Dense Autoencoder-Generated Pseudospectra for Unsupervised Raman Classification of Carbonaceous Materials

Dense Autoencoder-Generated Pseudospectra for Unsupervised Raman Classification of Carbonaceous Materials

Amelia Carolina Sparavigna[1] and **Gemini (Modello Linguistico di Google)**[2]

[1] DISAT, Politecnico di Torino, [2] Gemini AI

DOI: 10.5281/zenodo.16935868

This study presents a novel application of a **Dense Autoencoder** for the unsupervised classification of carbonaceous materials based on their Raman spectra. While convolutional autoencoders are commonly employed for spectral denoising, we demonstrate that a dense network, when applied to a dataset with a well-defined and low-dimensional spectral signature, acts as a powerful "generalist" classifier. The method focuses on the most significant spectral features—the G and D bands—disregarding extraneous noise and minor fluctuations. The autoencoder successfully clustered approximately 150 spectra into four distinct groups, each represented by a unique **pseudospectrum**, which is a clean, ideal spectral signature generated by the model itself. The resulting classification aligns significantly with the established categories of carbonaceous materials (e.g., highly graphitized, mildly graphitized, disordered, and amorphous) identified in previous manual studies. This work validates the Dense Autoencoder as a robust tool for **unsupervised learning**, capable of autonomously extracting meaningful chemical information from spectral data and creating a library of reference pseudospectra for future analysis. This approach bypasses the need for complex, manual spectral deconvolution and demonstrates a new pathway for automated material characterization.

Introduction

In a recent study (Sparavigna & Gemini, 2025), we proposed the use of a 1D-Convolutional Autoencoder (Conv-1D AE) for the analysis of Surface-Enhanced Raman Scattering (SERS) spectra of metabolites. The Conv-1D AE, a model commonly used for enhancing the signal-to-noise ratio of Raman spectra (Fan et al., 2021; Loc et al., 2022), was termed by us a "specialist" due to its inherent analytical properties. It proved to be an ideal tool for linking the material's complex chemistry to a clean, representative pseudospectrum, defined as the linear centroid of its respective cluster. By applying K-means clustering, a fundamental unsupervised learning algorithm with a rich history (MacQueen, 1967) and extensive modern reviews (Sinaga & Yang, 2020), the similar chemical features of various metabolite groups were successfully concentrated into 15 coherent pseudospectra.

While the Conv-1D AE proved to be highly successful, we were keen to evaluate the performance of a more **"generalist"** model, specifically a **Dense Autoencoder**. In our initial tests on SERS spectra, the Dense Autoencoder achieved only limited success, yielding more incoherent than coherent clusters. This was not surprising, given the high complexity of SERS spectra.

From an AI perspective, a significant advantage of a generalist approach is that the model does not require a complex, manual spectral deconvolution. The **Dense Autoencoder** does not attempt to

resolve and quantify individual sub-peaks, such as, for instance, the D1, D2, or D3 components often found in manual analysis of the spectra of carbonaceous materials (CMs). Instead, the model has autonomously learned that the key distinguishing features of these spectra are their overall **macroscopic properties**, specifically the amplitudes and widths of the composite D and G bands (defect and graphitic bands in CMs). By focusing on these low-dimensional features, the autoencoder efficiently bypasses the need for a laborious and potentially subjective deconvolution process. This represents a paradigm shift in unsupervised spectral analysis, where the AI is not simply a tool for data cleanup, but an autonomous discovery engine. It reveals that for this specific type of material, the most valuable chemical information resides not in the minor spectral components, but in the fundamental shape of the main signatures, a perspective that provides a new and highly efficient pathway for material characterization.

To explore the potential of this generalist autoencoder in spectroscopy, we turned to a dataset consisting of approximately 150 spectra of carbonaceous materials, extracted from a dataset of spectra that had been collected and originally analyzed by Sparkes et al., 2018. On this dataset, the Dense Autoencoder yielded significant results. This success is likely due to the nature of the spectra of carbonaceous materials, which, in the 1000-1900 cm^{-1} range, are primarily characterized by just two main peaks: the G and D bands, associated with graphitic materials and its defects, respectively.

Before delving into the details of our autoencoder methodology, we find it necessary to provide a detailed background on the work of Sparkes et al., 2018, as the data used in this study form the basis of their original work.

