Supplementary Material

Specificity of carbon nanotube accumulation and distribution in cancer cells revealed by K-Means clustering and principal component analysis of Raman spectra

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Lena Golubewa,*[a] Igor Timoshchenko,[b,c] and Tatsiana Kulahava [c]

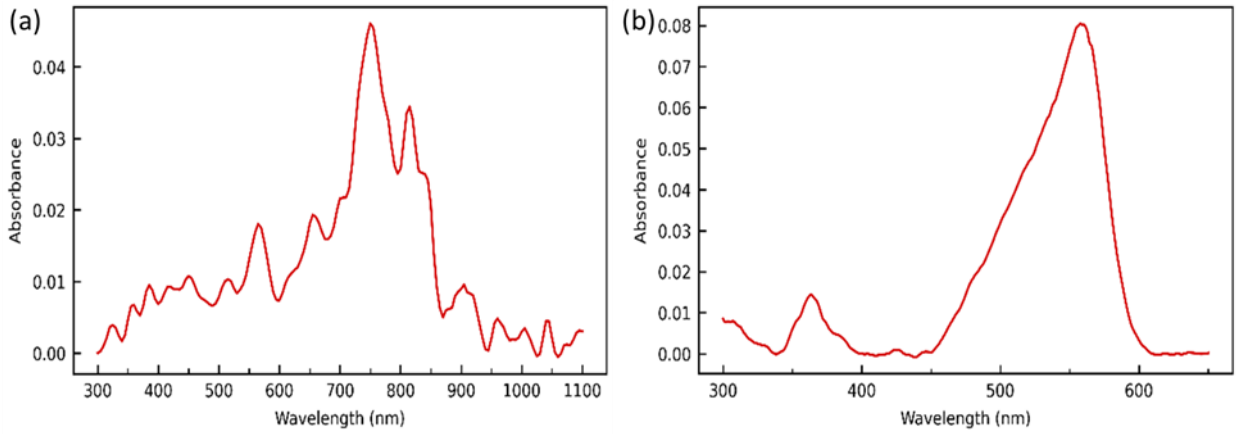


Figure S1. UV-vis spectrum of (a) SWCNT-DNA and (b) SWCNT-ON suspensions.

[a] *Department of Molecular Compounds Physics, State research institute Centre for Physical Sciences and Technology, Saulėtekio av. 3, Vilnius, 10257, Lithuania.*
[b] *Department of Computer Modelling, Physics Faculty, Belarusian State University, Nezavisimosti av. 4, Minsk, 220030, Belarus.*
[c] *Laboratory of Nanoelectromagnetics, Institute for Nuclear Problems of Belarusian State University, Bobruiskaya str. 11, Minsk, 220006, Belarus.*

Supplementary Material

Determination of the optimal number of PCs and clusters.

1. Dimensionality reduction

Normally, the number of PCs sufficient for the analysis is selected according to the Kaiser-Jolliffe rule, which states that the selected PCs should cover 70-90% of the total data variance [S1] [S2]. In practise, however, the selection is determined by the type and quality of the data. In our case, the cell spectra are mainly noisy and have an unknown noise level. The original data set consists of 1024 variables and the PCA generates 1024 PCs, most of which cover the variance caused by the noise. The lower the useful Raman signal, the higher the variance caused by the noise (see Figure 3a and 3b, main text). However, the main variance arises from the cells, the SWCNTs and their interaction with the cellular compartments, and this information is covered by the first 3 PCs. The loadings of the first three PCs are informative in all data sets, regardless of the degree of noise. This can be seen, for example, in Figure S6-S8 (right column) in the Supplementary Material. The loadings of PC4 and higher are informative in the case of a higher signal-to-noise ratio (Figure S2a), while in the case of a low signal-to-noise ratio no sufficient information can be extracted except about the noise (Figure S2b). Therefore, the first three PCs are used for data interpretation, while the PCs with higher numbers are used for noise reduction.

2. Denoising of Raman spectra

As shown in [S3] for Raman spectra of biological objects (phantom samples, human fingernail, leukaemia cells), the number of PCs for the denoising of Raman spectra with unknown noise level can be freely selected between 5 and 9 PCs. The more PCs are used, the more information is recovered in the reconstructed spectra. In our study, we used 10 PCs for noise reduction. As can be seen in Figure S2, 10 PCs provide a good approximation to the G peak of the SWCNTs for both high (Figure S2 a, b) and low intensities of the Raman signal (Figure S2 d, e). Twenty PCs also describe the G-peak well. However, as can be seen in Figures S2c and S2f, which show the "silent" spectral region, more noise is recovered.

