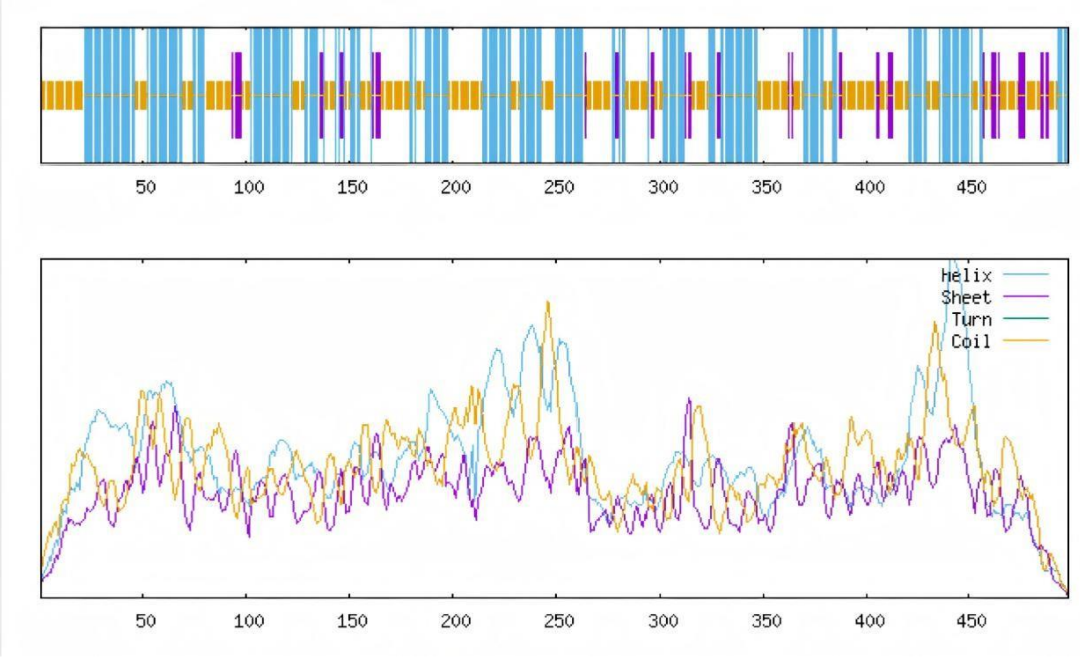

Figure 3.4 I6TAH8_I68A0_Doud_2015 ProtSSN mutation predicted Spearman rank correlation coefficient: 0.1470226068222732

In the secondary structure of KCNE1_HUMAN_Muhammad_2023_expression (Figure 3.5), the α-helix accounts for 37.21% (48 amino acids), the extended chain accounts for 8.53% (11 amino acids), and the irregular curl accounts for 54.26% (70 amino acids).

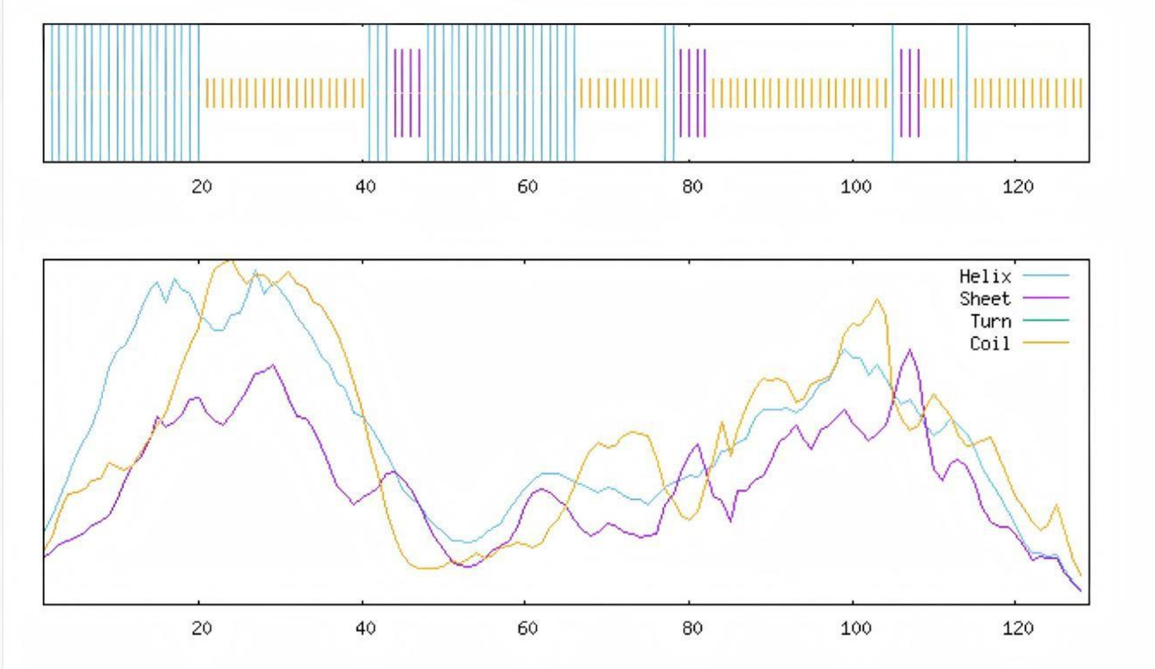

Figure 3.5 Spearman rank correlation coefficient for prediction of ProtSSN mutation in KCNE1_HUMAN_Muhammad_2023_expression: 0.070158095359

In the secondary structure of TADBP_HUMAN_Bolognesi_2019 (Figure 3.6), the α-helix accounted for 12.56% (52 amino acids), the extended chain accounted for 15.9% (66 amino acids), and the irregular coiled-up accounted for 71.50% (296 amino acids).

