Symphony of Fate: Weaving Life through the Music of Amino Acids

Anonymous authors
Paper under double-blind review

Abstract

Understanding the intricate relationship between protein sequences and their biological functions is a significant challenge in bioinformatics. Traditional structural methods capture only static features of proteins, often overlooking their dynamic and functional aspects. In this paper, we present a novel approach that encodes protein sequences into musical compositions, transforming amino acid properties, secondary structures, and tertiary dynamics into musical elements. Using multidimensional musical encoding, we map the physicochemical properties of proteins to musical features like pitch, timbre, rhythm, and harmony. We then apply machine learning models, including Random Forest, Support Vector Machines, and XGBoost, to analyze over 2,000 protein sequences, achieving a classification accuracy of 91.04% and a cross-validation score of 99.68%. Our results demonstrate a significant correlation between the harmony features of music encoding and protein functionality, such as enzymatic activity. At the same time, we combine the GFP dataset to learn feature-directed evolution using the diffusion model. This work introduces a novel methodology for protein function prediction and design, offering new insights into protein dynamics and structure-function relationships, and paving the way for future research in protein engineering and directed evolution.

1 Introduction

The relationship between protein sequence, structure, and function is central to understanding biological processes. Proteins are key molecules that regulate and execute a vast array of biological functions, from metabolism and signal transduction to immune responses and structural integrity. Traditionally, research in this field has relied heavily on methods like X-ray crystallography, nuclear magnetic resonance (NMR) (Clore & Gronenborn (1987); Bax & Grzesiek (1993)), and cryo-electron microscopy (Cryo-EM) (Yip et al. (2020) Nogales & Scheres (2015)), which provide valuable insights into protein structures and interactions. However, these methods often focus on static aspects of proteins, leaving the dynamic and functional properties less explored. In recent years, the application of computational techniques, such as machine learning, has shown promise in bridging this gap, offering a deeper understanding of how protein sequences relate to their biological functions (Jumper et al. (2021); Varadi et al. (2022)).

While significant advancements have been made in protein structure analysis, existing methods often fall short in addressing the dynamic and functional aspects of proteins. Techniques such as X-ray crystallography and Cryo-EM provide static, high-resolution structural data, but they fail to capture the protein's dynamic behavior and interactions, which are essential for understanding function. Additionally, traditional computational methods, like sequence alignment and homology modeling, do not fully account for the complex, multidimensional nature of protein functions (Clore & Gronenborn (1991); Whitford (2013)). Deep learning-based approaches, such as AlphaFold, have made strides in predicting protein structures with remarkable accuracy; however, they remain limited in their ability to predict highly flexible regions and complex protein-protein interactions (Jumper et al. (2021); Varadi et al. (2022)). Furthermore, these methods tend to focus on isolated aspects of proteins, often neglecting the interplay between sequence, structure, and function. There is, therefore, a pressing need for innovative approaches that can holistically capture the multidimensional characteristics of proteins, particularly their functional dynamics, and provide a deeper understanding of their biological roles (Buehler (2023); Martin et al. (2021)).

In this study, we propose a novel interdisciplinary approach that bridges the gap between protein structure and function by encoding protein sequences into musical compositions. We map the physicochemical properties of amino acids, secondary structures, and tertiary dynamics into musical features such as pitch, timbre, rhythm, and harmony. This musical encoding allows us to represent the multidimensional characteristics of proteins in a dynamic and interpretable form. We then employ machine learning models, including Random Forest, Support Vector Machines, and XGBoost, to analyze these encoded protein sequences, achieving classification accuracies of 91.04% and a cross-validation score of 99.68%. Our findings reveal a significant correlation between musical harmony and protein functionality, particularly enzymatic activity. This work introduces an innovative method for protein function prediction and offers new insights into protein design and directed evolution, leveraging the powerful combination of bioinformatics and music theory to enhance our understanding of protein dynamics.

2 METHOD

In this study, we introduce a novel interdisciplinary approach that encodes protein sequences and their structural information into musical compositions (Yu et al. (2019); Giesa et al. (2011)). The overall method consists of three main components: (1) protein-to-music mapping, (2) spectral feature extraction, and (3) machine learning-based function prediction. The protein-to-music mapping transforms the physicochemical properties of amino acids, secondary structures, and tertiary dynamics into musical elements such as pitch, timbre, rhythm, and harmony. This step provides a multidimensional representation of protein data that captures both sequence and structural information in a dynamic form. Spectral feature extraction then analyzes the musical compositions using tools such as spectrograms, log-Mel spectrograms, and Mel-frequency cepstral coefficients (MFCC), revealing frequency patterns that correspond to structural and functional features of the proteins. Finally, we apply machine learning models—including Random Forest, Support Vector Machines, and XGBoost—to predict protein functionality based on the extracted spectral features. The method offers a new way to analyze proteins, revealing deeper insights into their function and structural dynamics, and demonstrates the ability to predict enzymatic activities with high accuracy.