3. The quality and performance of the clustering algorithm

The quality and performance of the clustering algorithms were evaluated using the following clustering metrics: Silhouette coefficient, Davies-Bouldin index and Calinski-Harabasz index S4. The metrics reflecting the clustering results are shown in Figure S3 and correspond to the case of K-Means performed with (i) preprocessed but not denoised, (ii) denoised Raman spectra and (iii) 10 PCs. As can be seen from the data shown in Figure S3, denoising the spectra or using PCs for clustering significantly improves the metrics. The K-Means cluster maps in the case of denoised spectra and PCs differ only slightly from each other but allow a better visualisation of the cellular compartments than in the case of non-denoised spectra (Figure S4). The number of clusters was chosen according to the most optimal combination of metrics (the highest Silhouette coefficient, the lowest Davies-Bouldin index and the highest Calinski-Harabasz index). The first such combination of metrics corresponds to $k = 2$ and represents a trivial data segmentation into cell and buffer (Figure S4). The next optimal combination corresponds to $k = 5$ clusters and reflects the main components of the analysed system: SWCNTs, nucleus, cytoplasm, membrane, buffer. The further optimal number of clusters is $k = 7$, but this corresponds to overclustering, as it generates two clusters for buffer, for example.
The clustering results may change slightly for different states of the random number generator, but such changes are virtually eliminated by the considerably large number (10) of runs of the K-Means algorithm with different centroid seeds.

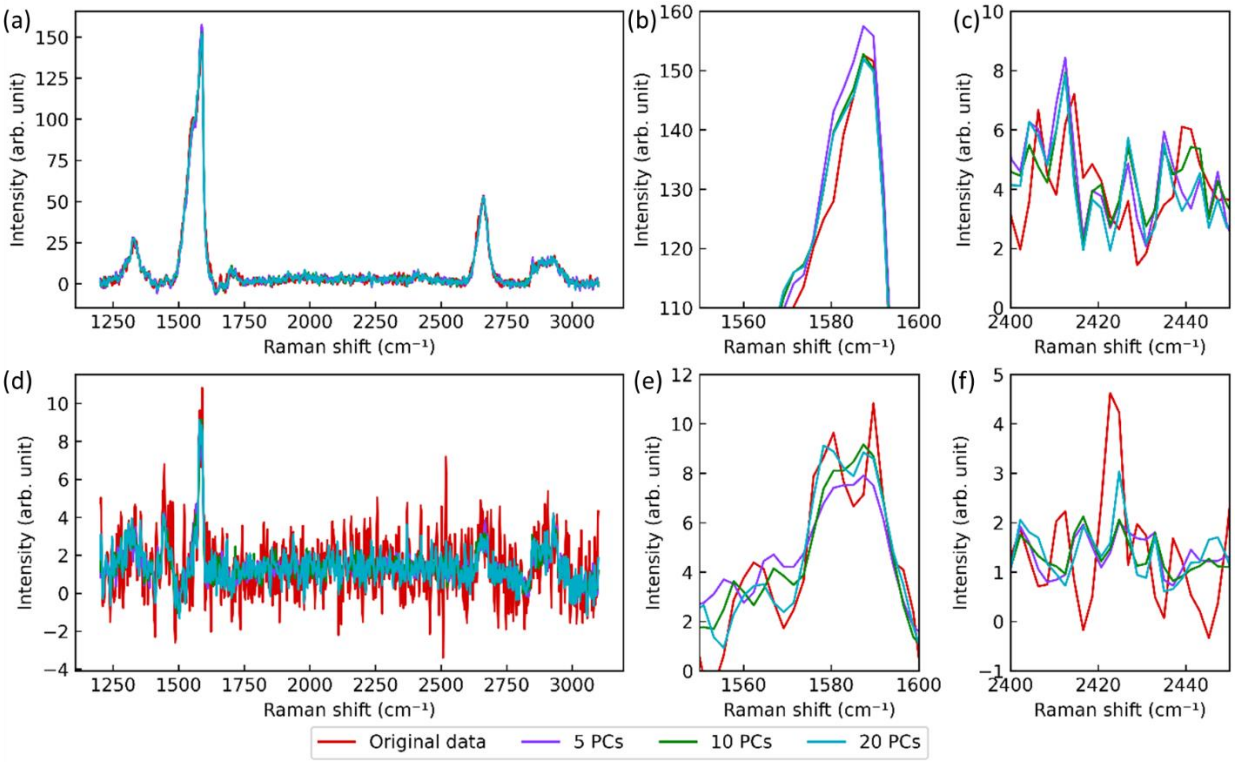


Figure S2. Recovery of spectra with different numbers of PCs: (a) and (d) full range, (b) and (e) G-band, (c) and (f) "silent" range, selected in high and low intensity Raman spectra, respectively.

