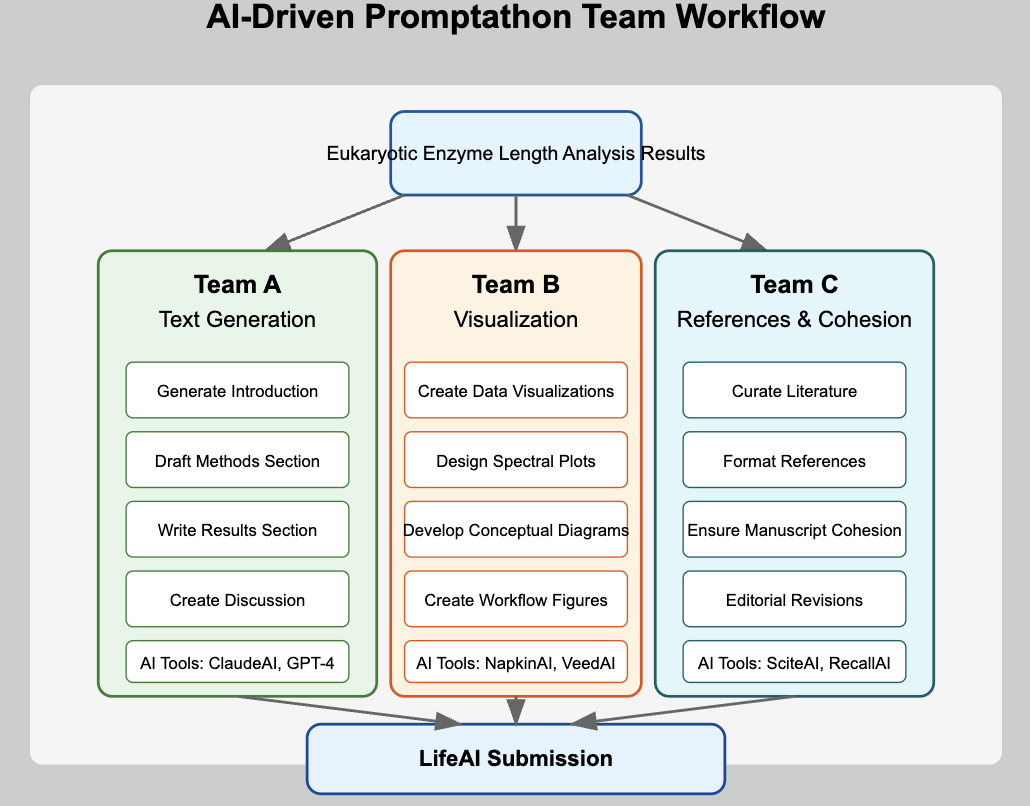
# **AI-Driven Promptathon Framework for Bioinformatics Research Publication**

## **Project Overview**

Inspired by the NYU Langone Health Prompt-a-thon methodology, we'll utilize a three-team approach to transform the eukaryotic enzyme length distribution analysis into a comprehensive research paper for BMC Bioinformatics, leveraging specialized AI tools for text generation, visualizations, and reference management.

## **Submission Timeline**

* Week of April 16-23: Team Setup and Prompt Development
* Apr 21, 2025: Title Due (Brightspace)
* Apr 23, 2025: Individual Team Drafts Due
* Week of April 24-30: Cross Team Content Generation and Initial Integration
* April 30, 2025: Initial Draft Due
* May 7, 2025: Final Draft Due Stretch Goal
* May 16, 2025: Final Draft Due Goal



## **Team Structure and AI Prompting Strategies**

### **Team A: Text Generation**

**Objective**: Develop comprehensive manuscript text using AI tools to articulate the enzyme length periodicity findings

**AI Tools**: ClaudeAI, GPT-4 (Team’s preferences)

**Prompt Types and Measurement Metrics:**

1. **Introduction Generation Prompts** (Success metric: Context quality)
   * Sample: "Generate a 750-word introduction for our paper on eukaryotic enzyme length periodicity that establishes the biological significance of protein length distributions, summarizes current knowledge gaps, and introduces our spectral analysis approach that identified periodicities at ~118.3, ~289.7, and ~490.4 amino acids."
   * Measurement: Citation integration rate, narrative flow score
2. **Methods Description Prompts** (Success metric: Technical precision)
   * Sample: "Create a detailed methods section describing our approach to analyzing enzyme length distributions, including: 1) UniProt data acquisition and filtering criteria, 2) LOWESS and Moving Average smoothing techniques, 3) FFT application with appropriate windowing, and 4) statistical validation using Chi-square tests and Gaussian Mixture Models."
   * Measurement: Reproducibility assessment by independent reviewer
3. **Results Articulation Prompts** (Success metric: Clarity of findings)
   * Sample: "Draft a results section that clearly communicates our major findings: 1) The identification of preferred enzyme lengths at ~118.3, ~289.7, and ~490.4 amino acids, 2) The taxonomic variation in these patterns, and 3) The statistical significance of these periodicities (p < 0.001)."
   * Measurement: Technical accuracy and clarity score
4. **Discussion Drafting Prompts** (Success metric: Interpretative depth)
   * Sample: "Generate a discussion section that interprets our findings in the context of: 1) Evolutionary constraints on protein size, 2) Structural modularity in enzyme design, 3) Comparison with previous studies on domain organization, and 4) Limitations of our spectral analysis approach."
   * Measurement: Insight generation, limitation acknowledgment score

