Cancer Cell Invasion Analysis Project Documentation

1. Biological Background: MDA-MB-231 Cell Line

MDA-MB-231 is a *triple-negative human breast carcinoma* cell line (ER^-, PR^-, $HER2^-$).
It's one of the most commonly used models for studying **invasion and metastasis**, because these cells:

- Are **mesenchymal-like** — highly motile, elongated morphology.
- Exhibit **3D invasion** when embedded in ECM (like collagen I or Matrigel).
- Often form **collective invasion streams** or finger-like protrusions from a spheroid core.

In your dataset, these cells were genetically engineered with a **pMSCV-GFP vector** so that:

- Each cell stably expresses **green fluorescent protein (GFP)**.
- They can be visualized using **confocal fluorescence microscopy** in 3D.

So, biologically, this model mimics **breast tumor cell invasion into the surrounding extracellular matrix**.

🔬 2. Experimental Setup (from R. Kamm Lab, MIT)

Feature	Description
Origin	Dr. Roger D. Kamm's lab, Dept. of Biological Engineering, MIT (USA)
Cell type	MDA-MB-231 human breast carcinoma cells

Transfection	pMSCV vector containing GFP sequence
Matrix	Collagen type I gel — mimicking ECM
Microscope	Olympus FluoView F1000 confocal
Objective lens	Plan 20×, NA 0.7
Voxel size	1.242 × 1.242 × 6.0 μm (anisotropic — z-step much coarser)
Time step	80 minutes between frames
Total timepoints	10–12 (depending on sequence)
Data type	3D time-lapse fluorescence stacks (grayscale intensity)
Ground truth (train)	Expert-labeled segmentation and tracking files
Ground truth (test)	No labels (for benchmarking algorithm accuracy)

The dataset is part of the **Cell Tracking Challenge (CTC)** repository, used for testing 3D segmentation and tracking algorithms across diverse biological systems.

💾 3. Dataset Structure

Once unzipped, the folders look like this:

```
Fluo-C3DL-MDA231/
│
├── 01/                 # Sample 1
│   ├── t000.tif
│   ├── t001.tif
│   ├── ...
│
├── 02/                 # Sample 2
│   ├── t000.tif
│   ├── ...
│
├── GT/                 # Ground truth annotations
│   ├── SEG/            # Binary masks (cell segmentation)
│   ├── TRA/            # Tracking data (track IDs per cell)
```

```
│
└── README.txt          # Imaging parameters and metadata
```

Each `.tif` file is a **3D z-stack** for one timepoint, with GFP intensity values corresponding to cell fluorescence.

🧩 4. What Your Project Does

Your project — the **3D Invasion Analysis Pipeline** — builds on this dataset to quantify *collective invasion metrics* from the segmentation outputs.

Let's break down its logic.

🧱 Step 1: Data Loading

You used two CSVs (`full_segmentation_features.csv` and `test_segmentation_features.csv`) derived from segmentation results — likely generated by **Cellpose3D** or **StarDist3D**.
Each row represents one detected cell nucleus or cytoplasm, with:

- Centroid coordinates (`centroid-0`, `centroid-1`, `centroid-2`)
- Timepoint
- Sample ID
- Morphological features (volume, area, etc.)

You label them as "train" or "test" to keep them organized.

🧮 Step 2: Compute Invasion Metrics (corrected version)

For each sample and timepoint, your script computes:

Metric	Description	Biological meaning

Mean Radius	Average distance from spheroid center	How far the bulk of the population has invaded
Median / 90th percentile radius	Distribution spread	Outer invasion front
Max radius	Farthest cell	Leading edge
Leader fraction	Top 10% farthest cells	Fraction of highly motile/invasive cells
Leader cell invasion depth	Distance of leader cells	How deep leaders penetrate the matrix
Nearest neighbor spacing	Mean cell–cell distance	Degree of dispersion or compaction
Cell density	Cells per unit volume	Population growth or compaction
Dispersion index	Standard deviation / mean distance	Heterogeneity of spread
Skewness	Asymmetry of distance distribution	Whether invasion is front-driven

Critical fix:
You now define a *fixed center of invasion* (based on t=0 centroid) and apply voxel scaling, so distances are in **µm**, not pixels.
This makes your measurements biologically accurate.

🎥 Step 3: Visualization

Your visualization script does two things:

1. Generates **3D scatter plots** (cells as dots, red dot for spheroid mean center) per timepoint and sample.

2. Combines these into **time-lapse videos** to show whether the population expands or stays compact.

📊 Step 4: Invasion Trend Analysis

You plot multiple metrics over time:

- If the invasion radius or leader depth **increases**, cells are migrating.
- If all metrics are **flat or oscillating**, invasion is negligible.

Your corrected figure shows that **radii, dispersion, and density remain nearly constant**, so your conclusion — *no strong invasion behavior detected* — is supported by the quantitative analysis.

🧬 5. Biological vs. Computational Insights

Aspect	Biological meaning	Computational reflection
Spheroid remains compact	Cells not invading the ECM	Mean radius stays flat
No directional protrusions	No leader-front formation	Skewness stays low
Only small local motion	Minor centroid fluctuations	High-frequency oscillations in leader fraction
Stable population	No proliferation or apoptosis	Constant cell counts

So computationally and biologically, your pipeline correctly detects a *non-invasive or weakly motile state* of the population during the imaging window.

🧠 6. Why This Project Is Valuable

Your pipeline does something that many standard tracking tools don't:
It converts raw segmentation outputs into **quantitative invasion dynamics metrics**, enabling:

- Objective comparison between samples or treatments
- Validation of model invasiveness
- Benchmarking of segmentation/tracking performance in 3D

You're effectively replicating (and improving) the analysis approach used in **Kamm et al.'s collective invasion models** — but in a reproducible, Python-based, data-driven way.

🗒️ 7. In summary

Component	Description
Dataset	3D time-lapse GFP fluorescence images of invasive breast cancer cells (MDA-MB-231) in collagen matrix
Source	Dr. R. Kamm, MIT, via Cell Tracking Challenge (Fluo-C3DL-MDA231)
Goal	Quantify invasion metrics (radius, leader fraction, dispersion, etc.) over time
Outcome	Current dataset shows minimal invasion during imaging; metrics stable
Contribution	Pipeline allows automated, quantitative assessment of collective cell invasion from 3D segmentation results