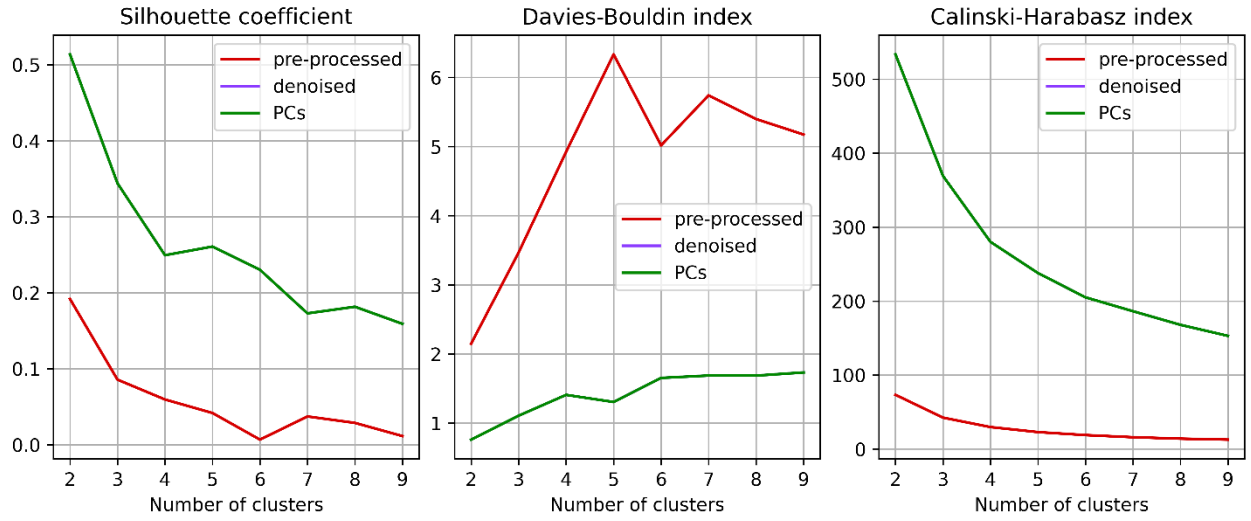


Figure S3. Clustering metrics calculated in the case of preprocessed but not denoised and denoised Raman spectra. The metrics for K-Means for PCs (green line) match the metrics of K-Means for denoised spectra (purple) and are indistinguishable in the Figure.