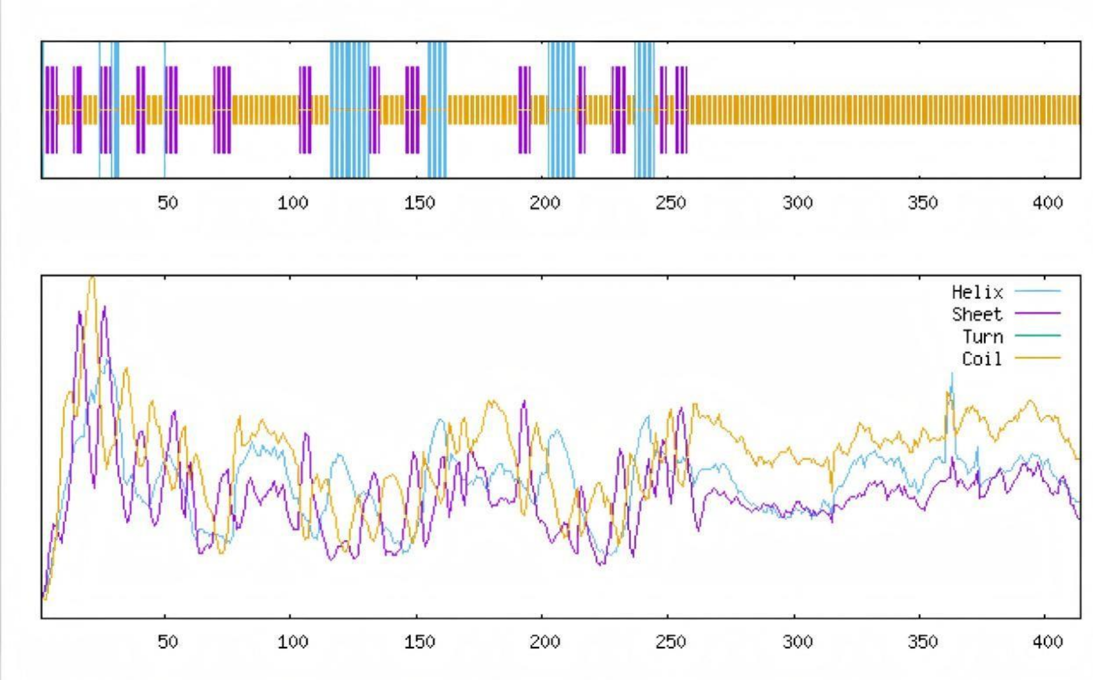

Figure 3.6 Spearman rank correlation coefficient for prediction of TADBP_HUMAN_Bolognesi_2019 ProtSSN mutation: 0.054450226965378754

(2) *Gaussian Pilman correlation coefficient Protein part:*

In the secondary structure of A4GRB6_PSEAI_Chen_2020 (Figure 3.7), the α-helix accounted for 31.58% (84 amino acids), the extended chain accounted for 20.68% (55 amino acids), and the irregular coiling accounted for 47.74% (127 amino acids).

In the secondary structure of BLAT_ECOLX_Jacquier_2013 (Figure 3.8), the α-helix accounts for 40.56% (116 amino acids), the extended chain accounts for 13.64% (39 amino acids), and the irregular curl accounts for 45.80% (131 amino acids).

In the secondary structure of HCP_LAMBD_Tsuboyama_2023_2L6Q (Figure 3.9), the α-helix accounts for 58.18% (32 amino acids), the extended chain accounts for 16.36% (9 amino acids), and the irregular coiling account for 25.45% (14 amino acids).

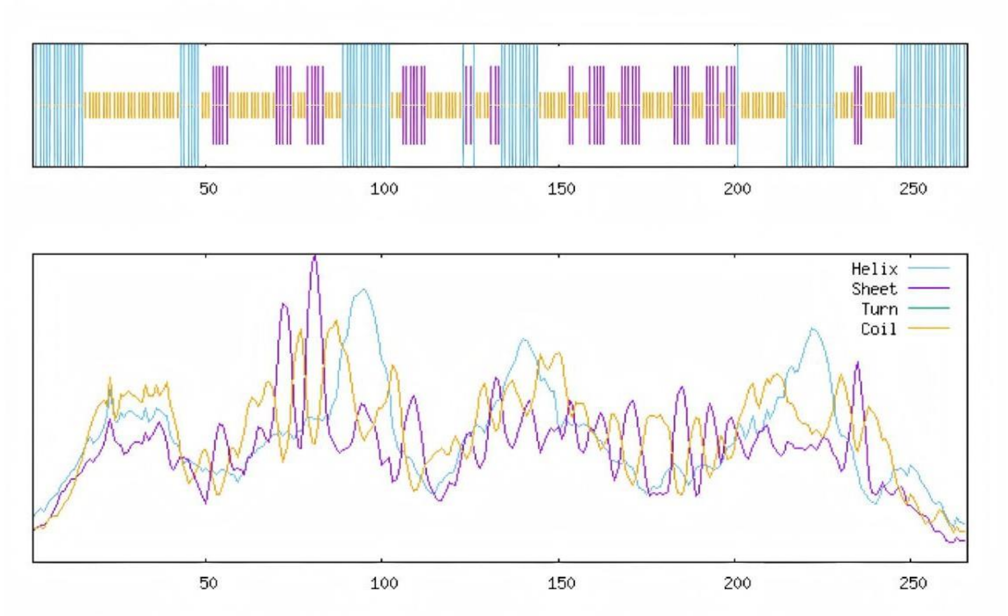

Figure 3.7 A4GRB6_PSEAI_Chen_2020 ProtSSN mutation predicted Spearman rank correlation coefficient: 0.7179293092233868

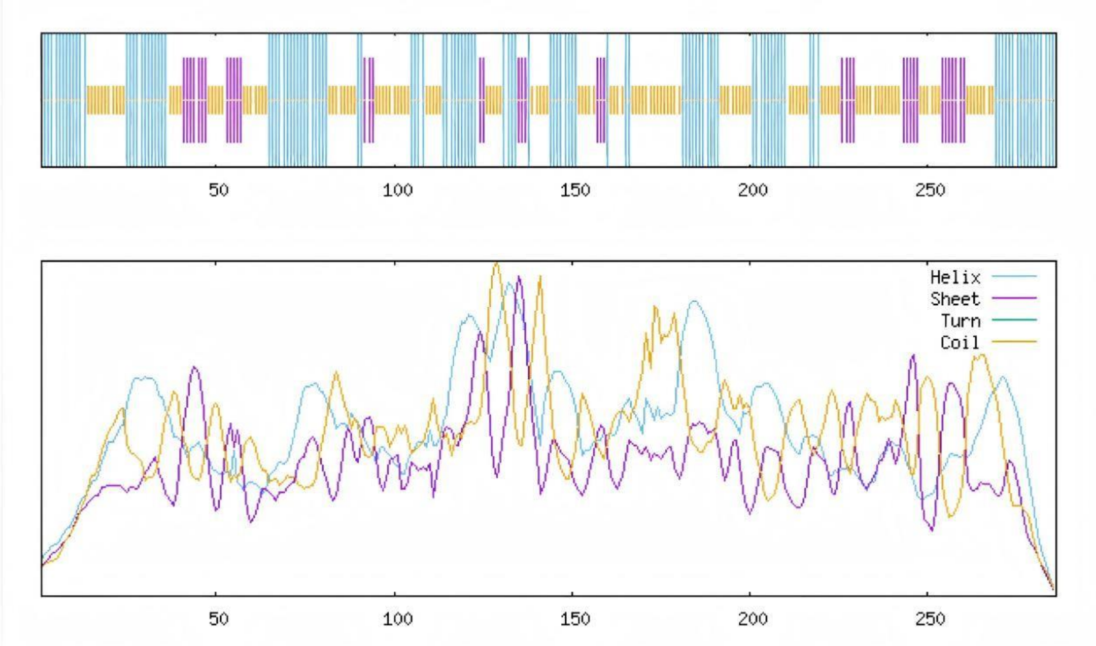

Figure 3.8 BLAT_ECOLX_Firnberg_2014 ProtSSN mutation predicted Spearman rank correlation coefficient: 0.7295504153600589

