APPENDIX – Results 1

PROMPT
<Can you interpret these plots? And provide some synthesis of the results? For context, the data analyzed are from measurements in mouse brain vasculature imaged using intravital 2-photon imaging and are derived from ROIs in 3D image stacks acquired using galvo-galvo scanner and 6X zoom with a Nikon 16x lens. There are two groups represented (control-red and experimental-black) but the data were analyzed blind to group (only represented in the PCA plot by text color of ROI#). Here are stattisal results when comparing group PC scores :

Results for PC1:
  Test: Mann-Whitney U test
  Statistic: 7534.0, p-value: 0.4224659152628756
Results for PC2:
  Test: Mann-Whitney U test
  Statistic: 6909.0, p-value: 0.062434681758270716
Results for PC3:
  Test: t-test
  Statistic: 0.7367184225588419, p-value: 0.4619655774546293>

RESPONSE
Based on the provided plots and statistical results, here's an interpretation and synthesis of the data:

1. PCA Analysis (Figure 1):
  - PC1 explains 52.94% of the variance in the data.
  - PC2 explains 30.20% of the variance.
  - Together, PC1 and PC2 account for 83.14% of the total variance.
  - The plot shows a spread of data points without clear clustering between the control (red) and experimental (black) groups.

2. PC1 vs Original Variables (Figure 2):
  - PC1 shows strong positive correlations with all three original variables:
    a. Zmicrons ($\rho = 0.822$, $p = 1.043e-64$)
    b. mean_FWHM_ums ($\rho = 0.470$, $p = 1.321e-15$)
    c. FeFv ($\rho = 0.832$, $p = 9.133e-68$)
  - FeFv has the strongest correlation with PC1, followed closely by Zmicrons.

3. Statistical Comparisons:
  - No significant differences were found between control and experimental groups for any of the principal components:
    - PC1: $p = 0.422$ (Mann-Whitney U test)
    - PC2: $p = 0.062$ (Mann-Whitney U test)
    - PC3: $p = 0.462$ (t-test)

Synthesis:

1. Variable Relationships: The PCA reveals that the three measured variables (Zmicrons, mean_FWHM_ums, and FeFv) are positively correlated. This suggests that these aspects of mouse brain vasculature tend to vary together.

2. Dominant Features: FeFv and Zmicrons are most strongly associated with PC1, indicating they contribute most to the primary axis of variation in the data. This could suggest that differences in vessel depth (Zmicrons) and a measure related to blood flow or vessel function (FeFv) are the most distinguishing features among the samples.

3. Lack of Group Separation: The PCA plot and statistical tests indicate no significant differences between control and experimental groups across the principal components. This suggests that the experimental condition did not produce large-scale, consistent changes in the measured vascular properties.

4. Potential Subtle Effects: The near-significant p-value for PC2 (0.062) hints at possible subtle differences between groups, which might become significant with a larger sample size or more sensitive measurements.

5. Data Structure: The PCA effectively reduced the dimensionality of the data, with the first two components explaining over 80% of the variance. This suggests that much of the variation in the data can be captured by these two composite measures.

6. Biological Interpretation: The strong correlations between variables might reflect underlying biological relationships in vascular structure and function. For example, vessel depth (Zmicrons) might be related to vessel size (FWHM) and blood flow characteristics (FeFv).

In conclusion, while the analysis reveals interesting relationships between the measured vascular properties, it does not provide evidence for significant differences between the control and experimental groups in this mouse brain vasculature study. The results suggest that any effects of the experimental condition on these vascular properties, if present, are subtle and not easily distinguished using this PCA approach.

****
Based on the information provided:

1. PC2 composition:
   PC2 is more likely to be influenced by mean_FWHM_ums (vessel width) rather than Zmicrons or FeFv. This is because:
   - Zmicrons and FeFv have very strong correlations with PC1 (0.822 and 0.832 respectively).
   - The correlation between mean_FWHM_ums and PC1 is notably lower (0.470).
   - PC2 often captures the next most important source of variation not explained by PC1.