Supplementary Material

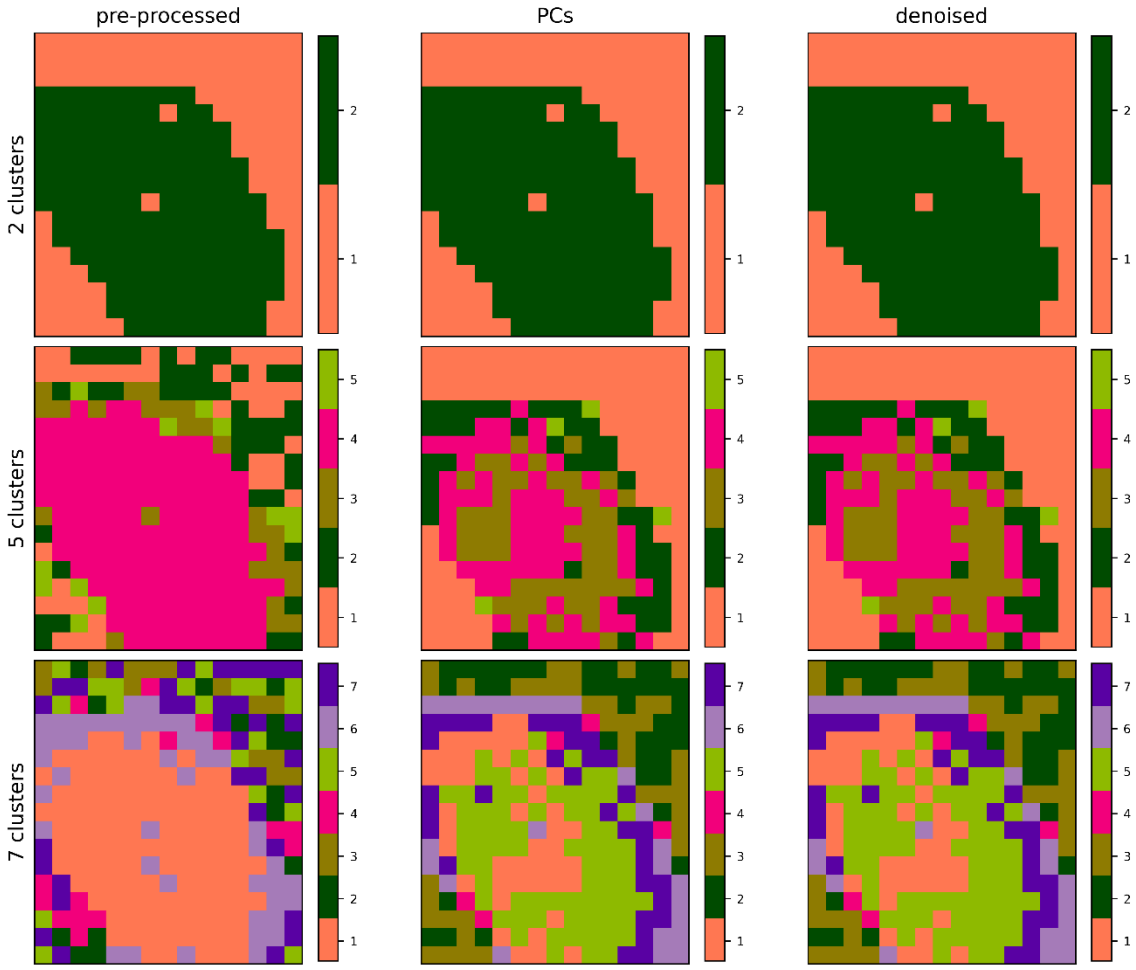


Figure S4. K-Means clustering calculated in the case of preprocessed but not denoised, denoised Raman spectra and PCs for two, five and seven clusters.

Supplementary Material

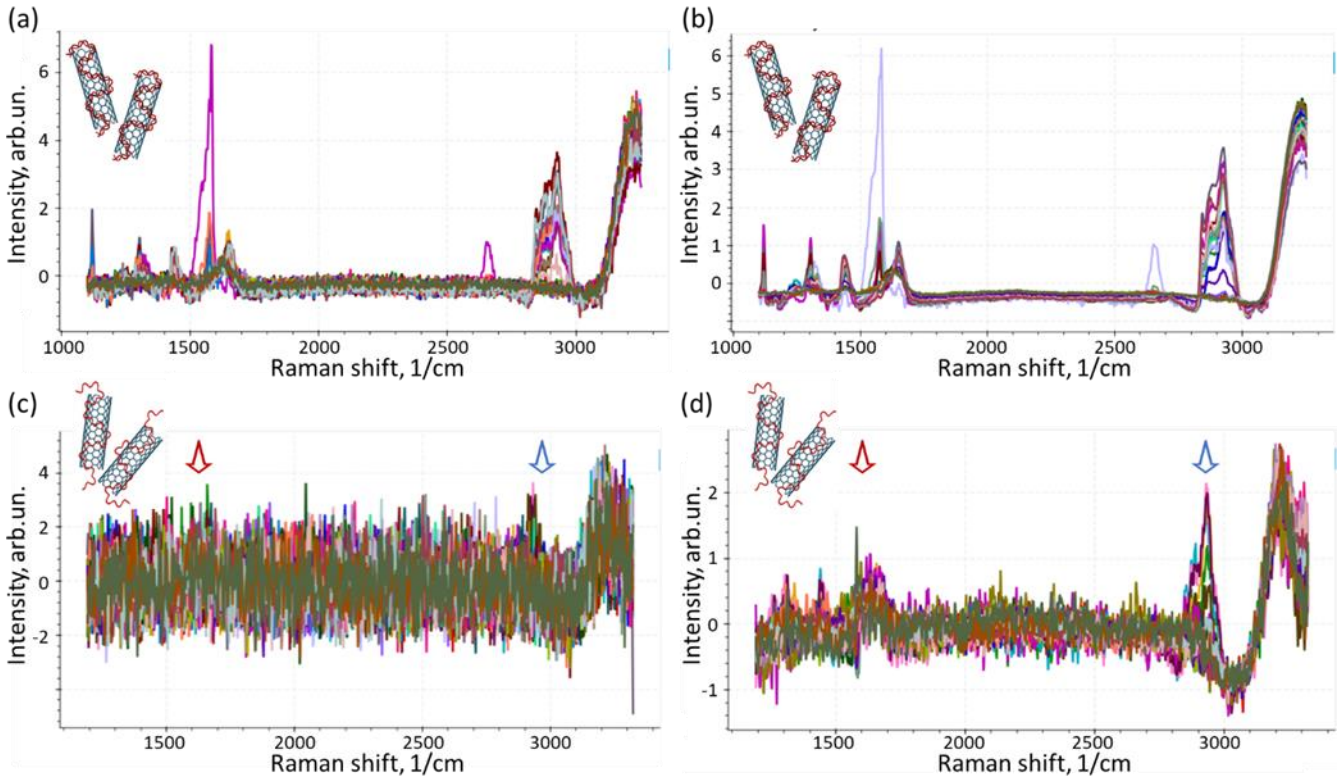


Figure S5. Several initial Raman spectra (a, c) and reconstructed by 10 PCs (b, d) from the scans of C6 glioma cells exposed to SWCNT-DNA complexes for 24 h (a, b) and to SWCNT-ON complexes for 72 h (c, d).