**Deliverables**:

* 20% or less similar (keep only one protein for analysis)
* Complete manuscript text (Introduction, Methods, Results, Discussion, Conclusion)
* Abstract optimized for journal requirements
* Supplementary methods text

### **Team B: Visualization Development**

**Objective**: Create publication-quality visualizations that effectively communicate the enzyme length distributions and spectral findings

**AI Tools**: NapkinAI, VeedAI, ClaudeAI for code generation (Team’s preferences)

**Prompt Types and Measurement Metrics:**

1. **Data Visualization Code Prompts** (Success metric: Code efficiency)
   * Sample: "Generate Python code using matplotlib and seaborn to create a publication-quality figure showing: 1) Histogram of enzyme length distribution (n=3,515) with kernel density overlay, 2) Taxonomic group distributions as small multiples, and 3) Highlighted peaks at the significant periodicities (~118.3, ~289.7, and ~490.4 amino acids)."
   * Measurement: Code functionality, visual clarity rating
2. **Spectral Analysis Visualization Prompts** (Success metric: Technical accuracy)
   * Sample: "Create visualization code for our spectral analysis results showing: 1) Original signal (enzyme length histogram), 2) Smoothed signal after LOWESS filtering, 3) Power spectrum after FFT with annotated peaks, and 4) Statistical significance thresholds at p<0.001, p<0.01, and p<0.05."
   * Measurement: Information density, technical accuracy score
3. **Conceptual Diagram Prompts** (Success metric: Explanatory power)
   * Sample: "Design a conceptual figure illustrating how the identified periodicities (~118.3, ~289.7, and ~490.4 amino acids) might relate to protein domain structure, folding constraints, and evolutionary selection pressures in eukaryotic enzymes."
   * Measurement: Expert comprehension assessment, visual appeal rating
4. **Workflow Diagram Prompts** (Success metric: Process clarity)
   * Sample: "Generate a workflow diagram illustrating our complete analysis pipeline: 1) Data acquisition from UniProt, 2) Filtering and preprocessing steps, 3) Signal processing with smoothing techniques, 4) FFT application and peak detection, and 5) Statistical validation approaches."
   * Measurement: Workflow comprehensibility score

**Deliverables**:

* 5-7 publication-ready figures with high-resolution exports
* Interactive visualization code for supplementary materials
* Comprehensive figure legends
* Graphical abstract for journal submission

### **Team C: References and Cohesion**

**Objective**: Curate comprehensive references and ensure manuscript cohesion through expert editing

**AI Tools**: SciteAI, RecallAI, ClaudeAI for integrative editing (Team’s preferences)

**Prompt Types and Measurement Metrics:**

1. **Literature Search Prompts** (Success metric: Reference quality)
   * Sample: "Identify 40 high-impact, recent (2020-2025) publications relevant to our eukaryotic enzyme length analysis in the following categories: 1) Protein domain organization (10 refs), 2) Spectral analysis methods in bioinformatics (10 refs), 3) Evolutionary constraints on protein size (10 refs), and 4) Structural biology of enzymes (10 refs)."
   * Measurement: Citation impact factor average, recency score
2. **Reference Integration Prompts** (Success metric: Contextual relevance)
   * Sample: "For each section of our manuscript, suggest 5-7 key references that should be cited to support our claims about: 1) The non-random nature of enzyme length distributions, 2) The statistical significance of the periodicities we identified, and 3) The biological implications of preferred enzyme sizes."
   * Measurement: Citation-claim alignment score, reference distribution
3. **Manuscript Cohesion Prompts** (Success metric: Narrative flow)
   * Sample: "Review our current manuscript draft for cohesion issues. Identify any logical gaps, repetitive content, or inconsistent terminology between sections. Suggest specific edits to improve flow between the introduction, methods, results, and discussion sections."
   * Measurement: Before/after readability scores, logical flow assessment
4. **Editorial Enhancement Prompts** (Success metric: Writing quality)
   * Sample: "Enhance the scientific writing quality of our results section by: 1) Improving sentence structure variety, 2) Eliminating redundancies, 3) Strengthening transitions between paragraphs, and 4) Ensuring consistent technical terminology throughout."
   * Measurement: Grammar/style improvement metrics, reviewer clarity rating

**Deliverables**:

* Comprehensive reference list formatted to BMC Bioinformatics standards
* Cross-reference check report ensuring all claims are properly cited
* Editorial revision document with tracked changes
* Manuscript cohesion analysis highlighting improved narrative flow

