Table S1. Gen AI Prompts created by Team C using GPT-4o

Step	Research Phase	Prompt Text
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 want 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 this 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	**Periodicity 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 121.4aa periodicity and significance.