The core of our approach lies in the protein-to-music mapping, where we translate the physicochemical properties of proteins into musical elements. Specific parts are shown in Figure 1 (Loy (2011); Benson (2006)). This step is divided into three key submodules:

- **Amino Acid to Musical Element Mapping:** Each amino acid is mapped to a musical note based on its physicochemical properties. The hydrophobicity and polarity of amino acids are represented through pitch: hydrophobic amino acids correspond to lower pitches, while more polar amino acids are mapped to higher pitches. Charged amino acids are represented through harmonic intervals to reflect their interactions within the protein structure. This mapping captures sequence-specific information, allowing us to translate the protein sequence into a series of musical notes.
- **Structural Mapping:** The secondary structures of proteins, such as alpha-helices and beta-sheets, are encoded through harmonies. Alpha-helices are associated with consonant intervals, representing their stability, while beta-sheets are mapped to dissonant intervals, reflecting their more rigid and complex structure. This structural information is integrated into the musical composition, enhancing the representation of protein functionality.
- **Tertiary Structure and Dynamic Features:** The tertiary structure of the protein is mapped through rhythmic variations. Stable regions of the protein correspond to steady rhythms, while more dynamic and flexible regions are represented by more complex or faster rhythms. These dynamic features reflect the functional regions of the protein and its flexibility, which are essential for understanding its biological role.

After the mapping process, the protein sequences are transformed into MIDI files, which contain the musical representation of the protein's sequence and structure.

Spectral Feature Extraction: Once the protein is mapped to music, we use a variety of spectral analysis techniques to extract relevant features. These techniques include the generation of spectrograms, log-Mel spectrograms, and power spectra, which provide insight into the frequency-domain

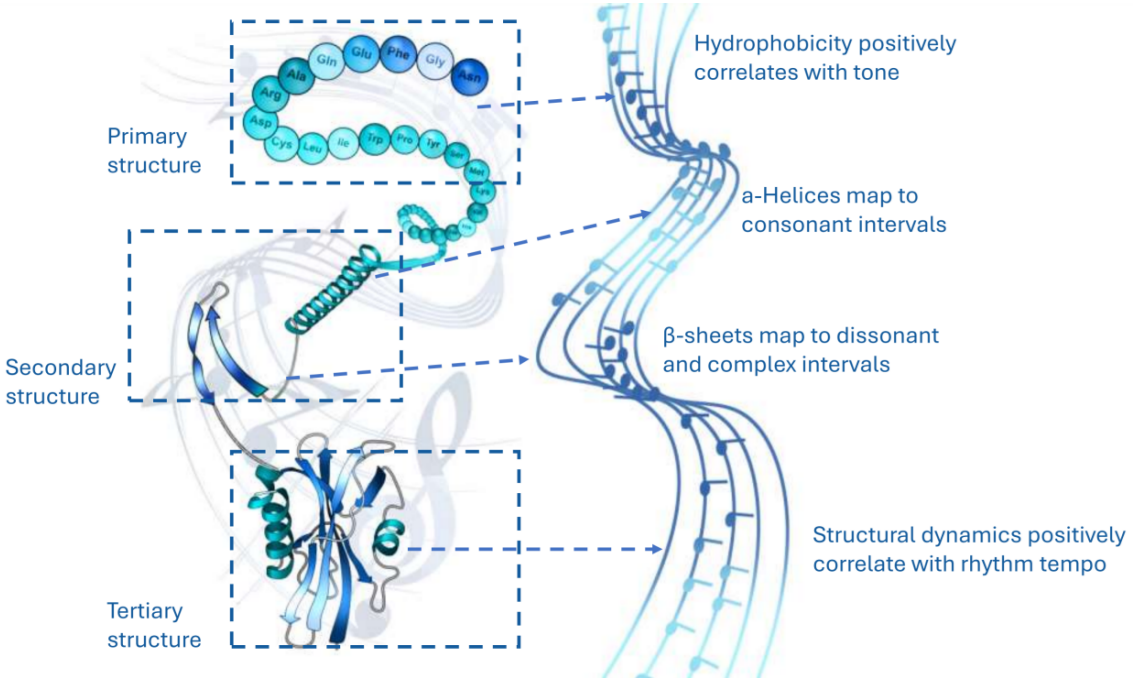


Figure 1: Schematic Diagram of the Mapping Rules

characteristics of the musical composition. These spectral features are then analyzed using Mel-frequency cepstral coefficients (MFCC), which are particularly useful for capturing both the timbral and rhythmic qualities of the encoded protein music. MFCC and other spectral features, such as spectral centroid and entropy, help identify functional areas of the protein and their corresponding dynamic behaviors.

Machine Learning for Protein Function Prediction: To predict protein functions, we use machine learning models, including Random Forest, Support Vector Machines (SVM), and XGBoost. First, we apply data enhancement techniques, such as synthetic data generation, to account for variations and noise inherent in biological systems. Then, we extract spectral features from the musical compositions, which are used as inputs to the machine learning models. These models are trained to classify protein functions based on their spectral signatures. Using techniques like recursive feature elimination (RFE) and synthetic minority over-sampling technique (SMOTE), we optimize the models and ensure they are robust against overfitting. The final model achieves high accuracy in predicting protein functionalities, including enzymatic activity, with a classification accuracy of 91.04% and a cross-validation score of 99.68%.

