Supplementary Materials

Table S1. Sample prompts utilized by Team C using GPT-4o OMNI.

Step	*Research Phase*	*Prompt*
1	*Project Initiation*	Hello I am working on this project and I need help with my part. I will upload the project plan for context now. We have three papers (see attached) that describe periodicity in protein/enzyme lengths using Cosine Fourier Transform and SAD. With AI tools now, I vant to see if we uncover periodic structure in eukaryotic enzymes?
2	*Data Validation*	A group member created this CSV using UniProt Advanced Search and filtered by EC number, evidence at protein level, and reviewed (SwissProt) for Eukaryota. I want to make sure this CSV correctly matches the downloaded proteins file I have. I also have t his notebook where I started analyzing the distribution. Help me check if the files align and the fields are correct.
3	*Data Enhancement*	I now want to add the actual amino acid sequences and their lengths. I'll need to go back to UniProt and get those fields. How do I create a file that includes columns like Sequence, Length, and match it using accession numbers?
4	*Filtering Rules*	Either filter the protein sequences by protein names [DE] or by sequences and apply either Rule **1** or Rule **2**. Rule **1**: For any two protein sequences with the same protein names [DE] or sequences, if the length difference is less than 20%, remove the shorter one, else, keep them both. Rule **2**: For protein sequences with the same protein names [DE] or sequences, if the difference of length between the shortest and second shortest is less than 20%, only remove the shorter and keep the rest, else, keep them both and the rest of the protein sequences with the same protein name or sequences.
5	*Visualization*	Help me generate histograms of protein length distributions across original and filtered datasets. Also generate pie charts of the top 10 organisms and top 10 protein names. I want to export all plots as images.
6	*Spectral Analysis*	We want to now test for periodicity using two methods: Cosine Fourier Transform and SAD (Spectral Analysis of Distributions) as done in the original Kolker and Berman papers. Use the filtered datasets and perform both FFT and SAD analysis. Visualize the power spectrum and frequency amplitude plots. Also find the most prominent peak.
7	*Statistical Modeling*	I want to now model the distribution of enzyme lengths using a mixture of a gamma background distribution and several normal peaks centered at integer multiples of a base period. I want to fit this model, estimate confidence intervals using bootstrap, and compare against a background-only model using AIC/BIC and a likelihood ratio test.
8	*Model Evaluation*	Print the final model evaluation summary for each filtered dataset. I want the likelihood ratio test statistics, p-value, AIC/BIC values, and 95% confidence intervals from bootstrap. Confirm that the results are significant and identify the preferred model based on both AIC and BIC. Also print the optimized weights and parameters used.
9	*Summary Generation*	Please come up with a short and well-written executive summary of our enzyme length analysis project that includes all key findings. Mention the hypothesis, data source, methodology (filtering, SAD, Fourier, and model), and key results like the 126 aa periodicity and significance.

Table S2. Sample prompts utilized by Teams A and B using Claude 3.7 Sonnet.

Team	*Prompt*
A	Looking at these instructions, how would you analyze this in google colab?: ASK: Analyze Eukaryotic Enzyme Length Distribution with regards to the potential preferred sizes. 2. READING: Read TWO key articles: https://www.pnas.org/doi/10.1073/pnas.91.9.4044 and https://www.liebertpub.com/doi/10.1089/153623102760092805. 3. DATA-1: Choose database (data preparation: e.g., EBI or NCBI) and compile target data set (data cleaning). 4. DATA-2: Apply Fourier-transform (data processing) and Exploratory Data Analysis (EDA). 5. HYPOTHESIS: Test hypothesis* and formulate conclusion/s. 6. DESCRIPTION: Provide a detailed description of the entire process.
A	Looking at these instructions, how would you analyze this in Google Colab? Please use the diverse enzyme .csv for analysis to start in Google drive. ASK: Analyze Eukaryotic Enzyme Length Distribution with regards to the potential preferred sizes (to generate our own new Fourier transformation equation and use the cosine equation from the paper that got us ~125aa). Apply Fourier-transform (data processing) and Exploratory Data Analysis (EDA). Please provide secondary analysis using updated approaches including Full Fast Fourier Transform (FFT) that captures both amplitude and phase information, Wavelet analysis for multi-scale detection of periodic patterns, and Short-Time Fourier Transform (STFT). Please also provide statistics comparing each analysis with each other and highlight p-values of any positive spectrogram peaks identified.
B	Help me to understand the Spectral Analysis of Protein Distribution using the attached paper as a guide. Create a step-wise approach to understanding this methodology used in the paper and break down each step. Presume that we could use R *tidyverse* and the R 'spectral' package if necessary.
B	As for the critique of the R script, yes. Adjust mixture model complexity, fixed peaks, parameter constraints, p-value interpretation. Window size limitation estimates, refer to the attached paper for guidance and cross-validation. Requesting to keep the Cosine model per the paper. Script Implementation issues aside, for now we need to fix the SAD and statistical modeling first. Regenerate another R script that addresses these issues, while also taking into account the attached. ALSO an after thought for further discussion - is the protein length normally distributed? What's up with the spikes in protein length in the dataset? Should we somehow take this into account?
B	We've created a complete, fully functional R script that performs spectral analysis on protein length distributions. The script follows the methodology from [13] and includes significant improvements to ensure reliability.