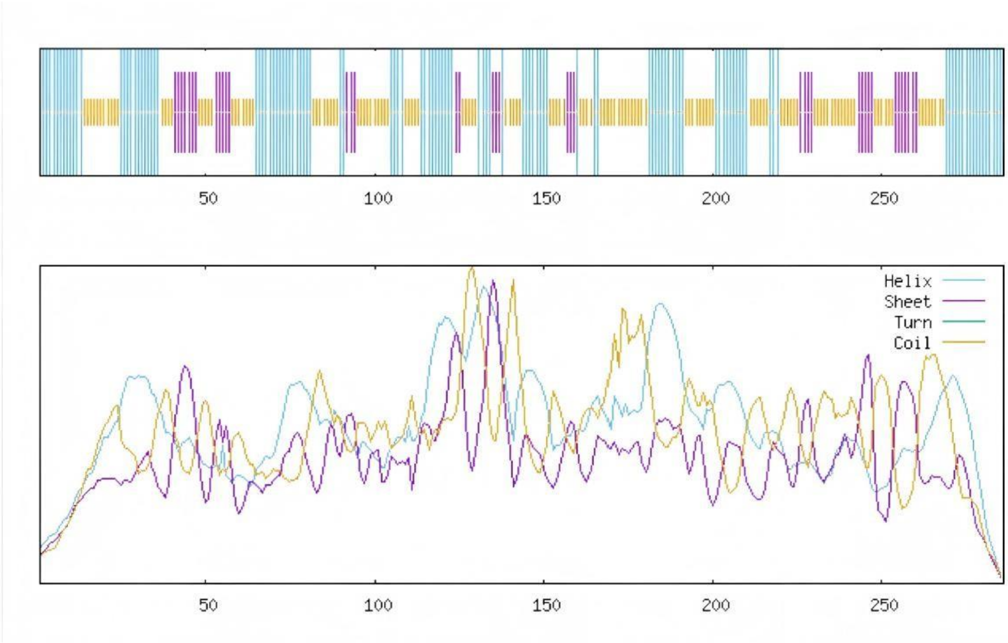

Figure 3.9 Spearman rank correlation coefficient for mutation prediction of BLAT_ECOLX_Jacquier_ 2013 ProtSSN: 0.6916445490953721

In the secondary structure of HCP_LAMBD_Tsuboyama_2023_2L6Q (Figure 3.10), the α-helix accounts for 58.18% (32 amino acids), the extended chain accounts for 16.36% (9 amino acids), and the irregular coiling account for 25.45% (14 amino acids).

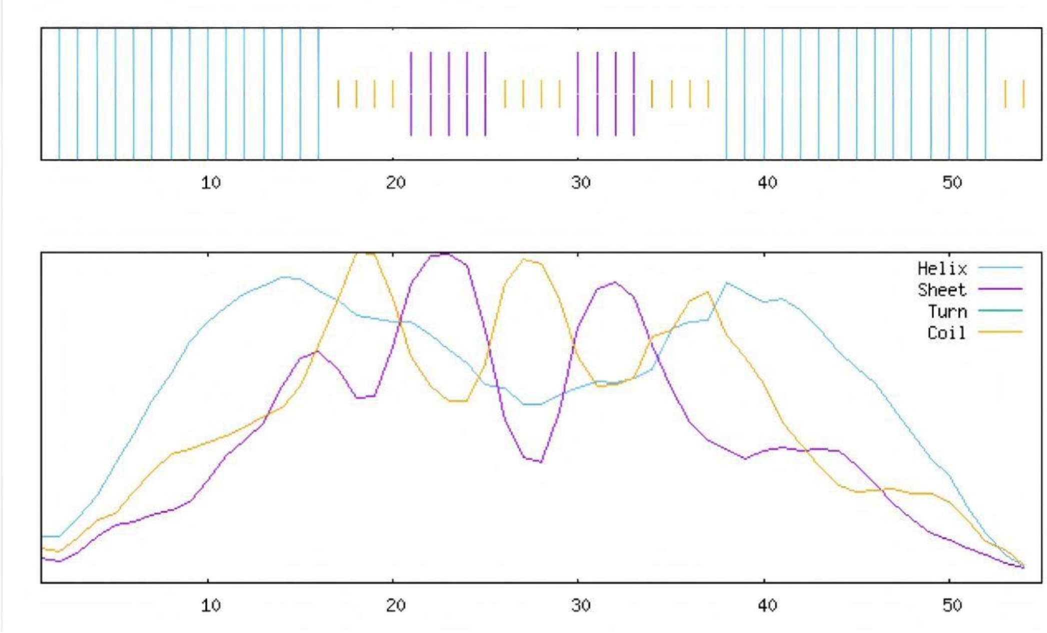

Figure 3.10 Spearman rank correlation coefficient for mutation prediction of HCP_LAMBD_Tsuboya ma_2023_2L6Q ProtSSN: 0.6867574588819246

Conformational prediction based on SOPMA shows that the highly accurate predicted mutation sites with a Spearman rank correlation coefficient of 0.65 or above are mostly distributed in the α -helix and extension chain regions, which is higher than the irregular curl. It is speculated that this distribution difference is correlated with the stability characteristics of the regular structure: The α -helix forms a tight hydrophobic core through the main chain i and the hydrogen bond at the i+4 position, while the extension chain maintains rigidity by the alternating arrangement of hydrogen bonds between layers and hydrophobic side chains. The integrity of both is highly sensitive to mutations. ProtSSN extracts the helix/extension chain formation tendency in the sequence through the pre-trained language model (ESM-2), and combines the equivariant graph neural network (EGNN) to encode the geometric features of the residue microenvironment, achieving precise analysis of regular structural mutations.

Summary of this part of the chapter III

This chapter studies and analyzes the deep correlation between the characteristics of secondary structures and the predictive performance of the ProtSSN model. The results show that the Spearman correlation between the predicted values and experimental values of regular structures such as β-bridge (B) and α-helix (H) is significantly higher than that of complex conformational regions (such as π-helix, rotation Angle, etc.). The speculative model has a stronger feature extraction ability for the geometric constraints and thermodynamic stability of regular structures. This advantage may stem from the coupling effect of sequence tendencies, spatial steric hindrance and energy changes within the regular structure, providing modellable physicochemical laws for deep learning models.