## **Cross-Team Prompting Strategy**

### **Prompt Engineering Assessment Framework**

For each team, we'll track prompt effectiveness using the following metrics:

1. **Prompt Specificity Score (1-10)**
   * How precisely the prompt targets the required information
   * Example: "Write an introduction" (3) vs. "Write an introduction that establishes the evolutionary context of enzyme length constraints, cites at least 3 landmark studies on protein domain organization, and clearly states our hypothesis about non-random periodicity in eukaryotic enzymes" (9)
2. **AI Response Quality Score (1-10)**
   * Technical accuracy, relevance, and usability of AI output
   * Requires expert validation
3. **Iteration Efficiency**
   * Number of prompt refinements needed to achieve desired output
   * Lower is better
4. **Novel Insight Generation**
   * Count of unexpected but valuable insights provided by AI
   * Qualitative assessment by team leads

### **Sample Specialized Prompts Based on Initial Analysis**

**Team A (Text Generation):**

“Generate a comprehensive methods section (800-1000 words) for our paper on eukaryotic enzyme length periodicity that includes:

1. Data acquisition subsection:

- Detail the UniProt database query parameters

- Describe filtering criteria (enzymes between 50-600 amino acids)

- Explain the deduplication process (18,076 to 3,515 unique proteins)

- Outline taxonomic classification approach across 12 major groups

2. Statistical analysis subsection:

- Describe basic statistical measures (mean: 283.6, median: 273.0)

- Explain non-normality testing (Shapiro-Wilk, p=2.36e-30)

- Detail cross-taxonomic statistical comparisons

3. Spectral analysis subsection:

- Explain transformation of histograms to signals (1-aa bin size)

- Detail smoothing approaches (LOWESS and Moving Average)

- Describe FFT application with appropriate windowing

- Likelihood

- Explain peak detection methodology

4. Hypothesis testing subsection:

- Outline Chi-square test for uniform distribution

- Detail Gaussian Mixture Model analysis identifying 3-5 components

- Explain significance testing of spectral peaks

- Describe cross-validation across taxonomic groups

Use precise technical language appropriate for BMC Bioinformatics while maintaining clarity.”

**Team B (Visualization):**

“Generate Python code for a multi-panel figure that effectively communicates our spectral analysis of eukaryotic enzyme lengths. The figure should include:

1. Panel A: Distribution overview

- Histogram of enzyme lengths (n=3,515) with 1-aa bins

- Kernel density estimate overlay

- Vertical lines marking the significant periodicities at ~118.3, ~289.7, and ~490.4 amino acids

- Color-coded by taxonomic group (12 categories)

2. Panel B: Signal processing visualization

- Raw histogram signal (gray)

- LOWESS-smoothed signal (blue)

- Moving average-smoothed signal (red)

- Annotated with smoothing parameters

3. Panel C: Spectral analysis

- Power spectrum after FFT application

- Highlighted peaks with significance values

- Significance thresholds at p<0.001, p<0.01, and p<0.05

- Labeled periodicity values for top 5 peaks

4. Panel D: Taxonomic comparison

- Small multiples showing length distributions for each taxonomic group

- Consistent x-axis scale across all plots

- Statistical summary (mean, median) for each group

Use matplotlib and seaborn with a publication-ready style (font size 9pt, figure width 7.5 inches, resolution 300 dpi).”

**Team C (References and Cohesion):**

“Create a curated reference collection for our eukaryotic enzyme length periodicity paper with the following specifications:

1. Identify 50 references meeting these criteria:

- 25 high-impact papers (impact factor >5) published since 2020

- 15 seminal papers on protein domain organization regardless of date

- 10 methodological papers on spectral analysis in bioinformatics

2. For each reference, provide:

- Complete citation in BMC Bioinformatics format

- 1-2 sentence summary of key findings relevant to our study

- Specific manuscript section where this reference should be cited

- Relevance score (1-5) to our specific findings

3. Special focus on references addressing:

- Previous studies of protein length distributions

- Structural constraints on enzyme size

- Evolutionary pressures on protein modularity

- Statistical methods for detecting periodicity in biological data

4. Identify any significant gaps in the literature regarding enzyme length periodicity that our paper will address.

Format the output as a structured reference database that can be easily incorporated into the manuscript.”

## **Collaborative Workflow**

1. Weekly prompt optimization meetings
2. Cross-team prompt sharing and refinement
3. Centralized prompt library with effectiveness metrics
4. Weekly integration of AI-generated content

## **Quality Control Process**

1. AI output verification against original data
2. Cross-team peer review of generated content
3. Expert validation of key claims and visualizations
4. Alignment check with journal requirements

## **Success Metrics**

* Prompt efficiency: Average iterations per successful output <3
* Content quality: Expert rating >8/10 for all sections
* Integration coherence: Seamless flow between team contributions
* Publication readiness: Compliance with all BMC Bioinformatics/LifeAI requirements

## **Titles**