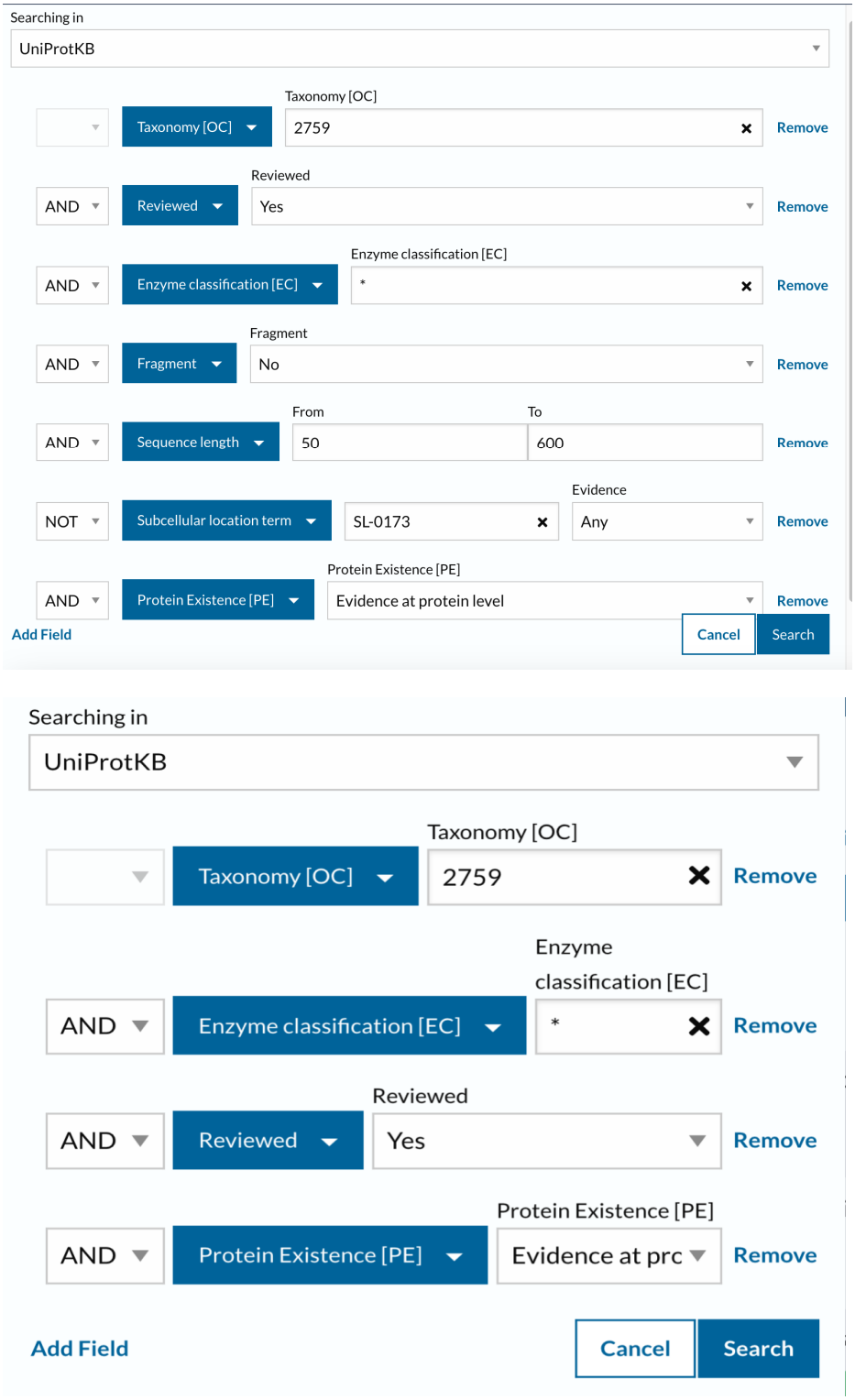


Figure S1. UnitProt APIs used for initial sets. A (Top): Datasets 1 and 2, B (Bottom): Datasets 3 and 4.