The analysis of prediction errors further supports the above conclusion. The error distribution of the regular structure is concentrated (interquartile range of the box <5), while the prediction stability of the complex conformation is poor (interquartile range >8). Through the secondary structure analysis of high/low correlation proteins by SOPMA, it was found that the proportion of regular structures (α-helical and extended chains) of proteins with high predictive performance was higher than that of the low performance group, verifying the potential mechanism of model performance differences from the perspective of structural biology. Meanwhile, this study also has certain limitations. The SOMPA secondary structure prediction software has certain errors, and the secondary structure correlation analysis based on SOMPA only takes the proteins with extremely high and extremely low Spearman correlations for analysis, which has certain limitations. In the future, the relevant technologies and theoretical basis will continue to be enriched, with the expectation of conducting more in-depth research on the secondary structure.

3.3.3 Overview of Amino Acid SASA

In the field of protein structure and function research, the Solvent Accessible Surface Area (SASA) is a key parameter reflecting the solvent exposure state of amino acid residues[11]. Its definition is to simulate the surface area of amino acid residues that a solvent molecule probe can come into contact with when rolling on the three-dimensional structure surface of a protein through theoretical calculation. SASA not only visually presents the burial depth of amino acid residues in the spatial structure of proteins, but also profoundly affects the folding dynamics, structural stability and biological functions of proteins.

From the analysis of chemical properties, the hydrophilic side chains of polar amino acids tend to be exposed on the protein surface to enhance the polar interaction with the solvent. The hydrophobic side chains of non-polar amino acids are usually buried inside the protein, forming a hydrophobic core that maintains the stability of the three-dimensional structure of the protein. This amino acid residue distribution law based on SASA constitutes the thermodynamic basis for the assembly of the secondary structure and the stability of the tertiary structure of proteins.

3.3.3.1 Research Methods

3.3.3.1.1 Acquisition of SASA data by DSSP and the data organization process

This chapter of the study extracted 148 deep mutation scan data of single-point mutant proteins from the Protein Gym database replacement unit. Each protein data file was independently stored in CSV format. The following research will be carried out on this basis.

The precise quantification of SASA is an important prerequisite for exploring the relationship between protein structure and function. The following are the methods involved in the SASA data acquisition and system data integration part of this study:

This chapter's research adopts the classic DSSP (Dictionary of Secondary Structure of Proteins) algorithm to carry out the SASA calculation work. The DSSP algorithm is based on the three-dimensional atomic coordinates of proteins. It simulates water molecules with a probe sphere with a radius of 1.4A. By analyzing the geometric relationships and spatial arrangements between adjacent atoms, it can not only accurately calculate the solvo-accessible surface area of each amino acid residue, but also simultaneously identify the secondary structural elements of proteins, such as α-helix and β-folding.

Integrate the SASA data calculated by DSSP with the output results of the mutation prediction software to prepare for the subsequent analysis part.

3.3.3.1.2 Data Analysis Process of Correlation between SASA and ProtSSN Prediction Errors

1. Verify data integrity and use automated scripts to verify whether each file contains the necessary fields (SASA value, relative prediction error);

2. Eliminate the observational data that meet any of the following conditions:

(2.1) The SASA value is missing or exceeds a reasonable range ($0 \leq SASA \leq 1$).

(2.2.) The relative error value is missing or exceeds the defined range

3. The hierarchical analysis method was adopted to evaluate the correlation between SASA and prediction error:

(3.1) Single protein analysis

Calculate the Spearman rank correlation coefficient (ρ) and its significance level (p-value) of each protein file

(3.2) Global Integration:

The overall correlation coefficient is calculated by the weighted average method:

Among them, ni represents the effective mutation number of the i-th protein

4. Build a visualization framework based on the Python ecosystem:

(4.1) Use Kernel Density Estimation (KDE) to plot the distribution of correlation coefficients.

(4.2) Design a bivariate scatter plot, with Spearman's ρ on the horizontal axis and the average SASA value on the vertical axis. Introduce color gradients to characterize statistical significance (-log10(p-value)) and the number of point size mapping mutations

5. All analyses were completed in the Python 3.10 environment. The main dependent libraries include:

(5.1) Scientific computing: NumPy 1.24, SciPy 1.10

(5.2) Data Processing: Pandas 2.0

(5.3) Visualization: Matplotlib 3.7, Seaborn 0.12

(5.4) Parallel Computing: Concurrent. Futures

3.3.3.2 Results and Analysis

This study systematically analyzed the association patterns between the solvent-accessible surface area (SASA) and mutation prediction errors of 148 single-point mutant proteins, revealing significant global negative correlation characteristics. Statistical analysis indicated that 86.5% of the proteins (128/148) exhibited statistically significant correlation ($p < 0.05$). The weighted average of the Spearman rank correlation coefficient (ρ) was -0.25 (standard deviation ±0.18), the median was -0.29, and the distribution range was [-0.59, 0.37]. Among them, 75% of the protein ρ values were concentrated in the range of [-0.38, -0.13], showing a clear left-biased trend. This result suggests that as the SASA value increases (characterizing the exposure of residues to the solvent environment), the prediction error shows a systematic decrease; Conversely, the low SASA residues located in the protein core may lead to a significant increase in the difficulty of prediction due to the complex co-conformational effect. It is worth noting that although most proteins follow the rule of negative correlation, individual cases (such as the maximum ρ value of 0.37) show reverse association, suggesting that there may be unconventional mechanisms of action in specific structural scenarios.

The following figure shows the box plot of the correlation between ProtSSN prediction error and SASA (Figure 3.11) and the landscape map of the correlation between ProtSSN prediction error and SASA (Figure 3.12).

