

SAD Analysis Pipeline - Quick Start Instructions

✅ What's Included

This bundle contains everything needed to reproduce the Spectral Analysis of Distributions (SAD) results:

- 4 preprocessed protein datasets (dataset_1.csv through dataset_4.csv)
- Complete R Markdown analysis script
- Reference output HTML for comparison

📋 Prerequisites

Ensure you have installed:

- R (version 4.0+)
- RStudio
- Required packages (run this in RStudio):

```
r  
install.packages(c("tidyverse", "knitr", "patchwork"))
```

🚀 Five Steps to Reproduce

Step 1: Extract Files

Extract this ZIP to a folder on your computer (e.g., `Desktop/SAD_Analysis/`)

Step 2: Open RStudio

- Launch RStudio
- Set working directory: `Session → Set Working Directory → Choose Directory`
- Navigate to your extracted folder

Step 3: Open the Analysis Script

- File → Open File → `SCRIPT_SAD_analysis_pipeline.Rmd`

Step 4: Run the Analysis

- Click the **Knit** button (top of script window)
- Or press `Ctrl+Shift+K` (Windows) / `Cmd+Shift+K` (Mac)
- Wait ~10-15 minutes for completion

Step 5: View Results

- Your output: `SCRIPT_SAD_analysis_pipeline.html` (newly generated)
- Reference output: `REFERENCE_OUTPUT_SAD_results.html` (for comparison)

✨ Success Indicators

You'll know it worked when:

- ✓ HTML output file is generated

- ✓ No error messages in RStudio console
 - ✓ Output contains 4 dataset analyses
 - ✓ Graphs display properly in the HTML
-

What You're Looking At

Dataset Summary:

- Dataset 1: 26,486 proteins → 14,685 non-redundant
- Dataset 2: 23,831 proteins → 13,328 non-redundant
- Dataset 3: 27,600 proteins → 14,881 non-redundant
- Dataset 4: 20,252 proteins → 12,378 non-redundant

Key Outputs Per Dataset:

1. Length distribution histogram
 2. Cosine spectrum (periodicity detection)
 3. Statistical mixture model with p-values
-

Troubleshooting

"Package not found" error:

```
r
install.packages("missing_package_name")
```

"File not found" error:

- Ensure all CSV files are in same folder as the .Rmd file
- Check working directory: `getwd()`

Out of memory:

```
r
# Increase memory limit (Windows)
memory.limit(size = 8000)
```

Need Help?

If you encounter issues not covered here, document:

1. The error message
 2. Your R version: `R.version`
 3. Your operating system
-

Optional: Deep Dive

For complete methodology including Clusterize preprocessing:

- See `OVERVIEW_interactive_methodology.html` in main repository
- Full reproducibility guide available in parent folder

Quick Start Bundle v1.0 - September 2025

Estimated time: 20-30 minutes