

SAD Reproducibility Guide - September 2025

Spectral Analysis of Distributions: Multi-Dataset Pipeline

Three Ways to Review

OPTION A: Full Verification with Preprocessing

For reviewers who want to recreate Dataset 4 from scratch

Part 1: Review Methodology

Step 1: Open ([00_OVERVIEW_interactive_methodology.html](#)) in browser

Step 2: Review Dataset 3 → Dataset 4 transformation process

Part 2: Run Clusterize Preprocessing

Step 1: Navigate to ([01_Clusterize_Preprocessing/](#))

Step 2: Download these files to a local folder:

- ([INPUT_uniprot_27600_proteins.csv](#)) (Dataset 3 source originally from Reviewer 3)
- ([SCRIPT_clusterize_85percent.R](#))

Step 3: Open RStudio and set working directory to your folder

Step 4: Open and run ([SCRIPT_clusterize_85percent.R](#))

- This will generate multiple output files
- The filtered dataset is specifically: ([filtered_enzyme_dataset.csv](#))

Step 5: Verify output file is created (~20,252 proteins)

- Look for ([filtered_enzyme_dataset.csv](#)) in your working directory
- This file will be renamed to ([dataset_4.csv](#)) for the main analysis

Step 6: Compare with provided ([OUTPUT_dataset4_20252_proteins.csv](#))

- These should be identical (both are the filtered enzyme dataset)

Part 3: Run Main Analysis

Step 7: Navigate to ([02_Main_Analysis/](#))

Step 8: Download all datasets and the analysis script

Step 9: Run ([SCRIPT_SAD_analysis_pipeline.Rmd](#)) in RStudio

Step 10: Compare your output with reference HTML

Note: Clusterize output is renamed 'dataset_4' in the Main Analysis pipeline

OPTION B: Quick Start Bundle - RECOMMENDED

The easiest path for most reviewers

Step 1: Navigate to ([03_Quick_Start_Bundle/](#))

Step 2: Download ([Quick_Start_Bundle.zip](#))

Step 3: Extract ZIP to a folder on your computer

Step 4: Open RStudio and set working directory to extracted folder

Step 5: Open `SCRIPT_SAD_analysis_pipeline.Rmd`

Step 6: Click Knit button (or press Ctrl/Cmd+Shift+K)

Step 7: Wait 10-15 minutes for analysis to complete

Step 8: View your generated HTML output

Why this is recommended:

- One download contains everything needed
- No complex folder navigation
- Pre-organized file structure
- Includes simple instructions
- Fastest path to verification

OPTION C: Browse Results Only

For reviewers who just want to see the outputs

No coding required! Simply view:

Step 1: Open `00_OVERVIEW_interactive_methodology.html` in your browser

- Interactive methodology explorer with clickable elements

Step 2: Open `02_Main_Analysis/REFERENCE_OUTPUT_SAD_results.html`

- Complete analysis results for all 4 datasets

Repository Structure

```
SAD_Reproducibility_Sept2025/
|
|  └── README_START_HERE.pdf (this file)
|  └── 00_OVERVIEW_interactive_methodology.html
|
|  └── 01_Clusterize_Preprocessing/
|      ├── INPUT_uniprot_27600_proteins.csv
|      ├── SCRIPT_clusterize_85percent.R
|      └── OUTPUT_dataset4_20252_proteins.csv
|
|  └── 02_Main_Analysis/
|      ├── dataset_1.csv
|      ├── dataset_2.csv
|      ├── dataset_3.csv
|      ├── dataset_4.csv
|      ├── SCRIPT_SAD_analysis_pipeline.Rmd
|      └── REFERENCE_OUTPUT_SAD_results.html
|
|  └── 03_Quick_Start_Bundle/
|      └── Quick_Start_Bundle.zip
```

System Requirements

Essential Software

- R version 4.0 or higher
- RStudio (recommended IDE)

Required R Packages

```
r

# Core packages
install.packages(c("tidyverse", "knitr", "patchwork"))

# For Clusterize preprocessing (Pathway B only)
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("DECIPHER")
```

Input Datasets Summary

Dataset	Initial Proteins	Non-redundant	Reduction
Dataset 1	26,486	14,685	44.5%
Dataset 2	23,831	13,328	44.0%
Dataset 3	27,600	14,881	46.1%
Dataset 4	20,252	12,378	38.9%

Processing Stages for Each Dataset:

1. **Data Standardization:** Harmonizing column names across different UniProt data sources
2. **Non-redundancy Processing:** Protein name-based deduplication (retaining longest variant per Kolker et al. 2002)
3. **Length Filtering:** Restricting to 50-600 amino acids range
4. **SAD Analysis:** Spectral Analysis of Distributions using cosine transform (Kolker et al. 2002)
5. **Statistical Mixture Modeling** (Team C Extension): Maximum likelihood fitting of gamma background + 4 normal peaks
6. **Likelihood Ratio Testing:** Statistical significance via χ^2 test with 6 degrees of freedom
7. **Bootstrap Analysis:** 100 iterations for 95% confidence intervals
8. **Visualization Generation:** Length distributions, cosine spectra, and probability density plots

Troubleshooting

Common Issues

"Package not found" error:

```
r

install.packages("missing_package_name")
# or for Bioconductor packages:
BiocManager::install("package_name")
```

"Cannot open file" error:

- Check working directory: `getwd()`
- Ensure all files are in same folder
- Check file names match exactly (case-sensitive)

Memory issues with large datasets:






```
r  
  
# Windows  
memory.limit(size = 8000)  
  
# Mac/Linux  
options(java.parameters = "-Xmx8g")
```

Verification Checklist

- ☐ R and RStudio installed
- ☐ Required packages installed
- ☐ Files downloaded to local folder
- ☐ Working directory set correctly
- ☐ Script runs without errors
- ☐ HTML output generated
- ☐ Results match expected patterns

Success Indicators

Your analysis is successful when:

-  All 4 datasets process without errors
-  HTML report generates with all plots visible
-  Periodicity detected around 115-125 amino acids
-  Statistical tests show significant p-values for datasets 3 & 4
-  Your results broadly match the reference output

Reporting Issues

If you encounter problems not covered here, please document:

1. Error message (exact text)
2. Step where error occurred
3. Your system info: `sessionInfo()`
4. Operating system and R version

Additional Resources

- **Original paper:** Kolker et al. OMICS 2002
- **Methodology details:** See `00_OVERVIEW_interactive_methodology.html`
- **Extended analysis:** Team C statistical extensions in main output

Estimated time varies by pathway chosen