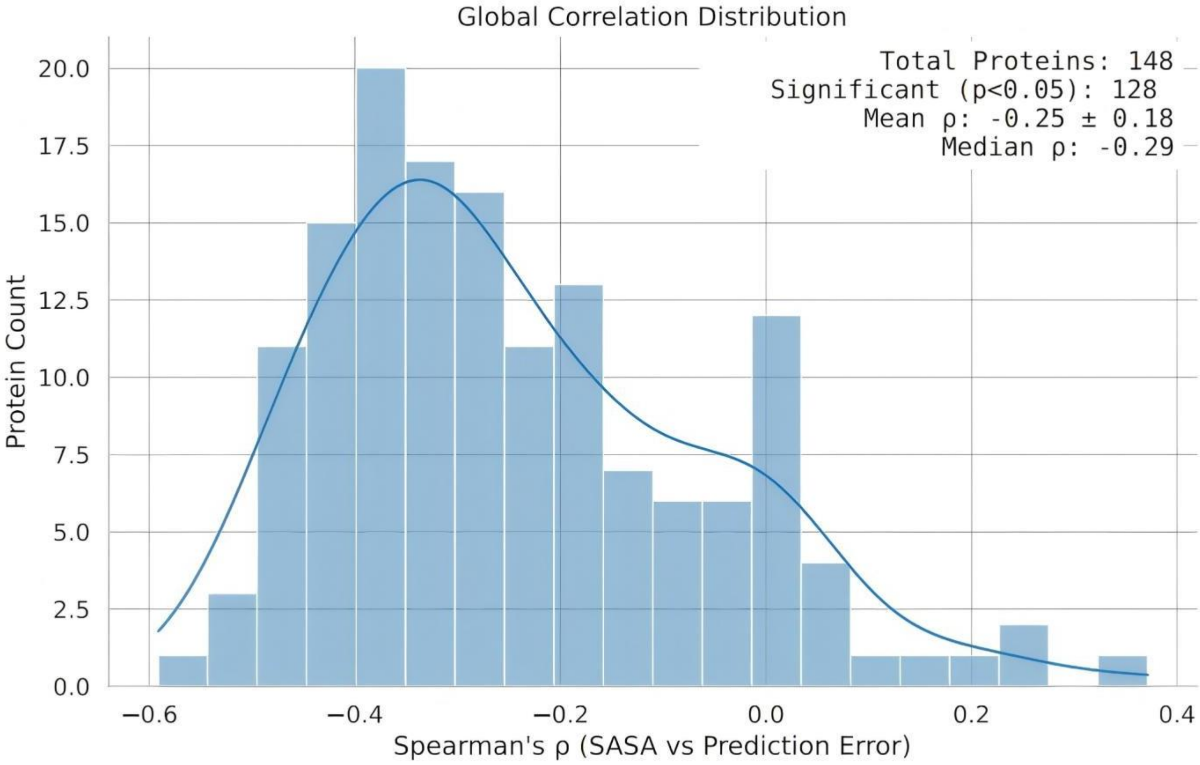

Figure 3.11 Box plot of the correlation between ProtSSN prediction error and SASA

Data quality analysis showed that the number of mutations covered by the study was significantly heterogeneous. The number of mutations of individual proteins ranged from 63 to 16,897 (median 2,384.5), among which proteins with a high number of mutations (such as n>10,000) might strengthen the global negative correlation trend through a leverage effect. The associated landscape map (Figure 3.12) further reveals that proteins with an average SASA value higher than 40 tend to cluster in strongly negatively correlated regions (ρ<-0.3), and large-sized data points (corresponding to high mutation numbers) are mostly distributed in this region, supporting the robustness of the statistical results. The kernel density curve shows that the distribution peak of Spearman's ρ is located around -0.3, which is highly consistent with the median value. The box plot (Figure 3.11) analysis indicates that the data dispersion is controllable (interquartile range [-0.38, -0.13]), and no extreme outliers. occur. From a biological perspective, the emergence of a negative correlation trend might imply that the current prediction model can better predict the impact of solvation on surface residues, but it has shortcomings in capturing the conformation changes of core residues. This conclusion points out the direction for improving protein engineering algorithms. For example, taking into account the parameters of the cooperative network among core residues may significantly enhance the prediction accuracy of the model.

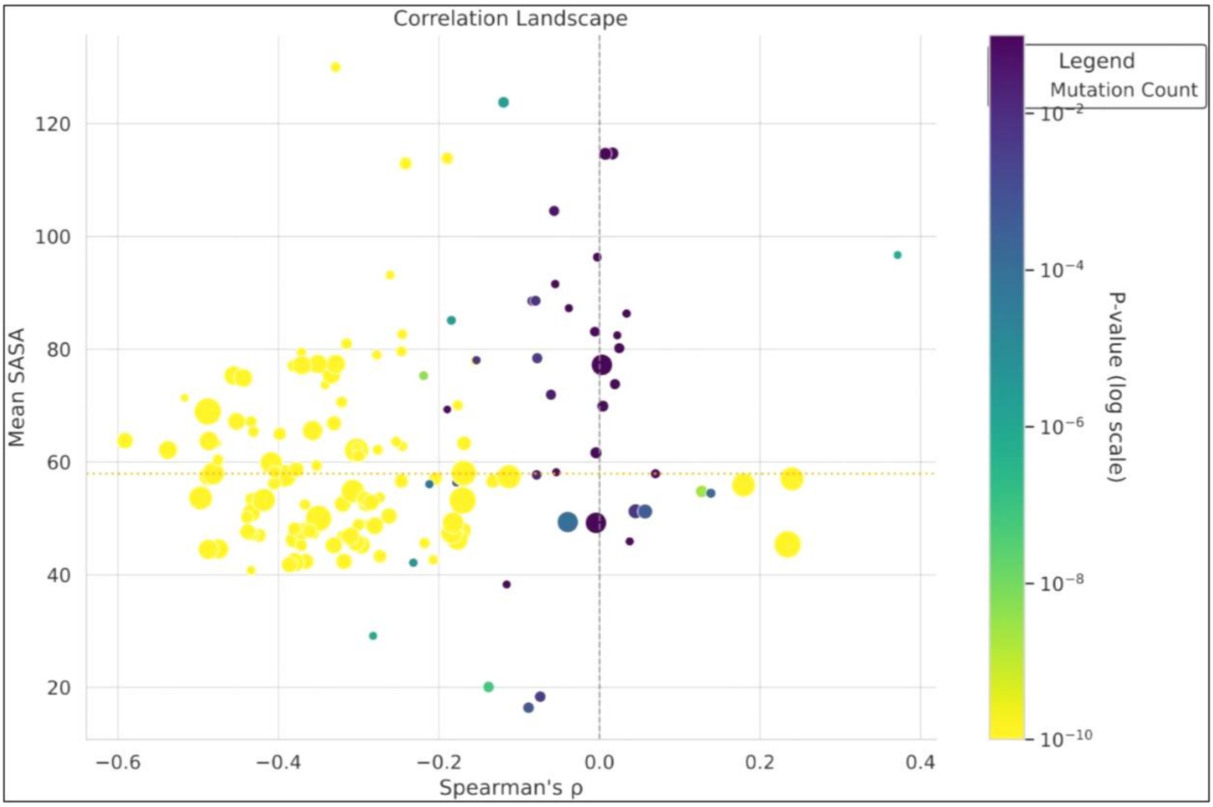

Figure 3.12 Landscape map of the correlation between ProtSSN prediction error and SASA

However, this study also has shortcomings. As the samples cover multiple species, this heterogeneity may interfere with the experimental results. The subsequent research can adopt the local weighted regression technique to explore the potential nonlinear laws in the data. Meanwhile, through subgroup analysis by species, explore how evolutionary conservation affects the relationship between prediction error and solvent accessible surface area (SASA). In addition, conducting independent verification experiments on proteins with a large number of mutations is helpful to determine whether such data will deviate from the overall trend, thereby improving the analysis system of large-scale mutation data.

