# GEMspa Analysis Pipeline Guide

## *Quickrun*

cd /Users/andrewbazley/Desktop/GEMspa-CLI

python3 -m venv venv && source venv/bin/activate

pip install -r requirements.txt

pip install -e .

./venv/bin/GEMspa-CLI.py --work-dir "/Users/andrewbazley/Desktop/20250716 RLTPSC ER GEM"

## 1. Introduction

This document describes the components of the GEMspa single-particle tracking (SPT) analysis pipeline, explains what each script does, and shows how to run the entire analysis from the CLI. It also explains the outputs generated and their interpretation.

## 2. Script Components

### 2.1 trajectory\_analysis.py

Per-replicate MSD and diffusion coefficient analysis:  
- Reads a Traj\_<condition>\_<rep>.csv file containing tracked x,y coordinates and frame numbers.  
- Computes mean-squared displacement (MSD) per track (Numba JIT accelerated) and fits to MSD = 4·D·t^α.  
- Saves `msd\_results.csv` with columns: track\_id, condition, D\_fit, alpha\_fit, r2\_fit.  
- Plots:  
 • D\_fit\_distribution.png: log‑spaced histogram of D (μm²/s).  
 • alpha\_vs\_logD.png: scatter of α vs. log10(D).  
- Optionally overlays colored tracks on the MAX\_\*.tif image using rainbow colormap.

### 2.2 msd\_diffusion.py

Core MSD computation and fitting utilities:  
- `\_msd2d\_jit`: fast Numba‐parallel function computing 2D MSD up to max lag.  
- `fit\_msd`: non‐linear least squares (SciPy) to fit MSD to power-law, returns D, α, R².  
- `fit\_msd\_linear`: fallback linear fit for purely diffusive tracks.  
- Step‐size export:  
 • `set\_track\_data` and `step\_sizes\_and\_angles` compute step‐size matrices and angles.  
 • `save\_step\_sizes` writes a “long” table with tlag & step\_size columns.

### 2.3 rainbow\_tracks.py

Overlay raw tracks on background TIFF images, color‐coded by diffusion coefficient:  
- Reads MAX\_<condition>\_<rep>.tif, converts to RGB canvas.  
- Normalizes each track’s D\_fit to [min\_D, max\_D], maps to specified colormap.  
- Draws line segments for each track on the image, saves high-resolution PNG.

### 2.4 step\_size\_analysis.py

Per-replicate and per-ensemble step-size analysis:  
- Expects `all\_data\_step\_sizes.txt` with columns [group, tlag, step\_size].  
- For each group and each tlag, plots log‑scale KDE of step\_size:  
 • step\_kde\_<group>.png (fade by tlag, inset α₂ parameter).  
- If two groups present, computes KS test per tlag and plots volcano plot:  
 • ks\_volcano\_<group1>\_vs\_<group2>.png

### 2.5 ensemble\_analysis.py

Pools replicate `msd\_results.csv` by condition and applies filtering:  
- Groups folders named <condition>\_<rep> and concatenates their x `msd\_results.csv`.  
- Writes `grouped\_raw/msd\_results.csv` and `grouped\_filtered/msd\_results.csv`.  
- Generates the same distribution & scatter plots for raw and filtered ensembles.

### 2.6 compare\_conditions.py

Cross-condition comparison on filtered ensemble data:  
- Overlaid, density-normalized histograms of D\_fit on log x-axis with mean lines and KS-test asterisks.  
- Overlaid, density-normalized histograms of α\_fit on linear x-axis with mean lines and KS-test asterisks.  
- Boxplot of replicate median D\_fit with jittered points and Mann–Whitney U (or KS) test asterisk.

### 2.7 GEMspa-CLI.py

Command-line entry point gluing everything together:  
- Discovers Traj\_\*.csv in --work-dir; processes each in parallel (--n-jobs).  
- Runs trajectory\_analysis, exports step sizes, runs step\_size\_analysis if requested.  
- After replicates, runs ensemble\_analysis and compare\_conditions.  
- Arguments control filtering, rainbow parameters, and parallelism.

## 3. CLI Usage

From your activated virtual environment, install and run:  
```  
pip install -e .  
gemspa-cli \  
 --work-dir /path/to/data \  
 --n-jobs 4 \  
 --rainbow-tracks \  
 --rainbow-min-D 0 \  
 --rainbow-max-D 2 \  
 --rainbow-colormap plasma \  
 --rainbow-scale 2 \  
 --rainbow-dpi 300 \  
 --step-size-analysis  
```

## 4. Outputs & Interpretation

- \*\*msd\_results.csv\*\*: per-track diffusion (D), anomalous exponent (α), fit quality (R²).  
- \*\*D\_fit\_distribution.png\*\*: shows spread of diffusion coefficients on log scale.  
- \*\*alpha\_vs\_logD.png\*\*: relation between α and D across tracks.  
- \*\*rainbow\_tracks.png\*\*: raw image with tracks color-coded by D.  
- \*\*all\_data\_step\_sizes.txt\*\*: long-form step-size data (group, tlag, step\_size).  
- \*\*step\_kde\_<group>.png\*\*: KDE curves of step sizes per tlag, log‐y.  
- \*\*ks\_volcano\_\*.png\*\*: p-value vs. tlag comparison between two groups.  
- \*\*grouped\_raw/msd\_results.csv\*\* & plots: pooled per-condition before filtering.  
- \*\*grouped\_filtered/msd\_results.csv\*\* & plots: pooled per-condition within filter bounds.  
- \*\*ensemble\_filtered\_D\_histograms.png\*\* & \*\*ensemble\_filtered\_alpha\_histograms.png\*\*: overlaid histograms comparing conditions.  
- \*\*replicate\_median\_D\_boxplot.png\*\*: boxplot of median D per replicate with significance.

Interpretation:  
- Low D\_fit values indicate confined or slow diffusion; high values indicate more mobile particles.  
- α < 1 indicates subdiffusion; α ≈ 1 indicates normal diffusion; α > 1 indicates superdiffusion.  
- Rainbow overlays let you visually link track mobility to spatial location.  
- Step-size KDE highlights heterogeneity in displacements across time scales.  
- KS‐volcano plots show which lags differ significantly between conditions.  
- Boxplots summarize inter-replicate consistency and statistical differences.