

DATA STRATEGY

DATA PIPELINE AND HIGH LEVEL ARCHITECTURE



DATASET PARAMETERS

Column Parameters:

- **Site:** The site of glycan measurement (**A, B, C, D, E**)
- **Glycan:** The specific glycan molecule (e.g., Glycan19, Glycan8, etc.)
- **Count1, Count2, Count3, Count4:** Likely represent the number of specific monosaccharide units in the glycan structure
- **Subject:** Subject identifier like, patient
- **Time (hours):** Time point of measurement
- **Area Percent:** Percentage of area occupied by the glycan
- **Protein Concentration:** Concentration of the protein
- **Glycan Concentration:** Concentration of the glycan



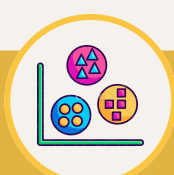
DATA PREPROCESSING AND FEATURE EXTRACTION

Numpy, Pandas:

The provided dataset is structured and **clean**. We used libraries to check for missing values, negative values, and duplicate rows

Numeric Descriptors:

- **Mean** glycan concentration
- **Peak** glycan concentration
- Area Under the Curve (**AUC**) using SciPy's trapz function (Trapezoidal rule)
- **Slope** of the concentration curve using NumPy's polyfit
- **Half-life** using SciPy's curve_fit for exponential decay



CLUSTERING

K-means Clustering:

- Unsupervised machine learning algorithm used for partitioning data into K distinct clusters
 - **Identifies groups of similar glycans** from each site (A,B,C,D,E) based on their **numerical features**
 - (**n_clusters=3**) using **Scikit-learn** (ML library): Specifies the number of clusters to form.
- Steps:**
- Randomly select K points as initial cluster centroids
 - Assign each data point to the nearest centroid based on Euclidean distance
 - Recalculate the centroids as the mean of all points in each cluster



DATA VISUALIZATION AND RESULTS

Matplotlib:

- Used "**matplotlib.pyplot**," glycan clearance curves were generated for various glycan molecules across different sites (A, B, C, D, E), resulting in five distinct plot curves
- Plot glycan clearance curves for each site based on the similarity in their numerical behaviour
- **PowerBI / R Studio:** For interactive dashboards or further insights
- The **y-axis scale** shows the *range of concentrations for different glycans*
- The **x-axis** allows for easy identification of *rapid vs. slow clearance*



OPTIMIZE MANUFACTURING AND REPORT FINDINGS

Manufacturing Optimization:

- Identify glycans that clear slowly (desirable) and those that clear quickly (undesirable)
- Develop recommendations to minimize the production of fast-clearing glycans during the manufacturing process
- **Recommendations for optimizing the drug manufacturing process**

Report Writing:

- Use **Microsoft Word** to document a comprehensive final project report documenting the entire study