Summary of chapter III

This chapter focuses on the application of ProtSSN software in the prediction and evaluation of protein secondary structure mutations. By introducing the DSSP (Precise Calculation of Secondary Structure) and SOMPA (Statistical Association Analysis) algorithms, a two-dimensional verification framework is constructed to analyze the influence mechanism of mutations on regular secondary structures (such as α-helix and β-fold).

The correlation analysis of the dual algorithms (DSSP/SOMPA) shows that the prediction accuracy of the software for regions with high structural order (such as α-helical cores) is significantly better than that for irregular structures (such as random curls), and it is speculated that it may have a dependent characteristic on stable conformational anchor points.

The multimodal analysis capability (sequence-structure-function coupling) of ProtSSN has unique advantages in structural bioinformatics, especially excelling in the scenario of co-mutation.

In the future, it is necessary to further optimize the prediction model for irregular structural regions and expand the adaptability of the algorithm to dynamic conformations (such as flexible ring regions) to enhance the full-scenario coverage capability. The research objective of this chapter focuses on the correlation mechanism between SASA and prediction accuracy, systematically exploring the quantitative relationship between the mutation prediction accuracy of ProtSSN and the amino acid solvent accessible surface area (SASA), aiming to analyze the influence law of protein surface accessibility on the prediction of mutation effects and reveal the prediction characteristics of the software in different structural environments.

The method framework and data analysis are based on the DSSP program to accurately calculate the SASA data and construct the correlation model between the prediction errors of SASA and ProtSSN. Statistical correlation analysis (such as Spearman correlation coefficient), kernel density estimation (peak -0.3), and box plot (IQR [-0.38, -0.13]) were used to quantify the data distribution and trend, covering 63 to 16,897 mutation samples (median 2,384.5) to ensure the comprehensiveness of the analysis.

The core findings reveal significant negative correlations and structural dependencies. SASA was globally significantly negatively correlated with the prediction error (weighted average ρ= -0.25), and 86.5% of the protein samples showed a statistically significant association.

The results clarified the advantages of ProtSSN in surface accessible residue prediction, providing a preferred target for stability design (such as antibody affinity optimization.

CONCLUSIONS

This study took the artificial intelligence protein mutation design software ProtSSN as the object. Through multi-dimensional analyses such as integrating deep mutation scan data, secondary structure prediction, and SASA correlation, it systematically revealed its technical characteristics, mechanism of action, and engineering application potential in the prediction of protein mutation effects. It provides theoretical and practical references for AI-driven protein design.

Studies have shown that ProtSSN achieves the deep integration of protein sequence semantics and three-dimensional structural topology through a dual-modal collaborative pre-training framework. Its core advantage lies in the precise analysis of the mutation effects of regular secondary structures (α-helical, extended chains). In the Protein Gym benchmark test, the Spearman correlation coefficient of the model for single-point mutations reached 0.429, and for multi-point mutation scenarios, it was 0.550, significantly superior to the traditional sequence model, reflecting its efficient capture ability for the synergistic effect of amino acid sites. This performance improvement stems from the quantitative coding of the geometric characteristics of the protein microenvironment, with high accuracy in predicting the distribution of mutations in the α-helix or extended chain regions. The model captures structural disturbances such as the breaking of hydrogen bonds in the main chain and the destruction of hydrophobic cores through the isovariant graph neural network (EGNN), achieving a nonlinear mapping from sequence mutations to thermodynamic stability (such as ΔTm, ΔΔG).

The secondary structure analysis based on statistical principles and SOPMA software further reveals that the prediction accuracy of ProtSSN is closely related to the structural orderliness: the prediction accuracy of regular structural mutations is improved compared with irregular occur. The essence of the difference lies in the multi-level coupling modeling of "sequence tendency - geometric feature - functional effect" by the model. The hydrophobic accumulation of the α -helical core and the hydrogen bond network between the extended chain sheets provide clear structural anchor points for the model. However, the flexible feature of irregular curling lacks stable conformational references, and the prediction accuracy is significantly affected by the synergy effect of adjacent regular structures. This in-depth analysis of the structure-function correlation provides precise computational guidance for targeted stability optimization and functional regulation.

Further analysis of SASA based on DSSP revealed that there was a significant global negative correlation between solvent accessible surface area (SASA) and mutation prediction error (weighted average ρ= -0.25), and 86.5% of the proteins showed a statistically significant association. The data covered 63 to 16,897 mutations (median 2,384.5), and the highly mutated samples (n>10,000) strengthened the negative correlation trend through the leverage effect. The kernel density distribution (peak -0.3) and the box plot (IQR [-0.38, -0.13]) indicate that the data dispersion is controllable. The associated landscape map reveals that the high SASA protein (>0.4) is concentrated in the strongly negatively correlated area (ρ<-0.3), supporting the reliability of surface residue prediction and the complexity of the conformation synergy effect in the core area. The results provide a structural feature basis for optimizing the protein prediction algorithm and suggest that the influence of highly mutated samples on the global trend needs to be evaluated specifically.

From the perspective of practical application, ProtSSN's lightweight architecture (110 million parameters) and efficient inference speed (≤2 hours per task) have broken through the computational bottleneck of traditional molecular simulation. Its open-source feature and visualization module have significantly lowered the technical threshold, and can increase the efficiency of mutant screening by more than three times in wet experiments. Effectively reduce the trial-and-error costs in research and development in fields such as biomedicine and biomanufacturing. For instance, in enzyme engineering modification, the model precisely locates high-potential mutation sites through site matching rate (SMR) and stability prediction error (SPE) indicators, promoting the transformation from an "empirical trial and error" design paradigm to a "computation-first" one.

This study innovatively explored the intrinsic connection between the prediction results of ProtSSN and the secondary conformational characteristics of proteins as well as the solvent contact area on the basis of previous research. It adopted the Protein Gym standard dataset combined with the diverse samples of the PDB structure database to ensure the wide representativeness of the analysis objects in terms of functional categories and species origin. In the experiment, the SOPMA platform was used to analyze the secondary conformation distribution characteristics of the samples, and the DSSP program was adopted to accurately calculate the solvent-accessible surface parameters of each amino acid residue. By integrating bioinformatics and statistical analysis methods, and focusing on the two key dimensions of the difference characteristics of prediction accuracy in different secondary conformation regions and the correlation pattern between solvent contact area and prediction error, the research results show the quantitative relationship between the characteristics of the local microenvironment of proteins and the prediction of mutation effects. This not only deepens the understanding of the mechanism of amino acid variation, it provides important theoretical support for optimizing the parameters of the prediction algorithm and guiding the rational design of proteins.