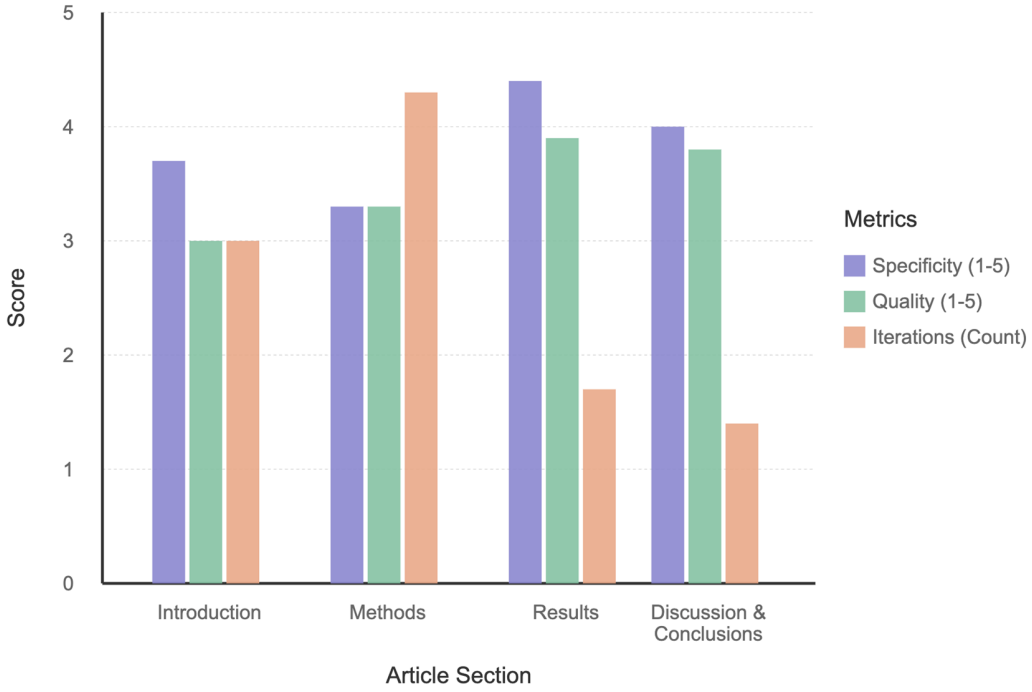


Figure S2. Comparative analysis of AI prompt performance metrics across the article.