3 Result

We conducted experiments on a dataset of 1,980 protein sequences, covering various types of proteins including enzymes, antibodies, and structural proteins (Amitai et al. (2004); Hellinga (1997); Tay et al. (2021)). The protein sequences were obtained from publicly available databases such as the National Center for Biotechnology Information (NCBI), and enzyme data, including catalytic efficiency (kcat/km), were sourced from the Brenda database. To ensure data quality, we employed the CD-HIT clustering algorithm to remove redundant sequences with over 90% similarity. The dataset was then used for musical encoding, followed by spectral feature extraction using spectrograms and MFCC. Machine learning models, including Random Forest, Support Vector Machines, and XGBoost, were applied to classify protein functions based on their encoded musical features. Model performance was assessed using classification accuracy and cross-validation scores.

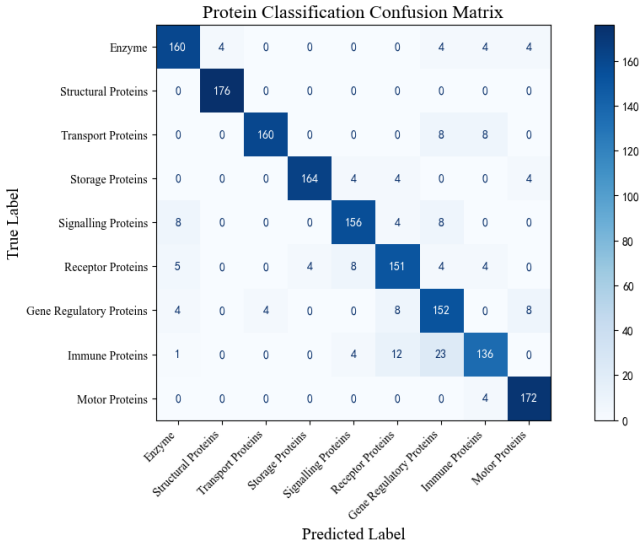
The results of our method demonstrate significant improvements in protein function classification when compared to traditional approaches. The model achieved a classification accuracy of 91.04%

on an independent test set, with a cross-validation score as high as 99.68%, indicating robust performance. To further validate the efficacy of our method, we conducted a comparative analysis between protein sequences encoded as music and traditional feature-based approaches. The results show that our method, based on spectral features extracted from the musical encoding, outperforms traditional methods in terms of both accuracy and the ability to capture dynamic protein properties. Figure 2a presents the confusion matrix for the nine protein classes, highlighting the model's high classification precision. Additionally, Figure 2b illustrates the learning curve, showing steady improvement in model performance during training, which reflects the model's ability to generalize well to unseen data. These results underscore the potential of our innovative music-based encoding method for protein function prediction.

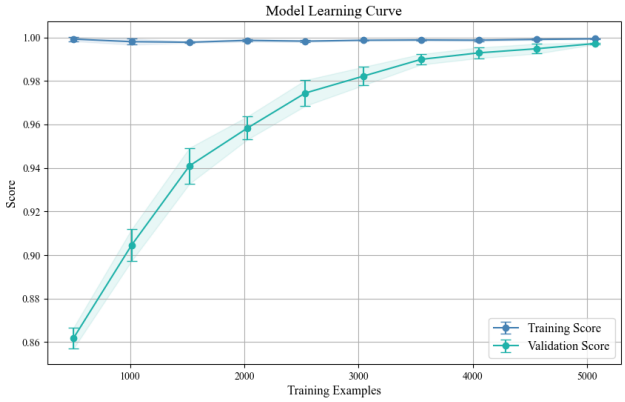
In this study, we present a novel approach that integrates an innovative method based on music-encoded green fluorescent protein (GFP) directed evolution with diffusion generation modelling. In the directed evolution experiments of GFP, we convert amino acid sequences into audio signals and optimise the function by extracting their spectral features. Utilising the Diffusion Model, we introduced noise and guided the generation process, thereby ensuring that the generated protein sequences approached the emission wavelength and brightness of the target, and the distribution of the sample space was sufficiently diverse to cover the original spatial range, as illustrated in Figure 2c. The diffusion model effectively guides the sequence towards a better functional outcome by progressively perturbing the input sequence (as shown in Figure 2d, there is a cumulative effect of multiple mutations generated by gradual evolution) and dynamically adjusting the noise amplitude according to the condition factor. Meanwhile, the results of the original wild-type sequences and the generated sequences are shown in Figure 2e, f, which shows that the space of the generated sequences is more widely covered by samples. Consequently, we were able to optimise the existing GFP mutant sequences and generate new sequences with superior performance. Specifically, we employed the conditional diffusion model to progressively generate GFP mutants that better match the target properties. The experimental results validate the potential of the diffusion model in protein-directed evolution and demonstrate the effectiveness of music coding in feature extraction and function optimisation.