2. Including additional data:

Yes, including X, Y coordinates and additional time series statistics could potentially increase the sensitivity and power of your analysis:

  - Spatial information (X, Y) could reveal location-dependent patterns in the vasculature.
  - Time series statistics (mean, stdev, min, max) for each metric at each Z-step could capture dynamic aspects of vessel behavior.
  - This additional information might reveal subtle differences between groups that aren't apparent in the current analysis.
  - However, be cautious of overfitting with too many variables. You may need to use feature selection techniques or regularization methods.

3. Using PC scores for clustering:
  Yes, you can definitely use the PC scores as input for clustering algorithms like K-means or UMAP:

  - This is a common approach in high-dimensional data analysis.
  - PCA reduces dimensionality while preserving most of the variance, making it easier for clustering algorithms to find patterns.
  - K-means could be useful if you expect distinct, spherical clusters.
  - UMAP might be better for capturing non-linear relationships and more complex cluster shapes.
  - Consider trying both and comparing results.

Additional suggestions:
1. Standardize your variables before PCA if you haven't already, especially if including new variables with different scales.
2. Consider using scree plots or explained variance ratios to determine how many PCs to retain for clustering.
3. If using K-means, try different numbers of clusters and use methods like the elbow method or silhouette score to determine the optimal number.
4. For UMAP, experiment with different parameter settings (n_neighbors, min_dist) to find the most informative representation.
5. Validate your clustering results using domain knowledge and additional statistical tests.

Remember, while these techniques can reveal patterns in your data, interpreting the biological significance of these patterns requires careful consideration of the underlying physiology and experimental context.

****

Including the grouping variable in the PCA analysis would not typically enhance the power or change the results in the way you might expect. Here's why:

1. Nature of PCA:

   PCA is an unsupervised method that doesn't consider group labels when determining the principal components. It focuses on maximizing variance across all data points, regardless of their group membership.

2. Potential Impact:
   - If you include the grouping variable as a binary feature (e.g., 0 for control, 1 for experimental), it would be treated as just another variable in the PCA.
   - This binary variable would likely have minimal impact on the overall variance structure, especially if the other variables have much larger ranges of values.
   - It might appear as a small contribution to one of the later principal components, but it's unlikely to significantly alter the main PCs.

3. Appropriate Use of Group Information:
   Instead of including the group variable in PCA, it's more common and usually more informative to:
   - Perform PCA without the group variable
   - Then analyze how the resulting PC scores differ between groups (as you've already done with your statistical tests)

4. Alternatives for Group-Aware Analysis:
   If you want to incorporate group information more directly, consider:
   - Linear Discriminant Analysis (LDA): This supervised method explicitly tries to find linear combinations of features that best separate the groups.
   - Partial Least Squares Discriminant Analysis (PLS-DA): Similar to PCA but takes group information into account.
   - ANOVA-simultaneous component analysis (ASCA): Combines ANOVA with PCA to analyze effects of experimental factors.

5. Enhancing Power:
   To potentially enhance the power of your analysis:
   - Ensure you have adequate sample sizes for each group
   - Consider using more sophisticated statistical methods that can handle high-dimensional data and group comparisons simultaneously (e.g., MANOVA, permutation tests)
   - Use dimension reduction techniques like PCA as a preprocessing step, then apply group comparison methods to the reduced data

6. Visualization:
   You can still use group information to enhance your visualization of PCA results, such as color-coding points by group in your scatter plots.

In summary, while including the grouping variable in PCA wouldn't directly enhance its power, there are other ways to leverage this information in your analysis. The approach you've taken - performing PCA and then comparing group differences in PC scores - is a valid and commonly

used method. If you want to more directly incorporate group information into your dimension reduction, consider the alternative methods mentioned above.