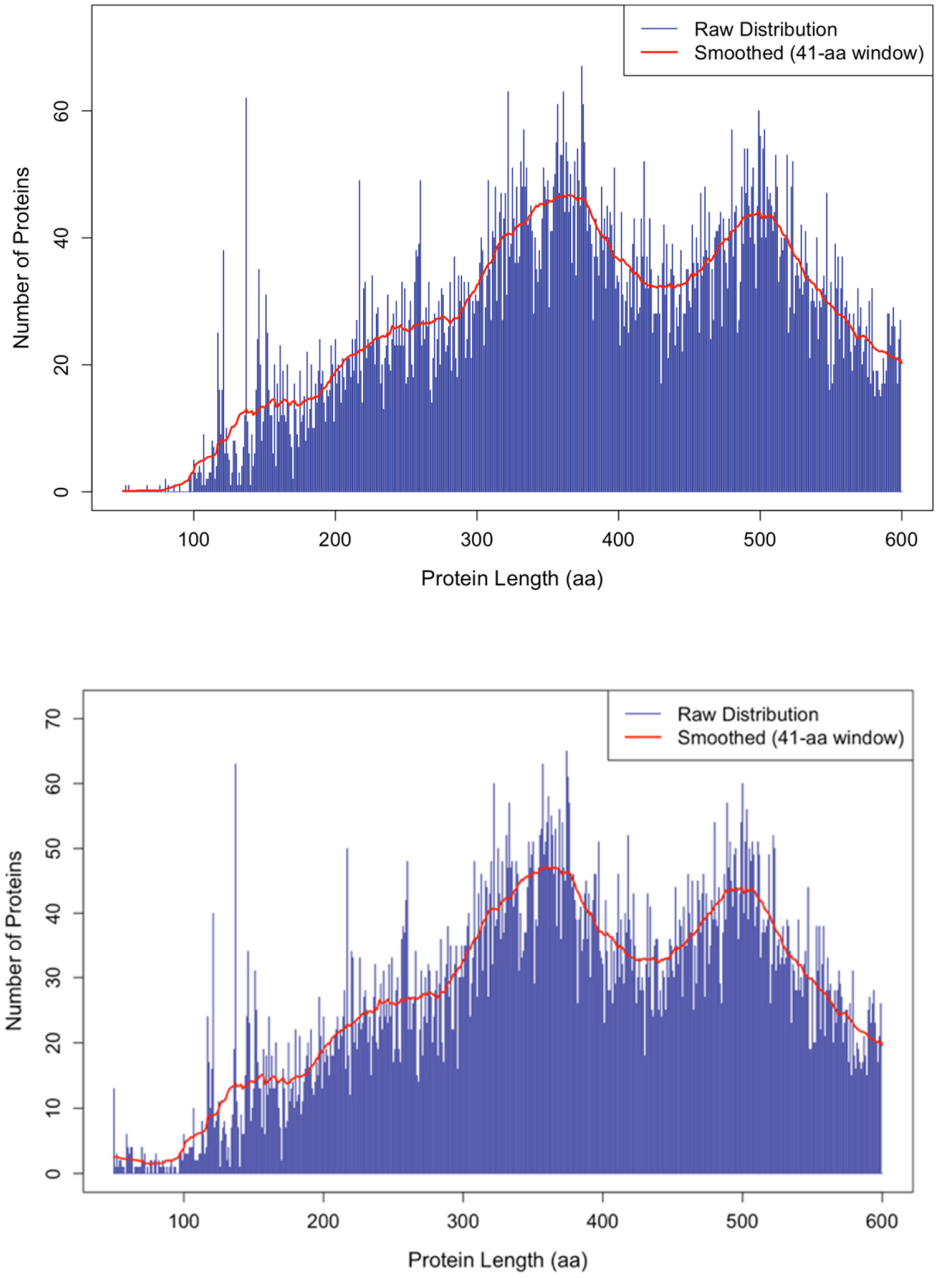


Figure S3. Distributions of eukaryotic enzyme lengths. A (Top): processed Dataset 1, B (Bottom): processed Dataset 3. The histograms correspond to the actual (raw) distributions, and the smoothed curves correspond to the running averages with a 41-aa window.