4 Conclusion

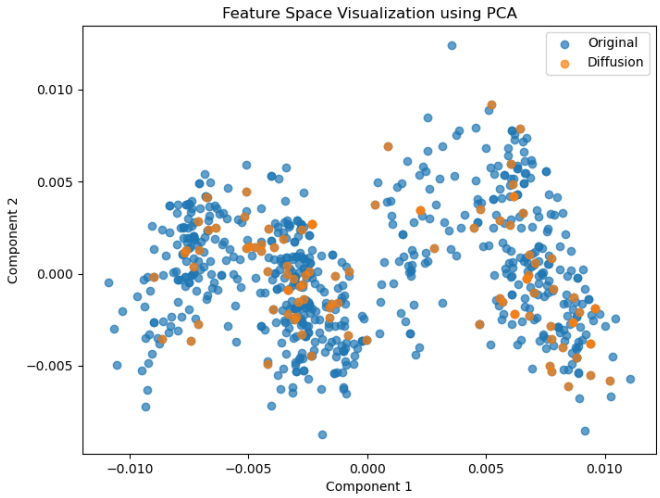
In this study, we introduced a novel method that encodes protein sequences and structures into musical compositions to predict protein functions. Our approach, combining music theory with machine learning, demonstrated high classification accuracy and a strong correlation between musical harmony and protein functionality. The results highlight the potential of this interdisciplinary framework for enhancing protein function prediction and design. Future work will focus on refining the protein-to-music mapping process, expanding the dataset to include a broader range of protein types, and further optimizing the machine learning models to improve prediction accuracy for complex protein interactions.



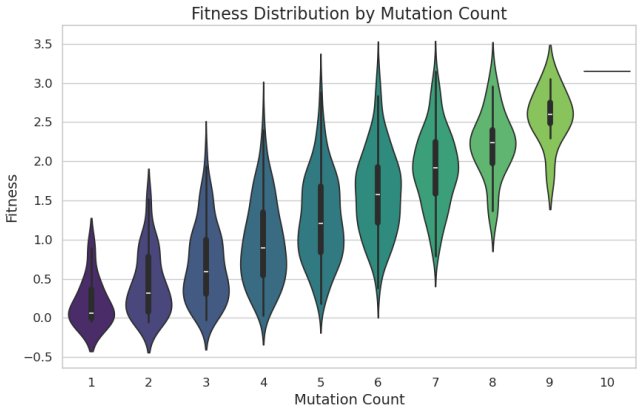
(a)



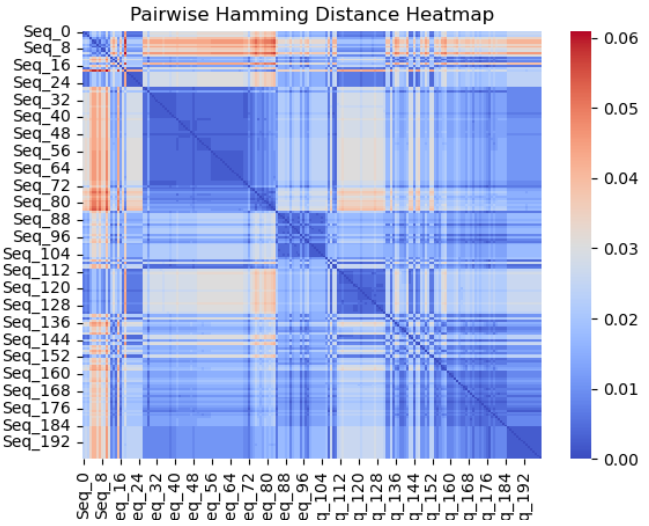
(b)



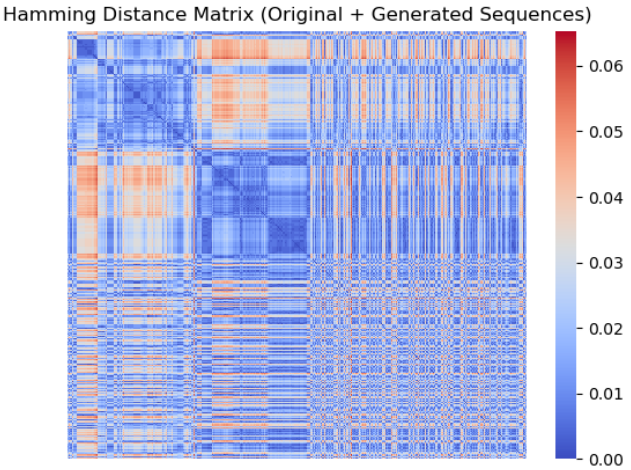
(c)



(d)



(e)



(f)

Figure 2: Test Results

References

G. Amitai, A. Shemesh, E. Sitbon, M. Shklar, D. Netanely, I. Venger, and S. Pietrokovski. Network analysis of protein structures identifies functional residues. *Journal of Molecular Biology*, 344 (4):1135–1146, 2004.

A. D. Bax and S. Grzesiek. Methodological advances in protein nmr. In *NMR of Proteins*, pp. 33–52. 1993.

D. Benson. *Music: A Mathematical Offering*. Cambridge University Press, 2006.

M. J. Buehler. Unsupervised cross-domain translation via deep learning and adversarial attention neural networks and application to music-inspired protein designs. *Patterns*, 4(3), 2023.

G. M. Clore and A. M. Gronenborn. Determination of three-dimensional structures of proteins in solution by nuclear magnetic resonance spectroscopy. *Protein Engineering, Design and Selection*, 1(4):275–288, 1987.