Supplementary Material

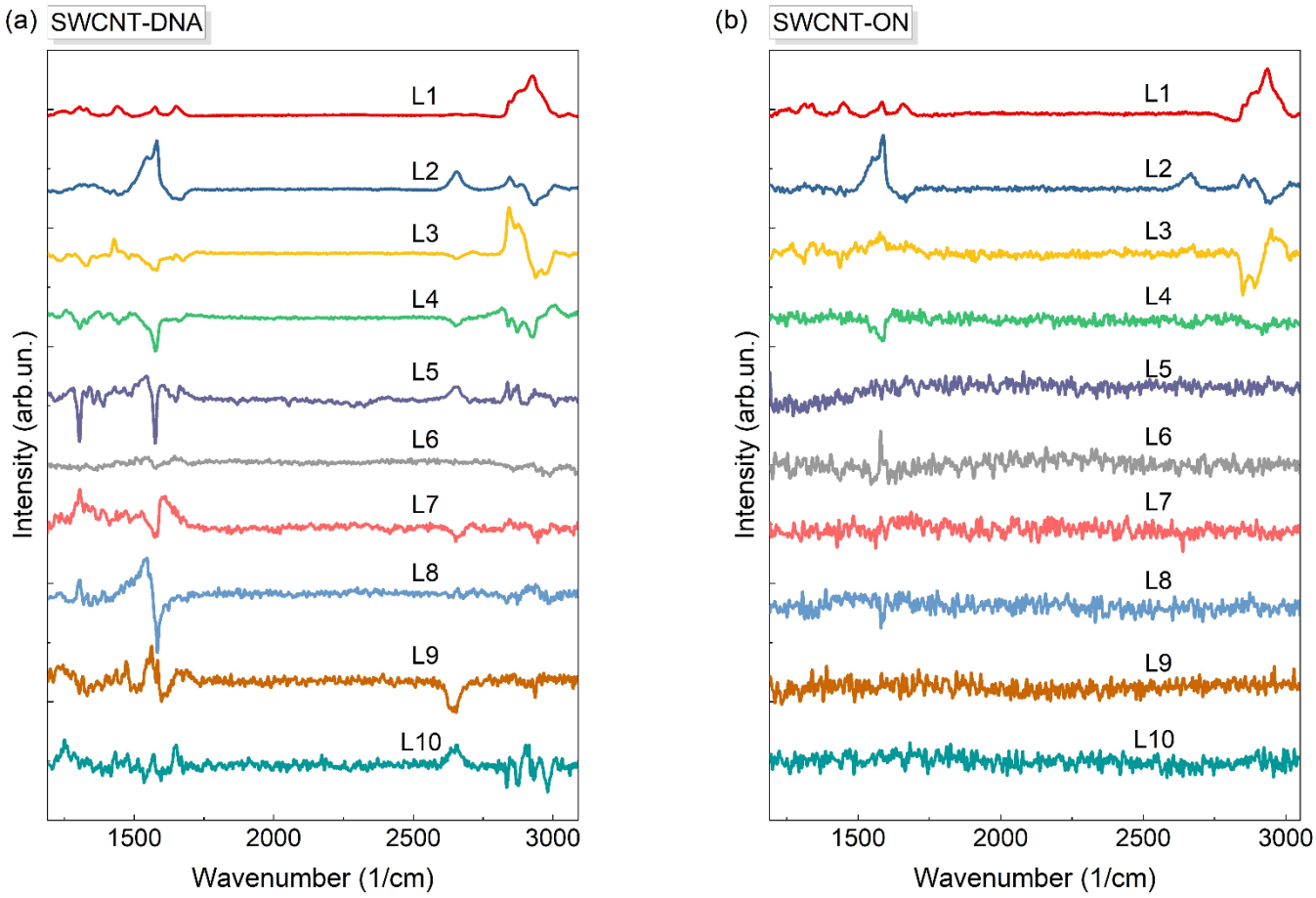


Figure S6. The first ten loadings for spectral datasets of cells exposed to SWCNT-DNA (a) for 24 hours and SWCNT-ON (b) for 72 hours.