Meanwhile, this study also has certain deficiencies. The research mainly focuses on the Protein GYM dataset and does not conduct larger-scale and more in-depth research and analysis on other datasets. Although the Protein GYM dataset has significant value and covers various protein mutation phenotypic data, However, different datasets have differences in data characteristics and covered contents. The lack of research on other datasets may lead to limitations in the research results. And for the secondary structure chapter, there are certain limitations due to issues such as technology and sample size. Future research can expand the scope of the dataset, incorporate proteins with broader correlation coefficients for comprehensive analysis, and learn more advanced techniques to make up for these deficiencies and promote the in-depth development of this field.

In conclusion, through the multi-dimensional system of "data-driven - structure analysis - SASA association", this study not only clarified the technical advantages of ProtSSN in structure-driven mutation design, but also the research results provided a new perspective for understanding the complex mapping relationship of "sequence - structure - solvent accessibility - function". Promoting protein engineering towards precision and efficiency is of great practical significance for accelerating the development of new enzyme preparations and the design of therapeutic antibodies, and it is hoped that it can contribute to the future development of the intersection of synthetic biology and computational biology.

1. The content of this chapter is a summary and analysis of the full text. ProtSSN fuses sequence and structural features based on a dual-modal framework. In the Protein Gym test, its single-point (ρ=0.429) and multi-point mutation (ρ=0.550) prediction performance is significantly better than that of traditional models. The EGNN network accurately models structural disturbances such as hydrogen bond breakage and hydrophobic core damage.

2. The regular secondary structure (α-helix/extended chain) has high prediction accuracy and relies on stable conformational anchor points (such as hydrogen bond networks); The high SASA region (>0.4) was strongly negatively correlated with the prediction error (weighted ρ=-0.25), and the conformational synergy effect in the core area increased the complexity.

3. The lightweight architecture (110 million parameters) and efficient inference (≤2 hours per task) increase the screening efficiency of wet experiments by three times, reduce the trial-and-error cost of antibody/enzyme design, and the open-source module promotes "computation-first" applications.

4. Limited to the insufficient coverage of the Protein GYM dataset, it is necessary to expand heterogeneous data such as membrane proteins. In the future, multi-scale algorithms will be integrated to optimize the prediction of the core area and enhance the adaptability of dynamic conformation.

5. Research and establish a new paradigm of dynamic mapping of "sequence - structure - function" to accelerate the precise design of enzyme preparations and therapeutic antibodies and promote the intelligent development of protein engineering.

REFERENCES

1. Jiacheng L. Research on the Thermodynamic Mechanism of Protein Folding and Docking Based on Entropy-Enthalpy Compensation [D]. Harbin: Harbin Institute of Technology, 2023.

2. Xiaohui C. Protein Design Based on Structural Bioinformatics [D]. Shanghai: Graduate School of Chinese Academy of Sciences (Shanghai Institute of Life Sciences), 2006.

3. Liting Z. Molecular Modification of Sucrose Synthase and Its Efficient Preparation of UDP-Glucose [D]. Wuxi: Jiangnan University, 2024.

4. Peng X. Protein Design Based on Statistical Energy Function [D]. Hefei: University of Science and Technology of China, 2015.

5. Weibin G. Major Breakthroughs in AlphaFold Structure Prediction and Its Impact and Challenges on Protein Research [J]. Progress in Biochemistry and Biophysics, 2024, 51(12): 3073-3083.

6. Yunan S, Chuan Y, Dongyu Z. Protein-related artificial intelligence algorithms in the AlphaFold era and their applications [J/OL]. Advances in Physiology Research, 1-15 [2025-05-10]. http://kns.cnki.net/kcms/detail/11.2270.R.20250311.1045.002.html.

7. Tan Y, Zhou B, Zheng L, et al. Semantical and geometrical protein encoding toward enhanced bioactivity and thermostability[J]. et al., 2024, 13: RP98033.

8. UniProt Consortium. UniProt: the universal protein knowledgebase in 2023 [J]. Nucleic Acids Research, 2023, 51(D1): D523-D531.

9. Notin P, Kollasch AW, Ritter D, et al. Protein Gym: Large-scale bench arks for protein fitness prediction and design[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems (Neu rIPS '23). New Orleans, LA, USA: NeurIPS Foundation, 2023: 64331-64 379.

10. Yao H, Heyu Z, Dongsong Y, et al. Research on the Water Retention P roperies of Sheep Meat from Beach Areas Mediated by Myofibrillar Protein Oxidation [J/OL]. Chinese Journal of Food Science, 1-12 [2025-05-10]. http://kns.cnki.net/kcms/detail/11.4528.TS.20250328.1403.014.html.

11. Xinping Y, Shiyao W, Qinggang M. Exploring the regulatory mechanism of Bái Yǎnjian Wen Pill on liver cell carcinoma based on bioinformatics [J]. Chinese Journal of Traditional Chinese Medicine, 2025, 43(04): 82-89 + 280-284.

12. Luwen W. Rational Design Combined with Fixation to Improve the Thermal Stability of CALB [D]. Wuxi: Jiangnan University, 2020.

13. Mao S, Cheng X, Zhu Z, et al. Engineering a thermostable version of D-allulose 3-epimerase from Rhodopirellula baltica via site-directed mutagenesis based on B-factors analysis[J]. Enzyme and Microbial Technology, 2020,132(C):109441.

14. Miao H, Xiang X, Han N, et al. Improving the Thermostability of Serine Protease PB92 from Bacillus alcalophilusvia Site-Directed Mutagenesis Based on Semi-Rational Design [J]. Foods, 2023,12(16):

15. Baoyan L, Shuai L, Hao S, et al. Protein computational simulation technology driven by the integration of artificial intelligence and physical principles [J]. Journal of Biotechnology Engineering, 2025, 41(03): 917-933.

16. Jiahao B, Guangyu Y. Artificial Intelligence-Assisted Protein Engineering [J]. Synthetic Biology, 2022, 3(03): 429-444.

17. Zhihao C, Menglin J, Yifei Q. Research Progress on Artificial Intelligence Algorithms for Protein Structure Design [J]. Synthetic Biology, 2023, 4(03): 464-487.

18. Wanshan N. Artificial Intelligence-Based Research on Post-Translational Modification of Proteins and Function Prediction in Biology [D]. Wuhan: Huazhong University of Science and Technology, 2021.

19. Biyan R, Lu L, Kunxian S, et al. Advances in the Application of Artifice ail Intelligence in Predicting Protein Interactions between Host and Patho gen [J]. Acta Virological, 2024, 40(05): 1121-1136.