G. M. Clore and A. M. Gronenborn. Structures of larger proteins in solution: three-and four-dimensional heteronuclear nmr spectroscopy. *Science*, 252(5011):1390–1399, 1991.

T. Giesa, D. I. Spivak, and M. J. Buehler. Reoccurring patterns in hierarchical protein materials and music: the power of analogies. *BioNanoScience*, 1:153–161, 2011.

H. W. Hellinga. Rational protein design: combining theory and experiment. *Proceedings of the National Academy of Sciences*, 94(19):10015–10017, 1997.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, and D. Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

G. Loy. *Musimathics, Volume 1: The Mathematical Foundations of Music*, volume 1. MIT Press, 2011.

E. J. Martin, T. R. Meagher, and D. Barker. Using sound to understand protein sequence data: new sonification algorithms for protein sequences and multiple sequence alignments. *BMC Bioinformatics*, 22:1–17, 2021.

E. Nogales and S. H. Scheres. Cryo-em: A unique tool for the visualization of macromolecular complexity. *Molecular Cell*, 58(4):677–689, 2015.

N. W. Tay, F. Liu, C. Wang, H. Zhang, P. Zhang, and Y. Z. Chen. Protein music of enhanced musicality by music style guided exploration of diverse amino acid properties. *Heliyon*, 7(9), 2021.

M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, and S. Velankar. Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022.

D. Whitford. *Proteins: Structure and Function*. John Wiley & Sons, 2013.

K. M. Yip, N. Fischer, E. Paknia, A. Chari, and H. Stark. Atomic-resolution protein structure determination by cryo-em. *Nature*, 587(7832):157–161, 2020.

C. H. Yu, Z. Qin, F. J. Martin-Martinez, and M. J. Buehler. A self-consistent sonification method to translate amino acid sequences into musical compositions and application in protein design using artificial intelligence. *ACS Nano*, 13(7):7471–7482, 2019.

Table 1: Source Databases for Protein Data

Types of Materials	Sources	URLs
Protein FASTA Sequences	NCBI Database	`https://www.ncbi.nlm.nih.gov/`
	UniProt Database	`https://www.uniprot.org/`
kcat	SABIO-RK Database	`http://sabio.h-its.org/`
km	EBI Database	`https://www.ebi.ac.uk/`

Appendix

A Experimental Setup and Data Details

A.1 Dataset

The protein dataset used in this study contains 1,980 protein sequence samples, covering various types of proteins including enzymes, antibodies, and structural proteins. All protein sequences are obtained from public databases, as shown in Table 1 below:

To ensure the diversity and accuracy of the experimental data, we employed the **CD-HIT clustering algorithm** to remove redundant sequences, retaining only those protein sequences with a similarity greater than 90%, thus ensuring the dataset's representativeness and comprehensiveness.

A.2 Data Preprocessing

During data preprocessing, each protein sequence underwent the following operations:

1. **Redundancy Removal**: The CD-HIT algorithm was used for sequence clustering to remove sequences with more than 90% similarity to other sequences.
2. **Handling Missing Values**: For sequences with missing information, linear interpolation was applied to fill in the missing amino acid positions.
3. **Normalization**: The physicochemical properties (e.g., hydrophobicity, polarity, charge) of each amino acid were normalized to ensure all features are on the same scale.

B Algorithm Details and Supplementary Formulas

B.1 Mapping Proteins to Music

Protein sequences are converted into musical elements by mapping the physicochemical properties of their amino acids. The detailed mapping process is as follows:

- **Mapping Amino Acids to Notes**: Each amino acid's physicochemical properties (such as hydrophobicity and polarity) are mapped to musical notes. Specifically, amino acids with higher hydrophobicity are mapped to lower pitches, while those with stronger polarity are mapped to higher pitches.

$$P(p_i) = \text{Pitch}\big(H(p_i)\big) \tag{1}$$

  where $P(p_i)$ denotes the pitch of amino acid p_i, $H(p_i)$ represents its hydrophobicity, and H is the hydrophobicity function.

- **Mapping Secondary Structures**: Protein secondary structures, such as α-helices and β-sheets, are mapped to harmonies. For stable α-helical structures, consonant intervals are used; for the more complex and unstable β-sheets, dissonant intervals are applied.

$$\text{Harmony}_\alpha = \text{Consonant Interval} \qquad \text{and} \qquad \text{Harmony}_\beta = \text{Dissonant Interval} \tag{2}$$

  where Harmony_α and Harmony_β denote the harmonic mappings for α-helices and β-sheets respectively.