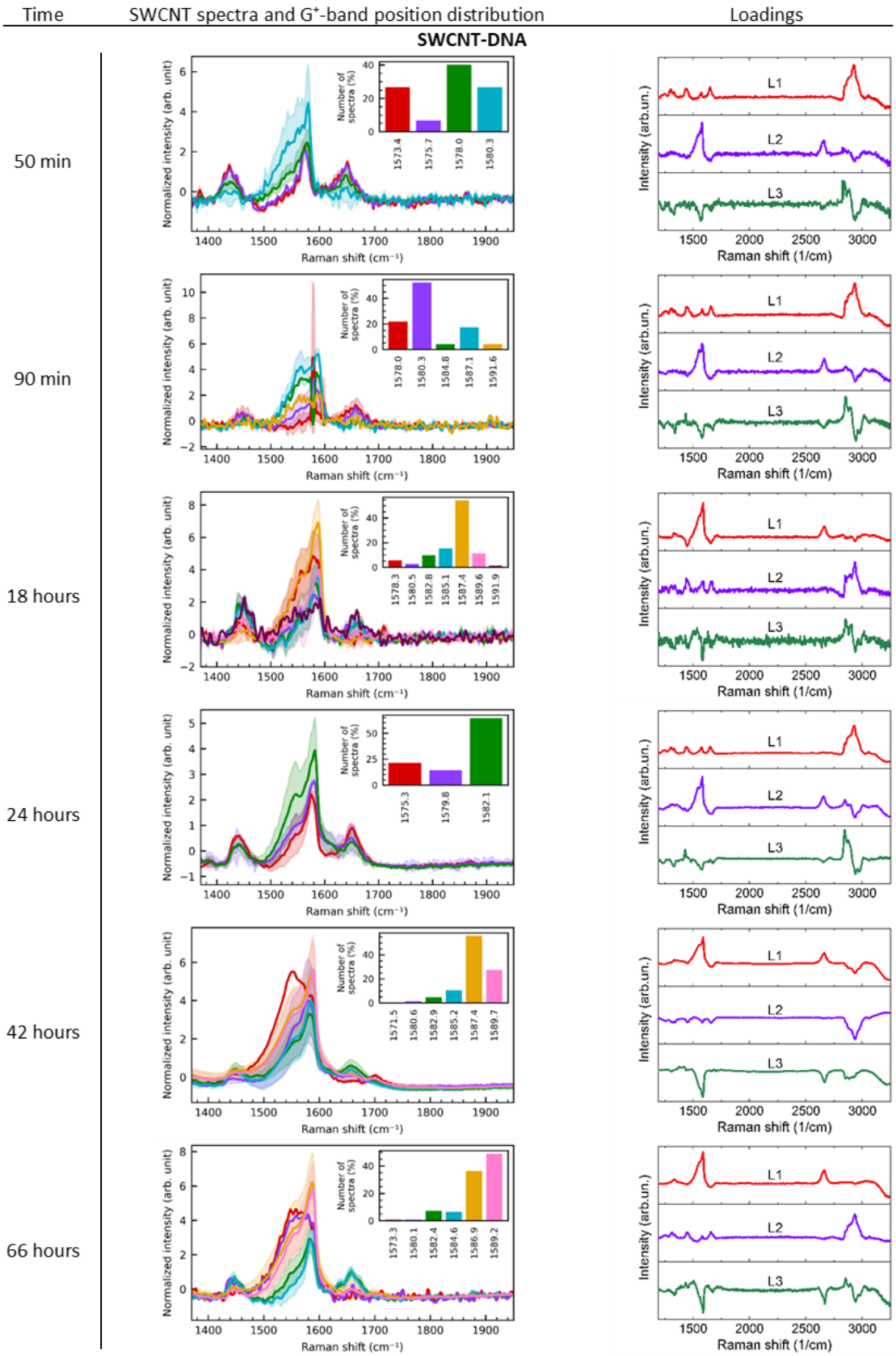


Figure S7. Spectra and Loadings calculated for cells shown in Figures 6-7 in the main text.

