GMM Dataset Preprocess

The initial steps of the data preprocessing involved the integration of secondary peaks information of NIR sequences from the **"230505_all_sequences.xlsx"** file into the **"230624_all_data_workup.xlsx"** file, specifically the **NIR Data** sheet.

The wavelength and LII were used to compute GMM components for the NIR peak, utilizing **"NIR Int Scaled"**. Since some sequences have secondary peaks, these secondaries were served as the second or third components of the NIR GMM.

The coefficients (a) and standard deviation (c) of the GMM components were determined using the equation **LII = a * c** with c set as sqrt(2)100.

Because we had previously calculated the **Cauchy-Schwarz distance** based on **Energy**, a decision was made to change wavelength to energy again. The formula **E = hc/λ** was used.

E = h * c / (λ 1e-9) # Using nanometers for wavelength

h = 6.626e-34 # Planck's constant in J.s

c = 3.0e8 # Speed of light in m/s

λ # Wavelength in nm

Note: the logarithm (log10) of all coefficients (a) was taken.

The NIR GMM results from **"NIR-GMM.xlsx"** were merged with the **"Gaussian fit"** sheet for visible sequences into one Excel file named **"All-sequences-GMM.xlsx"** for all four classes. For visible sequences, any sequences that showed negative values for either coefficients or standard deviation were excluded. These were saved in a file named **"visible-GMM.xlsx".**

The next phase involved the calculation of the **pairwise Cauchy-Schwarz distance**, followed by the **clustering** of the sequences.

For spectral clustering, a **"precomputed"** setup for the affinity matrix was used. However, as the distance was available, it had to be transformed into a **similarity matrix**. This was achieved by applying a **Gaussian kernel** to the distance matrix using the formula:

exp(-dist_matrix ** 2 / (2. * delta ** 2))

Here, delta served as a free parameter representing the width of the Gaussian kernel. (Delta= 0.01)