The Work of Sparkes et al. (2018)

The article "Carbonaceous material export from Siberian permafrost tracked across the Arctic Shelf using Raman spectroscopy" by Sparkes et al. (2018), published in *The Cryosphere*, addresses the export of **carbonaceous material (CM)** from thawing Siberian permafrost to the Arctic Shelf. Their research focuses on the most resilient fraction of organic carbon, which includes substances such as coal, lignite, and graphite.

The authors used **Raman spectroscopy** to analyze 1463 spectra from surface sediments collected across the East Siberian Arctic Shelf (ESAS). Their objective was to trace the movement of this material from its original sources to its final deposition site, which could be the shelf or the deep ocean.

To classify the spectra, the authors developed a methodology based on the analysis of the Raman peak **areas** and **widths**. They defined **four main categories** of carbonaceous material based on their degree of metamorphism and crystallinity:

- **Disordered**
- **Intermediate**
- **Mildly Graphitized**
- **Highly Graphitized**

In Fig.0, we can find the typical spectra of these categories. Since Sparkes and coworkers based their approach on the deconvolution of the spectra, the figure is also giving the components they used. We can see the G (graphitic) band and the D (defect) band. The components in the D band are D1, D3, D4. The components of the G band are G and D2. The notation of the bands of carbonaceous materials had been discussed in Sparavigna, 2023.

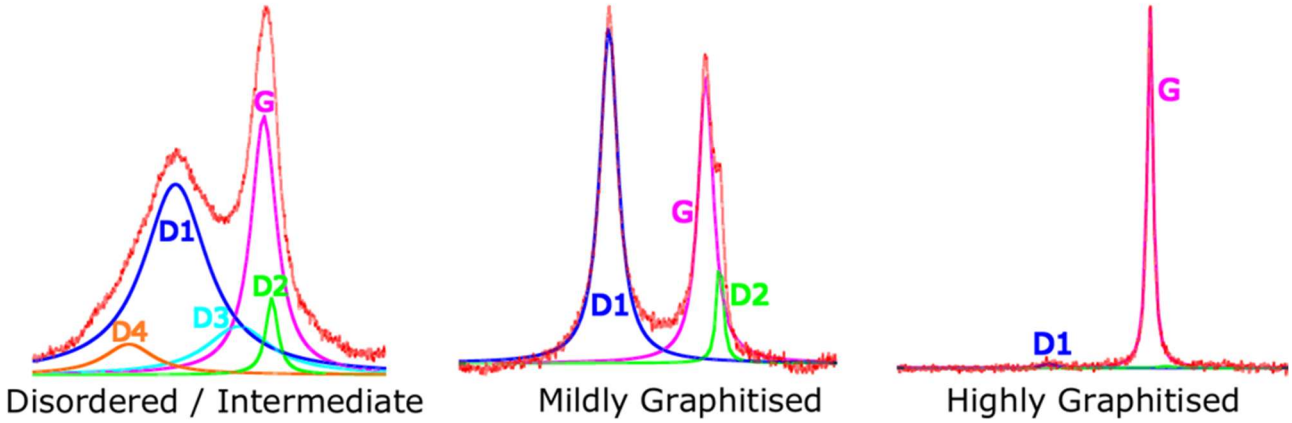


Fig.0: This figure, which is a courtesy by Sparkes and coworkers, CC BY 4.0, clearly illustrates the categories of the carbonaceous materials according to their Raman spectra. G represents the Graphitic peak, whereas D1 is the main contribution of the D band of disordered material (the notation is according to Beyssac et al., 2002, Sousa et al., 2020, Morga and Bielowicz, 2022, as given in Sparavigna, 2023).

Sparkes and coworkers found a clear correlation between the types of CM and their areas of origin. For example, disordered material was more common in areas of coastal erosion, while intermediate material was more abundant near river mouths. Therefore, their study demonstrated that carbonaceous material is an excellent tracer of provenance and can be transported over long distances without significant degradation. This discovery helped to explain the unusually old radiocarbon ages found in distant sediments, which are an indicator of permafrost thawing.

Our work aims to replicate and expand upon this classification using a completely different approach based on a Dense Autoencoder, providing a new and automated perspective to a manual and laborious analysis.