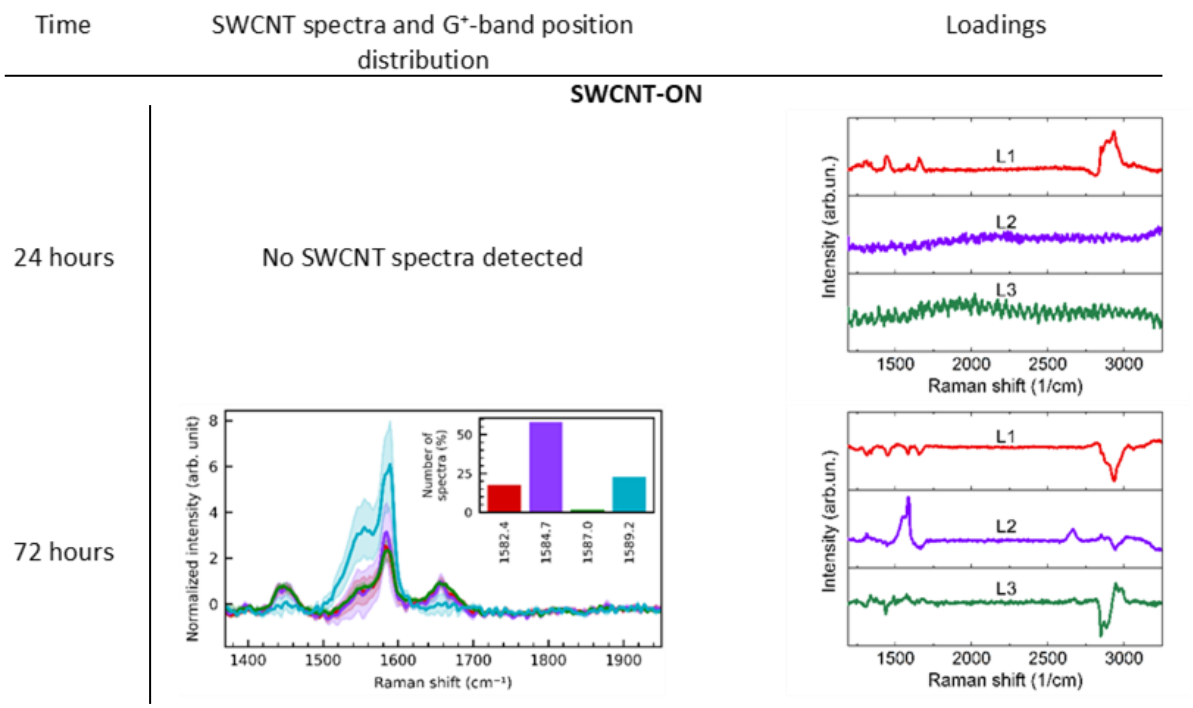
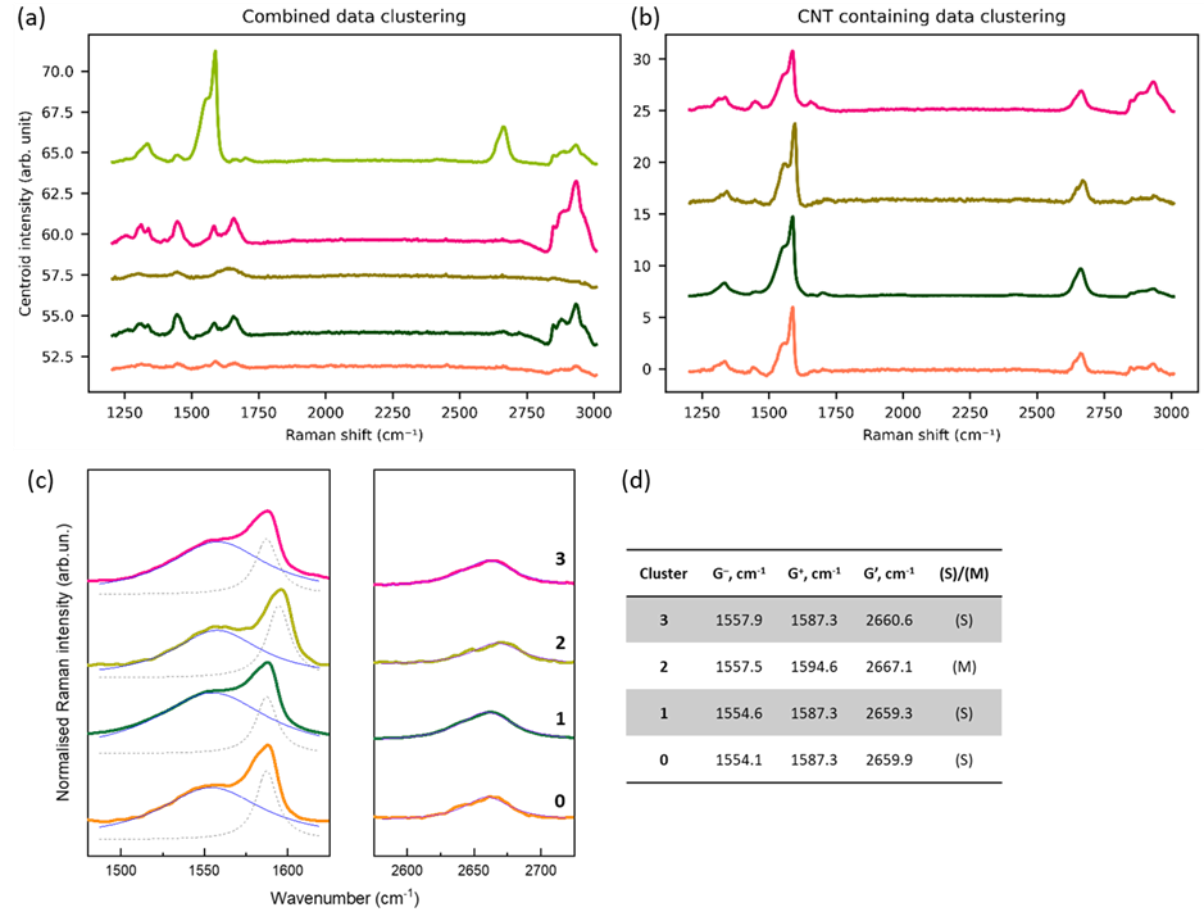


Figure S8. Spectra and Loadings calculated for cells shown in Figure 8 in the main text.

Supplementary Material



Cluster	G⁻, cm⁻¹	G⁺, cm⁻¹	G', cm⁻¹	(S)/(M)
3	1557.9	1587.3	2660.6	(S)
2	1557.5	1594.6	2667.1	(M)
1	1554.6	1587.3	2659.3	(S)
0	1554.1	1587.3	2659.9	(S)

Figure S9. Analysis of the heterogeneity of chirality of SWCNTs accumulated in C6 glioma cells: (a) the centroids of clusters extracted from the Raman spectral data set of cells incubated from 50 min to 24 h with SWCNT DNA and to 72 h with SWCNT-ON; (b) the centroids of SWCNT-related clusters; (c) approximation of the G^+, G^- and G'-bands; (d) the position of the maxima of the G^+, G^- and G'-bands.

