**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.

Then, I used the wavelength and LII to compute GMM components for the NIR peak, utilizing **"NIR Int Scaled"**. Since some sequences have secondary peaks, these secondaries serve 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**, we decided to change wavelength to energy again. We used the formula **E = hc/λ.**

**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:** we took the logarithm (log10) of the coefficients (a).

The NIR GMM results from **"NIR-GMM.xlsx"** was 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, I excluded any sequences that showed negative values for either coefficients or standard deviation. 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 applying spectral clustering, we used a **'precomputed'** setup for the affinity matrix. However, as we have the distance, we must transform it 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** is a free parameter representing the width of the Gaussian kernel. (Delta= 0.01)