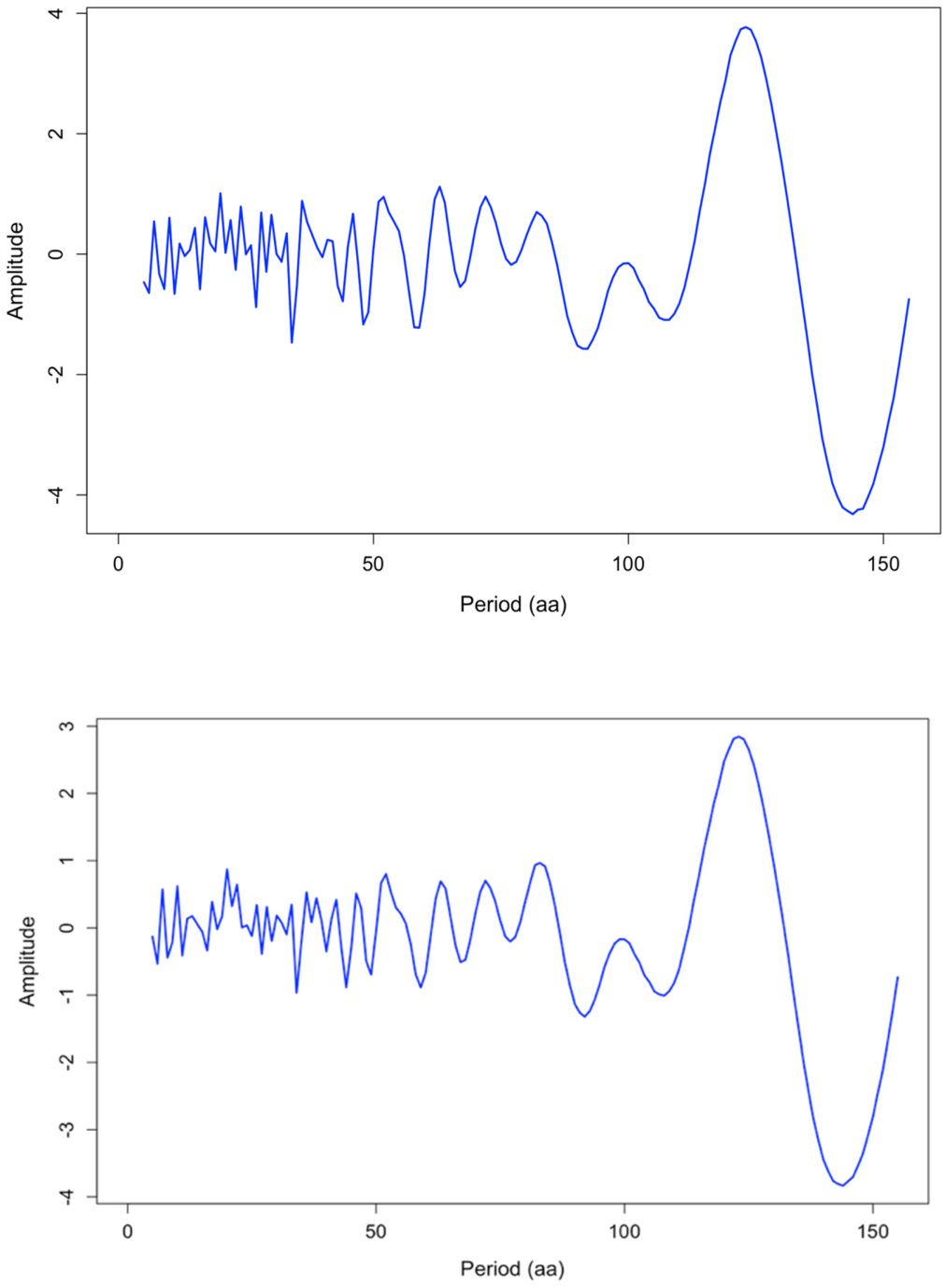
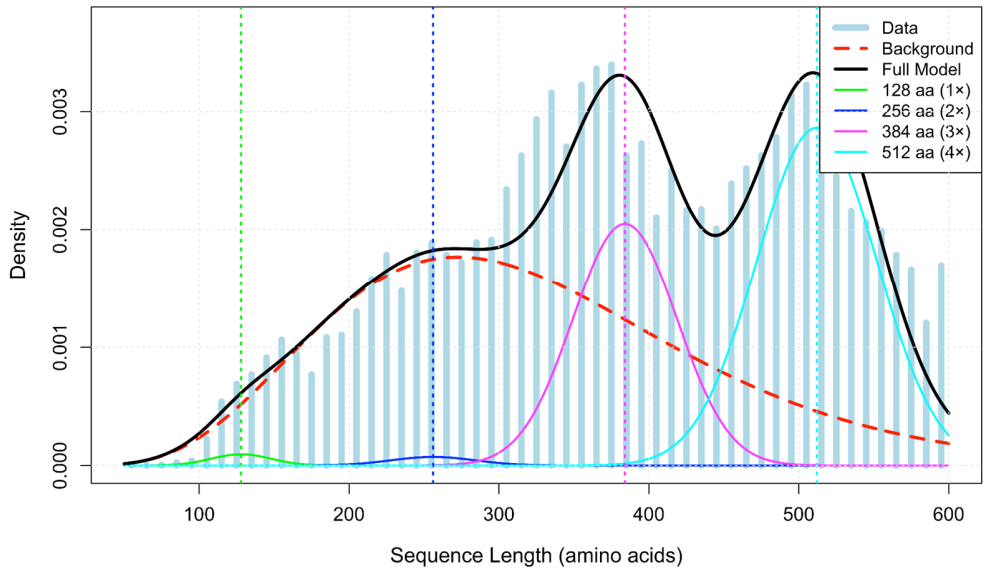


