🧬 Alignment of S1 and S2 Between Reports

Metric / Behavior	Processing_features.py Interpretation	Quantitative_invasion_analysis.py Interpretation	Conclusion: Matched / Divergent
S1 / Sample 1	*Aggressive dispersive invasion*: large radius (~900 µm), strong leader population, low density, high invasion depth	*Burst-type invasive*: high anisotropy, cyclic morphological oscillations, lower density, rapid leader-driven advance	✅ Matched — both describe S1 as highly invasive, leader-driven, low-density, burst-like
S2 / Sample 2	*Heterogeneous core-leader invasion*: small mean radius (~570 µm), high dispersion/skewness, compact core with few long leaders	*Coordinated growth*: high density, moderate anisotropy, proliferation-dominant invasion	✅ Matched — both describe S2 as compact, less migratory, proliferation-focused
Leader Cell Fraction	~10–12% leaders, stable across samples	10–11% leaders, stable across timepoints	✅ Perfect numerical and conceptual match
Cell Density	Low in S1 (migration-dominant), high in S2 (proliferation-dominant)	S1: low baseline density, S2: higher density and proliferation	✅ Matched
Invasion Radius	S1 larger (~900 µm), S2 smaller (~570 µm)	S1 expands slightly, S2 stable or shrinking	✅ Matched
Directional Behavior	Not explicitly measured, inferred via dispersion and skewness	Quantified explicitly as "anisotropy" (S1 ≫ S2)	✅ Complementary — confirms implied directional invasion in S1
Morphological Behavior	Described indirectly via dispersion and skewness	Explicit compactness–distance correlation oscillations (cyclic invasion–pause)	✅ Complementary — same biological basis, now quantified temporally
Proliferation vs Migration Trade-off	Central conclusion (Train S2 proliferation-dominant vs Test S1 migration-dominant)	Central conclusion (S2 proliferative, S1 migratory)	✅ Identical biological interpretation
Invasiveness Ranking	Test S1 > Train S1 ≈ Train S2 ↓ t S2	Sample 1 > Sample 2 (only two replicates)	✅ Matched hierarchy when mapped to S1/S2 equivalents

Detailed Interpretation of PCA and Clustering Analysis

Overview

Your analysis used 16 quantitative invasion metrics extracted from MDA-MB-231 cell populations embedded in collagen matrix to perform dimensionality reduction (PCA) and unsupervised clustering. This comprehensive approach reveals latent phenotypic structure in your dataset beyond what individual metrics can show.

Panel-by-Panel Breakdown

1. PCA: By Condition (Top Left)

What it shows:

- Each point represents one sample at one timepoint (48 total: 4 conditions × 12 timepoints)
- Colors represent experimental groups:
    - Blue (train_S1) and Green (train_S2): Training dataset samples
    - Red (test_S2) and Light green (test_S1): Test dataset samples

Key observations:

- Strong left-right separation along PC1 (X-axis, 40.9% variance): Train samples cluster on the left (negative PC1), test samples on the right (positive PC1)
- Vertical spread along PC2 (Y-axis, 25.9% variance): Captures additional within-group variation

- Clear clustering by condition: Groups are well-separated, indicating reproducible batch/sample-level differences

Biological interpretation:

- PC1 captures the dominant axis of invasion phenotype variation—likely related to overall spatial spread, radius, and leader cell behavior
- The separation suggests that train vs test, or Sample 1 vs Sample 2, have fundamentally different baseline invasion profiles
- This could reflect experimental batch effects, initial spheroid size differences, or genuine biological variability between replicates

---

2. PCA: K-Means Clusters (Top Center)

What it shows:

- Same PCA projection, but colored by K-means cluster assignment (4 clusters: 0-3)
- Clustering was performed on standardized feature space, not just PC1/PC2

Key observations:

- Cluster 0 (dark blue): Concentrated in lower-left quadrant—predominantly train samples
- Cluster 1 (teal/green): Middle-left and center—mixed train/test
- Cluster 2 (yellow): Upper region—primarily test_S1 early timepoints
- Cluster 3 (purple): Right side—test_S2 samples

Biological interpretation:

- Unsupervised clustering naturally separates samples into phenotypic subtypes without using condition labels
- These clusters likely represent distinct invasion modes:
    - Some clusters show compact, non-invasive phenotypes
    - Others show higher spatial spread (though still limited overall invasion based on your prior analysis)
- The alignment with experimental conditions suggests systematic differences between samples, validating both the PCA and clustering approaches

---

3. Scree Plot (Top Right)

What it shows:

- Blue line: Variance explained by each individual principal component
- Red line: Cumulative variance explained

Key observations:

- PC1 explains ~41% of total variance
- PC2 explains ~26%
- Together, PC1 + PC2 capture ~67% of total dataset variation
- Diminishing returns after PC3 (each subsequent component explains <10%)

Interpretation:

- Most meaningful biological signal is captured in the first 2-3 dimensions
- The "elbow" around PC3-4 suggests these are the dimensions needed for complete representation
- This validates focusing analyses on PC1 and PC2 for visualization and interpretation

---

4. Feature Loadings Heatmap (Bottom Left)

What it shows:

- Each row is one of the 16 invasion metrics
- Each column is a principal component (PC1, PC2, PC3)
- Color intensity shows the loading (contribution) of each feature to each PC
    - Red (positive): Feature increases with PC value
    - Blue (negative): Feature decreases with PC value

Key observations:

PC1 (strongest contributors):

- Positive (red): max_radius, mean_leader_radius, p90_radius, n_leader_cells
    - These are all "invasion-like" features (spatial spread, leader cell emergence)
- Negative (blue): cell_density, cell_count, mean_extent
    - These represent compact, dense clusters

PC2:

- Positive: dispersion_index, radius_skewness
- Negative: mean_nn_dist, median_nn_dist
    - Captures local cell organization and spatial distribution irregularity

PC3:

- More complex mix, capturing secondary morphological features

Biological interpretation:

- PC1 is the "invasion axis": High PC1 = large radius, many leader cells, low density (more invasive-like). Low PC1 = compact, high density (non-invasive)
- PC2 is the "spatial organization axis": Distinguishes uniform vs. heterogeneous cell distributions
- Your data shows that train samples have lower PC1 (more compact), while test samples have higher PC1 (slightly more spread, but still not truly invasive based on your volume/density analysis)

---

5. Hierarchical Clustering Dendrogram (Bottom Center)

What it shows:

- Tree diagram showing how samples group based on similarity across all 16 metrics
- Height of branches indicates dissimilarity (Ward linkage method)

- Labels show condition and timepoint (e.g., "train_S1_t00")

Key observations:

- Four major branches corresponding roughly to the four K-means clusters
- Orange branch (left): train_S2 samples cluster together
- Green branch (center-left): train_S1 samples
- Green branch (center-right): test_S1 samples
- Purple branch (right): test_S2 samples

Biological interpretation:

- Samples from the same condition tend to cluster together across timepoints, indicating stable phenotypic signatures
- Some mixing within branches suggests temporal dynamics (certain timepoints share features across conditions)
- The hierarchical structure validates the K-means clustering—both methods independently identify similar groupings

---

6. Cluster Characteristics (Bottom Right)

What it shows:

- Bar plot comparing mean values of three key metrics across the 4 clusters:
    - Blue: Mean radius (÷100 for scale)
    - Orange: Dispersion index
    - Green: Number of leader cells

Key observations:

Cluster 0:

- Moderate mean radius (~530 μm)
- Low dispersion (~3)
- Moderate leader cells (~4.5)
- Interpretation: Compact, cohesive clusters with some spatial spread

Cluster 1:

- Highest mean radius (~900 μm)
- Low dispersion (~1.5)
- Highest leader cells (~7)
- Interpretation: Largest spatial extent—most "invasive-like" phenotype in your dataset

Cluster 2:

- High mean radius (~900 μm)
- Moderate dispersion (~2)
- Moderate leader cells (~5)
- Interpretation: Spread phenotype with moderate organization

Cluster 3:

- Moderate mean radius (~750 μm)
- Low dispersion (~1.5)
- Low leader cells (~4)
- Interpretation: Intermediate phenotype

Biological interpretation:

- Cluster 1 represents the most invasive-like state in your experiment (highest radius and leader cells)
- However, even Cluster 1 doesn't show robust invasion based on your earlier volume/density analysis—it's "relatively more invasive" compared to other clusters, but not truly invading
- This aligns with your conclusion that cells are primarily compact/non-invasive under these conditions

---

Summary: What This Analysis Tells You

1. Your data has clear phenotypic structure

The strong separation in PCA space and consistent clustering across methods (K-means, hierarchical) demonstrates that your invasion metrics capture real, reproducible biological variation—not just noise.

2. Condition-level differences dominate

PC1 (the main axis of variation) separates train vs test, and Sample 1 vs Sample 2. This suggests:

- Systematic batch effects or baseline differences between experimental groups
- Possible differences in initial spheroid size, matrix preparation, or cell state

3. Four distinct phenotypic modes exist

Unsupervised clustering identifies four natural groupings characterized by different combinations of radius, dispersion, and leader cell metrics. However, none show robust invasion dynamics.

4. Invasion-related features drive most variation

The feature loadings show that radius, leader cells, and spatial spread metrics (not just cell count or volume) are the primary discriminators between phenotypes.

5. Validation of your earlier conclusions

This multivariate analysis confirms your single-metric findings: the cells show heterogeneous spatial organization but minimal true invasion (no sustained volume expansion, density reduction, or leader cell progression over time).

Biological Context for Your Project

Your MDA-MB-231 cells in collagen matrix display:

- Phenotypic heterogeneity: Different samples/timepoints adopt distinct spatial configurations
- Limited invasion: Even the most "invasive-like" cluster (Cluster 1) doesn't exhibit hallmarks of active invasion
- Stable phenotypes: Clustering suggests cells maintain characteristic states rather than transitioning from compact to invasive over time

This reinforces that the collagen matrix conditions (density, stiffness, lack of chemoattractants) are likely suppressing invasion, and your analytical framework successfully captures this biological reality across multiple complementary perspectives.