20. Jiang F, L i M, Dong J, et al. A general temperature-guided language mod el to design proteins of enhanced stability and activity. [J]. Science advances,2024,10(48): eadr2641.

21. Ziyi Z, Liang Z, Yuanxi Y, et al. Enhancing efficiency of protein language ge models with minimal wet-lab data through few-shot learning [J]. Nature Communications, 2024, 15(1):5566.

22. John J, Richard E, Alexander P, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021,596(7873):583-589.

23. Zhao W, Wang C. Protein designer David Baker: I like doing things that seem like magic [J]. Natl Sci Rev, 2020, 7(8):1410-1412.

24. Qiu Y, Wei GW. Artificial intelligence-aided protein engineering: from topological data analysis to deep protein language models[J]. Brief Bioinform, 2023, 24(5): bbad289.

25. Xiaojing L, Chunrun W, Meihuan Z, et al. Secondary structure of the amide III band of proteins in natural rubber and its thermal denaturation [J]. Synthetic Rubber Industry, 2024, 47(05): 389-393.

26. Xiaoyi W, Hao W, Can Y, et al. Research Progress on Improving Thermal Stability of Enzymes through Protein Engineering [J]. Food Science, 2024, 45(19): 263-271.

27. Mengyu L. Research on Protein Thermal Stability Prediction Method Based on Deep Learning [D]. Beijing: Beijing University of Chemical Technology, 2024.

28. Xiaoyang X. Research on Protein Thermal Stability Prediction Model Based on Deep Learning [D]. Wuxi: Jiangnan University, 2024.

29. Jianjun Z. Research on Prediction Model for Thermal Stability of Multiple Species Proteins [D]. Suzhou: Soochow University, 2023.

30. Yuanlin G. Rational Design to Enhance the Thermal Stability of Zearalenone Hydrolase and Analysis of Its Mechanism [D]. Jiangnan University, 2023.

31. Gao Y, Wang H, Zhou J, et al. An easy-to-use three-dimensional protein-structure-prediction online platform "DPL3D" based on deep learning algorithms [J]. Current Research in Structural Biology, 2025, 9100163-100163.

32. Javier J L, Luna R, Alejandro S, et al. Structural Protein Effects Underpinning Cognitive Developmental Delay of the PURA p. Phe233del Mutation Modelled by Artificial Intelligence and the Hybrid Quantum Mechanics–Molecular Mechanics Framework [J]. Brain Sciences, 2022, 12(7):871-871.

33. Jonghanne P, Hyung Gyo C, Jewel P, et al. Artificial Intelligence-Powered Hematoxylin and Eosin Analyzer Reveals Distinct Immunologic and Mutational Profiles among Immune Phenotypes in Non-Small-Cell Lung Cancer. [J]. The American journal of pathology, 2022, 192(4):701-711.

34. Fabrizio P, Martin S, Marianne R. Artificial intelligence challenges for predicting the impact of mutations on protein stability. [J]. Current opinion in structural biology, 2021, 72161-168.

35. Xu C, Zhidong C, Daiyun X, et al. De novo Design of G Protein-Coupled Receptor 40 Peptide Agonists for Type 2 Diabetes Mellitus Based on Artificial Intelligence and Site-Directed Mutagenesis [J]. Frontiers in Bioengineering and Biotechnology, 2021, 9694100-694100.

36. Zhongju Y, Tao S, Sheng X, et al.AF2-mutation: adversarial sequence mutations against AlphaFold2 in protein tertiary structure prediction[J].Acta Materia Medica, 2024, 3(4):462-476.

37. Inoue A, Zhu B, Mizutani K, et al. Prediction of Single-Mutation Effects for Fluorescent Immunosensor Engineering with an End-to-End Trained Protein Language Model. [J]. JACS Au, 2025, 5(2):955-964.

38. V E D P, M H C R, S T A A, et al. Exploring Protein Super Secondary Structure Through Changes in Protein Folding, Stability, and Flexibility. [J]. Methods in molecular biology (Clifton, N.J.), 2019, 1958173-185.

39. Topolska M, Beltran A , Lehner B. Deep indel mutagenesis reveals the impact of amino acid insertions and deletions on protein stability and function. [J]. Nature communications, 2025, 16(1):2617.

40. Anusuya S, Jeyakumar N. Combination of site directed mutagenesis and secondary structure analysis predicts the amino acids essential for stability of M. leprae Mur E. [J]. Interdisciplinary sciences, computational life sciences, 2014, 6(1):40-7.

41. Zhang L, Pang H, Zhang C, et al. Venus Mut Hub: A systematic evaluation of protein mutation effect predictors on small-scale experimental data [J]. Acta Pharmaceutica Sinica B, 2025, 15(5):2454-2467.

42. Samanta P, Ghorai S. Prediction of Safed adhesin strong binding peptides for pilus proteins assembly suppression in the prevention of Salmonella-induced biofilm formation using virtual mutagenesis studies[J]. In Silico Pharmacology, 2025, 13(1):25-25.

43. Liao J, Wu M, Meng F , et al. Studying the Protein Thermostabilities and Folding Rates by the Interaction Energy Network in Solvent. [J]. Journal of computational chemistry, 2025, 46(11): e70113.

44. Rossi I, Barducci G, Sanavia T, et al. Mass balance approximation of unfolding boosts potential-based protein stability predictions[J]. Protein Science,2025,34(5):e70134-e70134.

45. Yan Q, Pan S, Cheng Z, et al. RAANMF: An adaptive sequence feature representation method for predictions of protein thermostability, PPI, and drug–target interaction [J]. Future Generation Computer Systems, 2025, 169107819-107819.

46. Komp E, Phillips C, Lee M L, et al. Neural network conditioned to produce thermophilic protein sequences can increase thermal stability [J]. Scientific Reports, 2025, 15(1):14124-14124.

47. Effect of overall charge of oligochitosan on structure and thermal stability of ovalbumin and thermodynamic properties ovalbumin/oligochitosan systems [J]. Journal of Thermal Analysis and Calorimetry, 2025, (prepublish):1-13.

48. Jiang Y, Yuan X, Zheng S, et al. The influence of reduced amino acid alphabets on prediction orthologous protein thermostability [J]. Biologia, 2025, (prepublish):1-11.

49. Bukovics P, Lőrinczy D. Deconvolution Analysis of G and F-Actin Unfolding: Insights into the Thermal Stability and Structural Modifications Induced by PACAP [J]. International Journal of Molecular Sciences, 2025, 26(7):3336-3336.

50. Yeh C, Hayes L R. Predicting Thermodynamic Stability at Protein G Sites with Deleterious Mutations Using λ-Dynamics with Competitive Screening. [J]. The journal of physical chemistry letters, 2025, 16(13):3206-3211.