Figure S4. Spectral analysis of eukaryotic enzyme lengths. A (Top): processed Dataset 1, B (Bottom): processed Dataset 3. The cosine spectra generated by SAD achieve maxima amplitude at ~125 aa.



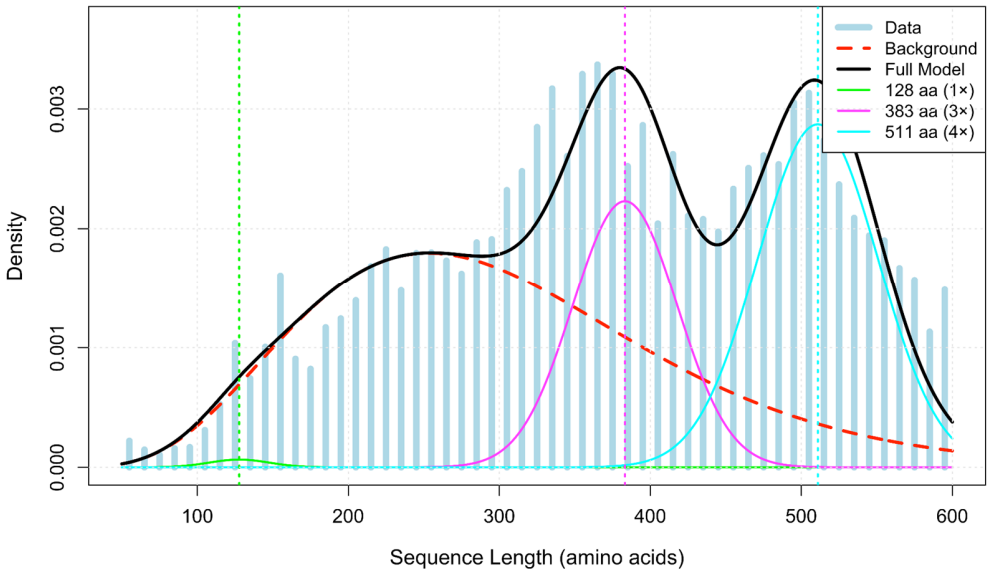


Figure S5. Statistical mixture modeling of enzyme lengths. A (Top): processed Dataset 1, B (Bottom): processed Dataset 3. A. For Dataset 1, the mixture model combines a gamma background distribution (red dashed) and overall model (sum of normal peaks, black) with two small peaks at 1× (green, *p1*) and 2× (blue, *p2*) multiples of the base period μ and two prominent peaks at 3× (pink, *p3*) and 4× (magenta, *p4*) multiples of μ (**Table 5**). **B.** Similarly, Dataset 3 showed one small peak at 1× (green, *p1*) and two prominent peaks at 3× (pink, *p3*) and 4× (magenta, *p4*) multiples of the base period μ (**Table 5**).