- **Mapping Tertiary Structures and Dynamics**: The protein's tertiary structure is represented through rhythmic variations. Stable regions correspond to slow rhythms, whereas dynamic regions are mapped to faster or more complex rhythms.

$$\text{Rhythm}_{\text{stable}} = \text{Slow Rhythm} \quad \text{and} \quad \text{Rhythm}_{\text{dynamic}} = \text{Fast or Complex Rhythm} \tag{3}$$

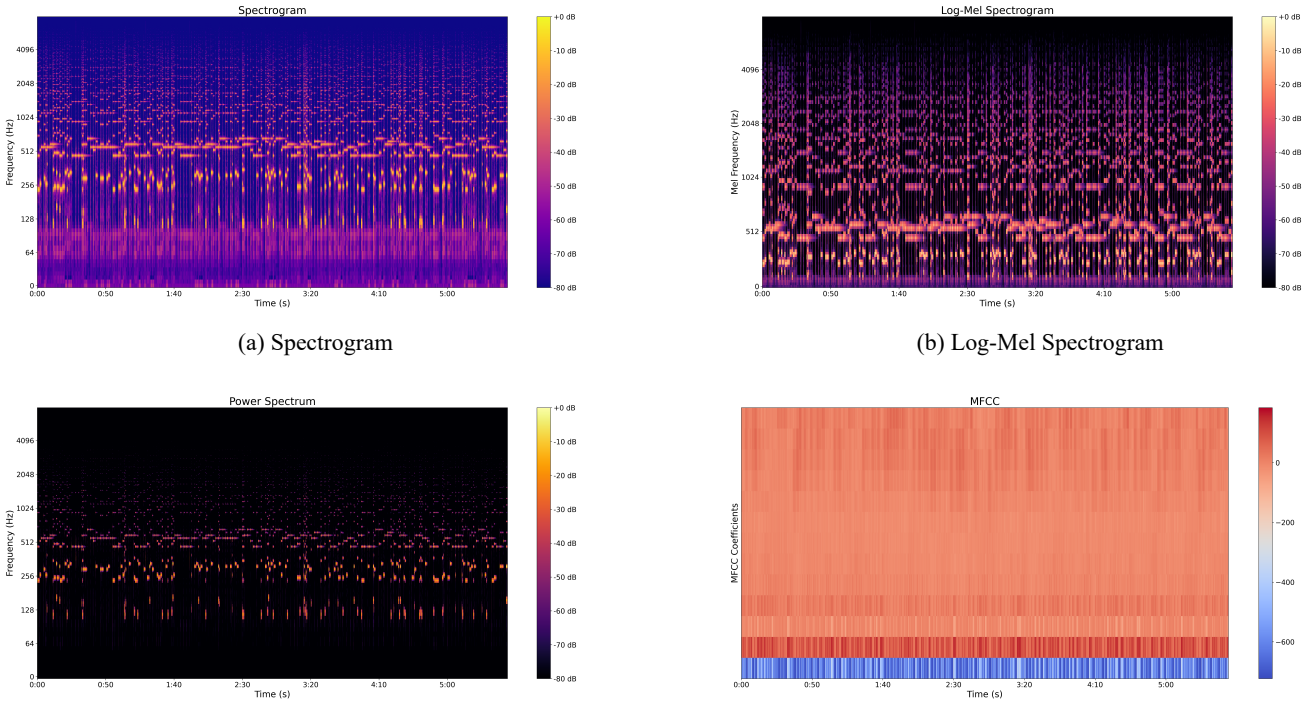
where $\text{Rhythm}_{\text{stable}}$ and $\text{Rhythm}_{\text{dynamic}}$ represent the rhythms for stable and dynamic regions respectively.

B.2 Spectral Feature Extraction

After converting the protein sequences into music, spectral features of the generated musical sequence are extracted using several common spectral analysis methods:

1. **Spectrogram**: Displays the distribution of the signal over time and frequency.
2. **Log-Mel Spectrogram**: Converts the spectrum to the Mel scale to emphasize important frequency features, particularly in the low-frequency range.
3. **Power Spectrum**: Describes the intensity distribution of the signal's frequency components, representing the power or energy distribution across different frequencies.
4. **Mel Frequency Cepstral Coefficients (MFCC)**: Captures the timbre and rhythmic features of audio signals; widely used in speech signal processing.

As illustrated in Figure 3, various spectral analysis graphs corresponding to the protein(6KGA)-to-music conversion are provided.



(a) Spectrogram

(b) Log-Mel Spectrogram

(c) Power Spectrum

(d) MFCC

Figure 3: Schematic Diagram of Spectrum Analysis

The MFCC is computed as follows:

$$\text{MFCC}(S) = \text{MFCC}(S, \text{num\_coefficients} = 13) \tag{4}$$

where S is the log-Mel spectrogram and `num_coefficients` indicates the number of cepstral coefficients extracted.

Four specific proteins were selected for spectral analysis: glucose-6-phosphate dehydrogenase (1DPG), α-lactalbumin (1A4V), outer membrane lipoprotein (1IWM), and sugar-binding protein-A (1AFA), representing various species and functional categories.

Table 2: Comparative Analysis of Frequency-Domain Features

Protein	Mean MFCC	Spectral Centroid (Hz)	Spectral Flatness	Spectral Entropy	Zero-Crossing Rate	Mean Delta MFCC
Glucose-6-phosphate Dehydrogenase	-34.28810	861.10952	0.00007	5.60367	0.06885	-0.00087
α-Lactalbumin	-33.95147	852.43250	0.00006	1.30511	0.06629	-0.00639
Outer Membrane Lipoprotein	-34.13306	872.84715	0.00006	2.14594	0.07173	-0.00451
Mannose-Binding Lectin-A	-38.04973	813.48966	0.00005	2.53378	0.06588	0.00126

B.3 Implementation of Machine Learning Models

In this study, several machine learning models were utilized to learn protein functions from the extracted spectral features. These models include the Multi-Layer Perceptron (MLP), Random Forest, XGBoost, and Support Vector Machine (SVM). The following provides a detailed explanation of each model along with its formula. The basic computational flow is illustrated in Figure 4.