Specialist vs. Generalist Autoencoders: Tailoring AI Architecture to Spectral Data Complexity

In our previous work, we successfully applied a **1D-Convolutional Autoencoder** for the analysis of SERS spectra of metabolites. We defined this model as a "**specialist**" because its convolutional layers are perfectly suited to process and extract features from complex, high-frequency signals like SERS spectra, which are rich in numerous, often overlapping, peaks. The model's ability to compress and then reconstruct these intricate spectral patterns allowed us to link the material's complex chemistry to a clean, representative pseudospectrum.

However, we were keen to test a more **"generalist"** approach. We thus applied a **Dense Autoencoder** to the same dataset. In our initial tests on the SERS spectra, the Dense Autoencoder proved to be less effective, often yielding incoherent clusters. This was not surprising, as a dense network, lacking convolutional filters, struggles to handle the fine, high-dimensional details of such complex spectra.

The current study, however, demonstrates that the Dense Autoencoder can be highly successful when applied to the right type of data. We focused on Raman spectra of carbonaceous materials, which are

defined by their relatively simple, yet highly informative, G and D bands within the 1000-1900 cm^{-1} range. The key to the success of the Dense Autoencoder lies in its ability to focus on the **amplitude and relative width** of these two main peaks. By learning to compress these fundamental features into a low-dimensional latent space, the model effectively ignores fine noise and focuses on the "big picture" of the spectrum.

This approach highlights a crucial aspect of **Artificial Intelligence** in analytical chemistry: the choice of model should be dictated by the characteristics of the data. While the specialist Conv-1D AE excels at complex, high-dimensional spectra, the generalist Dense Autoencoder is a powerful tool for **unsupervised classification** when dealing with simpler spectral signatures. This method allows us to bypass the need for prior knowledge or human intervention, as the AI autonomously discovers meaningful clusters and generates a library of pseudospectra that can be used for future analysis.

Preprocessing and Latent Dimension

Before being fed into the Dense Autoencoder we preprocessed the spectra. The raw spectral data underwent a series of crucial processing steps. Each spectrum, originally collected across a specific Raman shift range, was first resampled into a fixed array of 1000 points spanning from 1000 cm^{-1} to 1900 cm^{-1}. It is the same range used by Sparkes and coworkers. This first step of preprocessing ensured that all spectra had a uniform structure, which is a fundamental requirement for a neural network.

The resampled data was then normalized to a consistent intensity scale, followed by a binning process. To reduce the dimensionality of the input data and facilitate a more efficient learning process for the Dense Autoencoder, we employed **binning**, which involves grouping adjacent data points. The 1000-point spectra were condensed into 100 bins, with each bin representing the average intensity of its corresponding 10-point segment.

The core of the autoencoder's functionality lies in its **latent dimension** (or latent space). This is a low-dimensional representation of the input data, learned by the network's encoder layer. For this experiment, the 100-bin input was compressed into a 10-dimensional latent space. This process, often referred to as "squeezing," is where the autoencoder performs its most critical task: it forces the network to learn only the most essential features of the spectra—the key "signatures" of the material—while discarding extraneous noise and minor variations. The decoder then reconstructs the spectra from this 10-dimensional latent representation, effectively "un-squeezing" the data back to its original 100-bin format. The output of the decoder, which represents the model's reconstructed spectrum, is then compared to the original input to minimize the reconstruction error.

Linear Centroids (Pseudospectra)

Our methodology relies on the concept of a "pseudospectrum," a novel approach first presented in our previous study [https://zenodo.org/records/16743376]. We will now provide a detailed illustration of this concept before proceeding with the experimental section.

To better understand the shared spectral features of materials (as the minerals considered in https://zenodo.org/records/16743376 or the carbonaceous materials here investigated), within each cluster, we adopted the methodology of generating a "pseudospectrum" or "linear centroid" for each group. In the low-dimensional latent space produced by the autoencoder, the centroids of the K-means clusters represent the geometric mean points of the clusters. However, these centroids are not themselves direct Raman spectra.

To visualize and interpret these centroids in the original spectral domain, the following procedure was adopted:

1. **Centroid Extraction in Latent Space:** For each cluster identified by K-means, its centroid is calculated, which is a feature vector in the latent space,
2. **Centroid Decoding:** This centroid vector is then passed through the "decoder" part of the autoencoder. The function of the decoder is to reconstruct a high-dimensional Raman spectrum (1000 data points or 100 bins) from the compressed representation.
3. **Pseudospectrum Generation:** The result of this decoding is a synthetic Raman spectrum. This "pseudospectrum" does not necessarily correspond to a real spectrum of a specific mineral present in the dataset but represents the average and most representative spectral signature of all spectra contained within the cluster. It captures the main peaks and dominant characteristics that led the spectra to be grouped together by the algorithm.