Supplementary Material

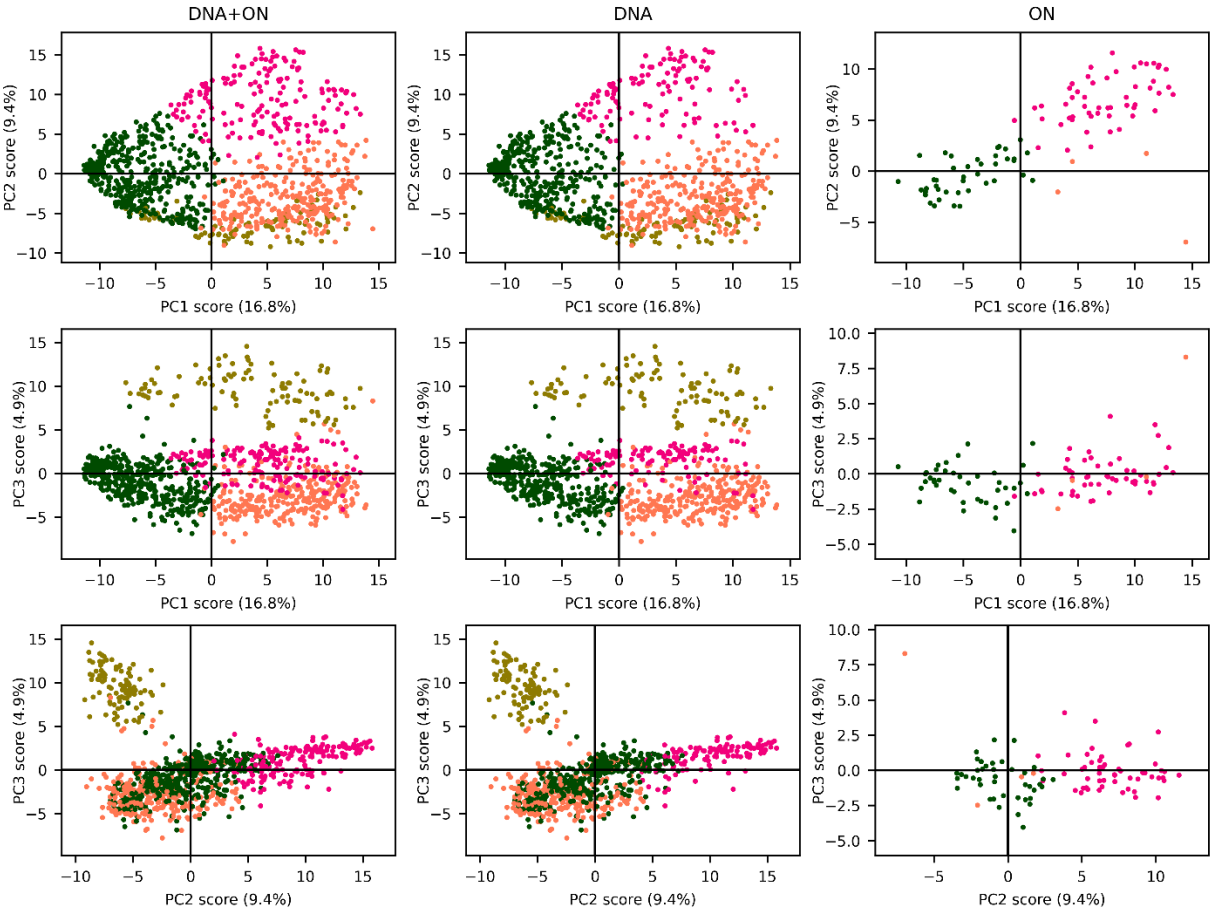


Figure S10. The score plots of the data relating to SWCNTs only, related to the data described by the light green centroid in Figure S9a: the left column represents the entire data set of spectra from cells exposed to SWCNT-DNA and SWCNT-ON, the centre column represents the data set of spectra from cells exposed to SWCNT-DNA, the right column represents the data set of spectra from cells exposed to SWCNT-ON. The colours of the data correspond to the colours of the centroids in Figure S9.

Supplementary Material

References:

S1 Daniel S. Wilks., *Statistical methods in the atmospheric sciences, second edition*, Academic press, Elsevier, Second., 2007, vol. 14.

S2 I. T. Jolliffe, *Encycl. Stat. Behav. Sci.*, 2002, **30**, 487.

S3 Y. Bai and Q. Liu, *Biomed. Opt. Express*, 2020, **11**, 200.

S4 L. E. Ekemeyong Awong and T. Zielinska, *Sensors*, 2023, **23**, 7925.