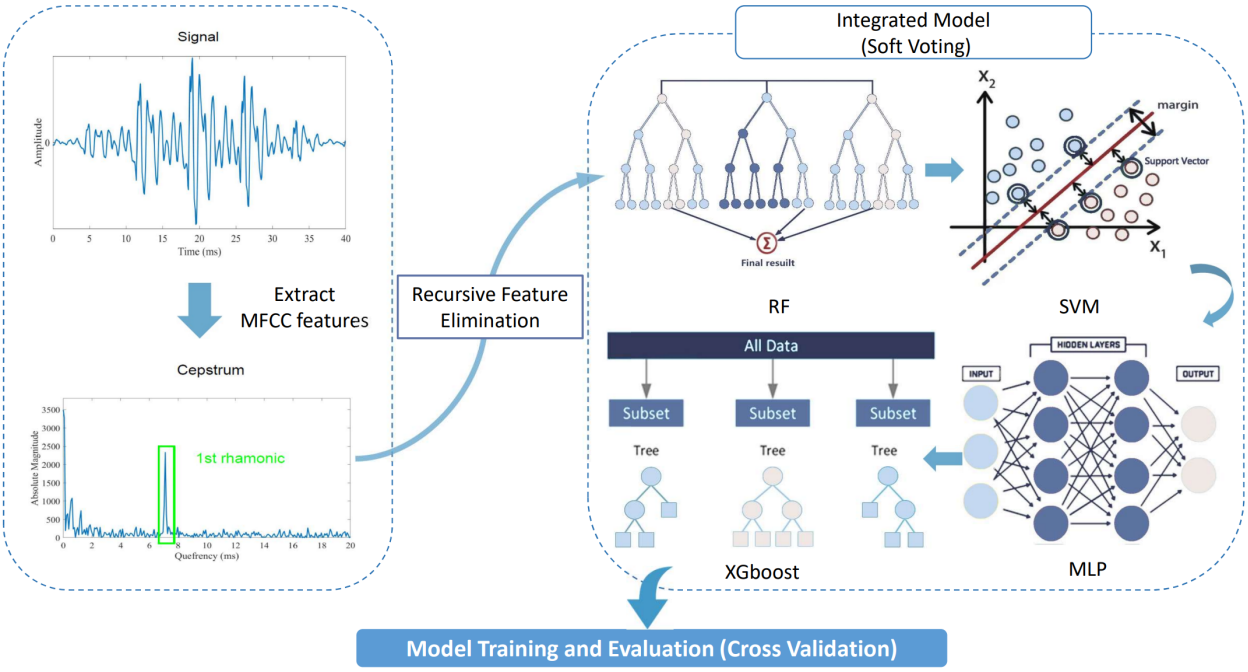


Figure 4: Basic Computational Flow of the Ensemble Process

1. **Multi-Layer Perceptron (MLP)**: MLP is a feedforward neural network that optimizes model weights using backpropagation. Its basic formula is:

$$y = \sigma(Wx + b) \tag{5}$$

where x is the input feature vector, W is the weight matrix, b is the bias term, σ is the activation function, and y is the output.

2. **Random Forest**: An ensemble learning method that constructs multiple decision trees and combines their predictions for classification. Its prediction formula is:

$$\hat{y} = \frac{1}{M} \sum_{i=1}^{M} T_i(x) \tag{6}$$

where M is the number of decision trees, $T_i(x)$ is the prediction of the i-th tree, and $\hat{y}$ is the final prediction.

3. **XGBoost**: Based on Gradient Boosting Trees (GBDT), XGBoost captures the nonlinear relationships among spectral features. Its prediction formula is:

$$\hat{y} = \sum_{k=1}^{K} \alpha_k h_k(x) \tag{7}$$

where K is the number of trees, α_k is the weight of the k-th tree, and $h_k(x)$ is the output of the k-th tree.

4. **Support Vector Machine (SVM)**: SVM constructs an optimal hyperplane to differentiate between classes. Its decision function is:

$$f(x) = w \cdot x + b \tag{8}$$

where w is the weight vector, b is the bias term, x is the input feature vector, and $f(x)$ is the SVM output.

B.4 Spectral Features and Function Prediction

By extracting spectral features and training the machine learning models, we effectively predict protein functions. Techniques such as Recursive Feature Elimination (RFE) and the Synthetic Minority Over-sampling Technique (SMOTE) are used to optimize the models and avoid overfitting.

C Application of Diffusion Generation Model in GFP Directed Evolution

C.1 Principles of the Diffusion Generation Model

The Diffusion Generation Model is a generative model that gradually introduces noise and then reverses the process to guide the generation toward the target distribution. Widely used in image and sequence generation tasks, this model is capable of producing protein sequences that meet target characteristics. In this study, the diffusion model is employed to optimize the Green Fluorescent Protein (GFP) sequence so that it gradually approaches the target emission wavelength and brightness.