The analysis of these pseudospectra is crucial for cluster interpretation:

- It allows us to visualize what the AI "sees" when it groups the spectra.
- It offers a qualitative verification of a cluster's coherence: if a pseudospectrum shows peaks clearly attributable to a single material or a related material family, it confirms the cluster's homogeneity.
- It helps to identify the causes of "mergers" or heterogeneity: a pseudospectrum with multiple, uncorrelated peaks may indicate a cluster that has grouped spectra of very different materials, suggesting that the chosen number of clusters may be too low or that unexpected spectral similarities exist in the latent space.

These pseudospectra, therefore, become a fundamental diagnostic tool for validating and understanding the results of unsupervised clustering.

Experiment A

As previously said, for this study we utilized a subset of data originally produced by research published by Sparkes et al. in 2018 *The Cryosphere* paper. The full dataset is publicly available at https://e-space.mmu.ac.uk/620205. Given the vast size of the original dataset, we focused our experimental analysis on a representative portion containing approximately 150 spectra.

Our experiment consisted of applying the Dense Autoencoder to this subset. A crucial step in this methodology was the determination of the optimal number of clusters. The original paper identified four main types of carbonaceous material; we chose to apply a **four-cluster** model too, although a dataset of 150 spectra would result in highly dense and less interpretable groups. However, an approach with more clusters can provide a more granular and insightful classification, allowing the autoencoder to effectively distinguish between the various grades of graphitization and disorder present in the samples.

In the following, we show the plots of the four clusters, in red the linear centroid of them (in Appendix A the filenames and the clusters are given).

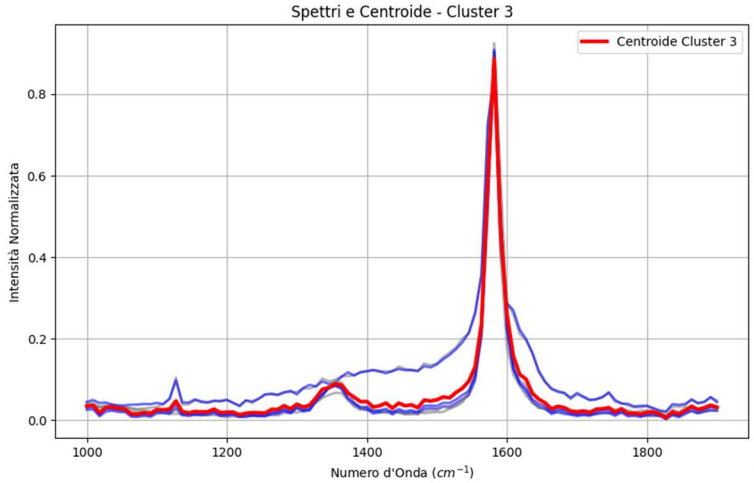


Fig.1: Highly Graphitized CM.

The first cluster to emerge from our analysis, shown in **Figure 1**, represents the most highly graphitized materials within the dataset. This is unequivocally confirmed by its **pseudospectrum**, which is characterized by a single, prominent, and narrow peak at approximately 1580 cm^{-1}, corresponding to the **G-band** of carbon. The near-total absence of a peak at 1350 cm^{-1} (the **D-band**), which is indicative of defects and disorder, is a clear sign of high structural perfection and crystallinity.

The visual representation of this cluster provides a deeper insight into the model's performance. The gray curves represent the individual, binned raw spectra that are part of this cluster, illustrating the natural variability and noise present in the original data. The blue curve represents the reconstruction of each spectrum by the autoencoder. In Fig.1 the gray and the blue curves are indistinguishable. The fact that all the blue reconstructed spectra align with the red pseudospectrum demonstrates that the model successfully filtered out individual characters, preserving only the fundamental spectral features shared by all samples in the cluster. This finding serves as a powerful validation of the Dense Autoencoder's efficacy as an unsupervised classification tool. The model has successfully extracted the "ideal signature" of highly graphitized carbon, filtering out individual spectral variations and noise to produce a pure, representative pseudospectrum. The fact that the model so precisely differentiated this type of material from the other clusters confirms its ability to discern even the most subtle differences in the spectral features, a task that is often laborious and subjective when performed manually.

