CytoAutoCluster:Enhancing Cytometry withDeep Learning

The notebook focuses on analyzing cytometric data using a semi-supervised autoencoder approach for dimensionality reduction and clustering. Here's a breakdown:

1. **Data Loading and Exploration:**
    - Loads the Levine_32dim dataset.
    - Performs initial exploration, including checking for missing values, duplicates, and data types.
2. **Data Cleaning and Preprocessing:**
    - Removes irrelevant columns and cleans column names.
    - Detects and handles outliers.
    - Calculates and visualizes the correlation matrix.
    - Examines the distribution, skewness, and kurtosis of features.
3. **Dimensionality Reduction with PCA and t-SNE:**
    - Applies PCA and t-SNE for initial dimensionality reduction and visualization.
4. **Data Corruption and Splitting:**
    - Corrupts the data using binary masking to create labeled (corrupted) and unlabeled (original) sets.
    - Splits the labeled data into training and testing sets for the autoencoder.
5. **Autoencoder Training (Implied):**
    - The notebook sets up the data for training a semi-supervised autoencoder (code not explicitly shown).
    - This autoencoder would learn from both corrupted and original data to extract meaningful features.
6. **Downstream Analysis (Implied):**
    - After autoencoder training, the encoded data would likely be used for:
        - Further dimensionality reduction and visualization.
        - Clustering to identify cell populations..

In essence, the notebook prepares cytometric data for a semi-supervised autoencoder approach, aiming to improve dimensionality reduction and clustering for subsequent biological analysis.