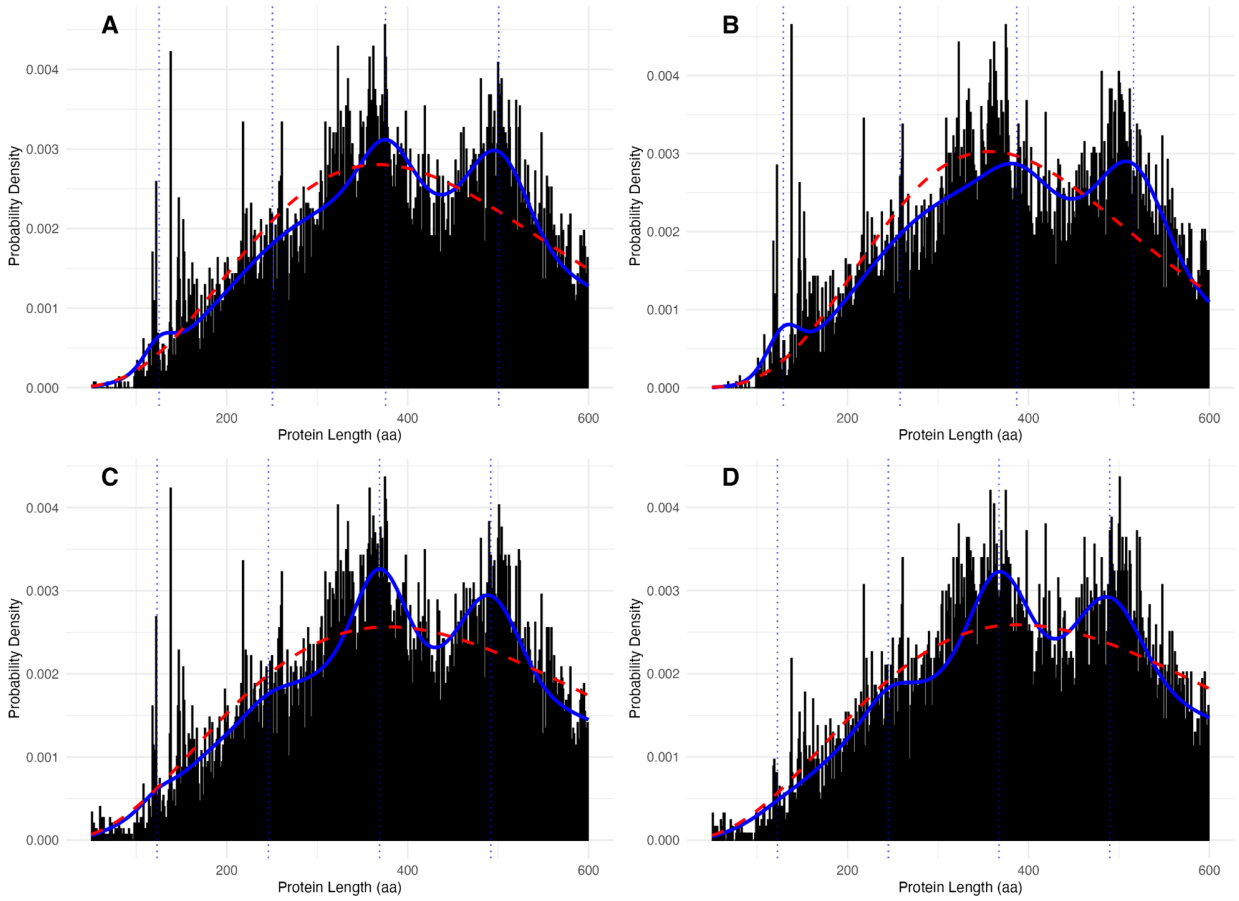


Figure S6. Four processed datasets: Results validation with mixture model analysis. A (Top Left): Dataset 1, B (Top Right): Dataset 2, C (Bottom Left): Dataset 3, and D (Bottom Right): Dataset 4. Each panel displays observed protein length distributions (gray bars), gamma background model (red dashed), and full mixture model combining gamma background with oscillating (periodic normal) components (blue solid). Dotted lines mark peaks at 1×, 2×, 3×, and 4× multiples of the base period μ. Specific estimates include the base period $\mu = 125.21$ aa with the probability $p = 2.36\times10^{-51}$ for Dataset 1, $\mu = 129.04$ aa with $p = 1.31\times10^{-50}$ for Dataset 2, $\mu = 122.98$ aa with $p = 3.05\times10^{-55}$ for Dataset 3, and $\mu = 122.48$ aa with $p = 3.23\times10^{-33}$ for Dataset 4. All four datasets show highly significant oscillating components, rejecting the null hypothesis of the gamma background distributions (**Table 5**).