1. **Initial Noise Matrix Generation**: The protein sequence is first converted into a one-hot matrix representation, where each amino acid is encoded as a unique vector. Noise is then added to these representations to facilitate gradual optimization during training.

$$X_{\text{noisy}} = \text{OneHot}(x) \tag{9}$$

where X_{noisy} is the initial noise matrix and x is the protein's amino acid sequence.

2. **Introducing Noise and Gradual Adjustment**: The diffusion process generates a new sequence by incrementally introducing noise and adjusting it. At each step, the model dynamically modifies the noise intensity based on the current noise and target characteristics (such as brightness and emission wavelength), guiding the generated sequence closer to the target.

$$\text{condition\_factor} = 1 + \frac{\text{conv\_value} - \text{predicted\_conv}}{100} \tag{10}$$

where conv_value represents the target brightness or emission wavelength, and predicted_conv is the currently predicted value. The condition factor adjusts the noise amplitude to control the generation process.

3. **Generating the Optimized Sequence**: Through iterative adjustment of the noise matrix, a protein sequence that approaches the target brightness and emission wavelength is eventually produced.

$$X_{\text{final}} = \text{Softmax}(X_{\text{adjusted}}) \tag{11}$$

where the Softmax function converts the adjusted noise matrix into a probability distribution to yield the final protein sequence.

C.2 GFP Directed Evolution Process

In the directed evolution of GFP, the diffusion generation model is used to iteratively optimize the GFP amino acid sequence to reach the target emission wavelength and brightness. The core idea of directed evolution is to iteratively generate and select the best sequences to optimize protein functionality. The process is detailed as follows:

1. **Target Fitness Evaluation**: Each generated sequence is evaluated based on the difference between its emission wavelength and brightness compared to the target values. Higher fitness indicates that the sequence's characteristics are closer to the target, guiding the adjustment of the noise matrix.

$$\text{fitness}(x) = \|\text{em\_max}(x) - \text{target\_em\_max}\| + \|\text{brightness}(x) - \text{target\_brightness}\| \quad (12)$$

   where $\text{fitness}(x)$ is the fitness of sequence x, $\text{em\_max}(x)$ is its emission wavelength, target_em_max is the target emission wavelength, $\text{brightness}(x)$ is the brightness, and target_brightness is the target brightness.

2. **Noise Adjustment and Generation of New Sequence**: Based on the fitness evaluation, the model adjusts the noise intensity to generate a new protein sequence. Low fitness leads to increased noise amplitude for more exploration, while high fitness reduces the noise amplitude to allow convergence.

$$X_{\text{adjusted}} = X_{\text{noisy}} \times \text{condition\_factor} \quad (13)$$

3. **Iterative Optimization**: This process is repeated over multiple iterations. In each iteration, the sequence with the highest fitness is chosen as the starting point for the next round. After several iterations, a protein sequence with the target functional characteristics is achieved.

4. **Result Visualization**: To further analyze the characteristics of the generated sequences, dimensionality reduction (using PCA or t-SNE) is performed, and the sequences are visualized in feature space.

$$Z = \text{PCA}(X_{\text{features}}) \quad (14)$$

$$Z = \text{t-SNE}(X_{\text{features}}, \text{perplexity} = 30) \quad (15)$$

   Finally, the generated sequences are mapped onto a 2D space and plotted using Matplotlib:

$$\texttt{plt.scatter}(Z(:,0), Z(:,1), \text{label} = \text{"Generated Sequences"}) \quad (16)$$

D Visualization and Analysis of Generated Sequences

After generating new protein sequences, a detailed analysis is conducted to ensure that the resulting sequences possess the expected functionalities. The following methods are used for dimensionality reduction and visualization:

1. **PCA Dimensionality Reduction**: Principal Component Analysis (PCA) is applied to reduce the dimensionality of the generated protein sequences, extracting the most important features and displaying their distribution in 2D space.

$$Z = \text{PCA}(X_{\text{features}}) \quad (17)$$

2. **t-SNE Dimensionality Reduction**: t-SNE is used for nonlinear dimensionality reduction to further analyze the relationships and diversity among the generated sequences, displaying their distribution in feature space.

$$Z = \text{t-SNE}(X_{\text{features}}, \text{perplexity} = 30) \quad (18)$$

3. **Visualization of Generated Sequences**: The Matplotlib library is utilized to create a 2D scatter plot to observe the distribution and clustering of the generated sequences.

$$\texttt{plt.scatter}(Z(:,0), Z(:,1), \text{label} = \text{"Generated Sequences"}) \quad (19)$$

Through these methods, we gain better insights into the features of the generated protein sequences and can further evaluate their optimization performance. By progressively introducing noise and adjusting its intensity based on fitness, we successfully generated protein sequences that closely match the target functional characteristics. Subsequent analysis using PCA and t-SNE reveals the diversity and effectiveness of the generation process, offering a novel approach to protein design and optimization with significant potential for practical applications.