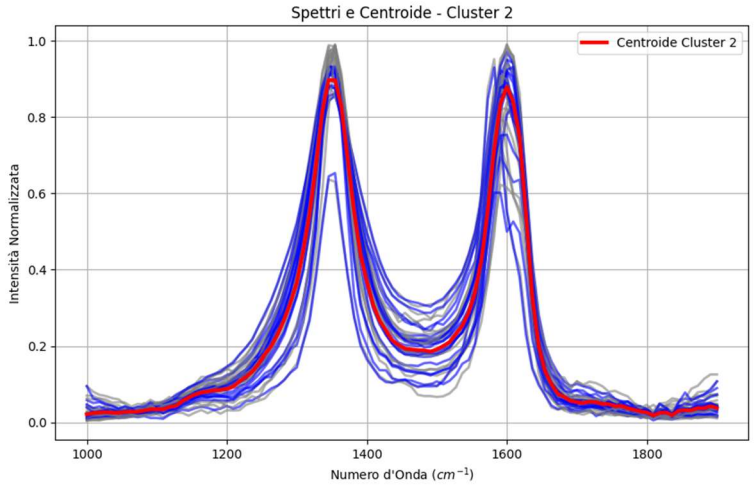


Fig.2: Mildly Graphitized CM.

The second cluster, illustrated in **Figure 2**, reveals a clear transition toward a more disordered carbon structure compared to the previous cluster. While the G-band peak remains at approximately 1580 cm^{-1}, a distinct **D-band peak** now appears at around 1350 cm^{-1}. This spectral signature indicates that the material is no longer near-perfect graphite but a **graphitized carbon with a noticeable degree of structural defects**. The presence of the D-band is a direct indicator of disorder, which can be attributed to factors such as lattice imperfections, edge defects, or smaller crystallite sizes.

The successful isolation of this transitional phase demonstrates the fine-grained discriminatory power of the Dense Autoencoder. The model is not merely separating graphite from non-graphite but is accurately classifying materials along a continuous spectrum of increasing disorder and decreasing crystallinity.

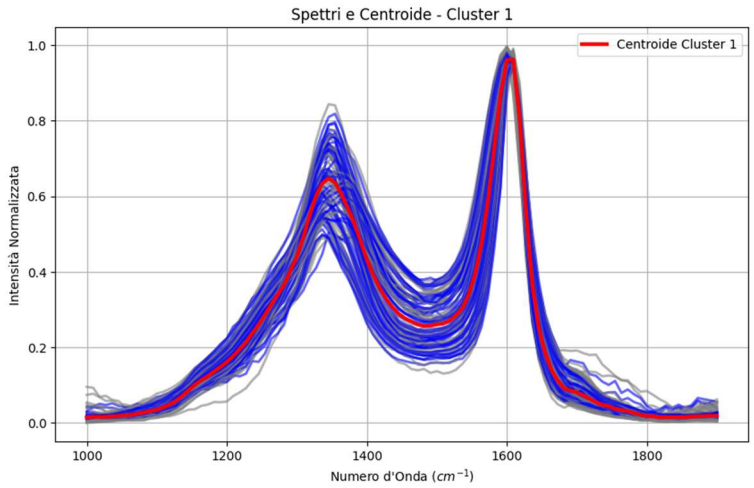


Fig. 3: Intermediate CM.

The third cluster, shown in **Figure 3**, presents a spectral signature of a different kind of disorder. While the relative height of the D-band may appear lower than in previous clusters, a crucial observation is the significant **increase in its width**. This broadening of the D-band, along with the increased width of the G-band, is the key indicator of a new level of complexity in the material's structure. This signature suggests an increase in the number of defective components that contribute to the overall D-band signal. If we were to perform a deconvolution of this pseudospectrum into its underlying Lorentzian or Gaussian components, or q-Gaussians, we would see a rise in the number of defect-related peaks, even if their individual intensities are not high. This implies a higher degree of amorphousness or more varied defect types within the carbon lattice. The material in this cluster is not simply "less graphitic," but rather exhibits a high degree of structural disorganization. For these reasons, it represents intermediate carbon material.

This finding highlights the Dense Autoencoder's ability to not only distinguish between the presence and absence of peaks but also to accurately interpret more subtle, yet scientifically significant, changes in their shape and width. This capacity allows the model to differentiate between various levels of structural disorder, providing a more nuanced classification than a simple intensity ratio would permit.

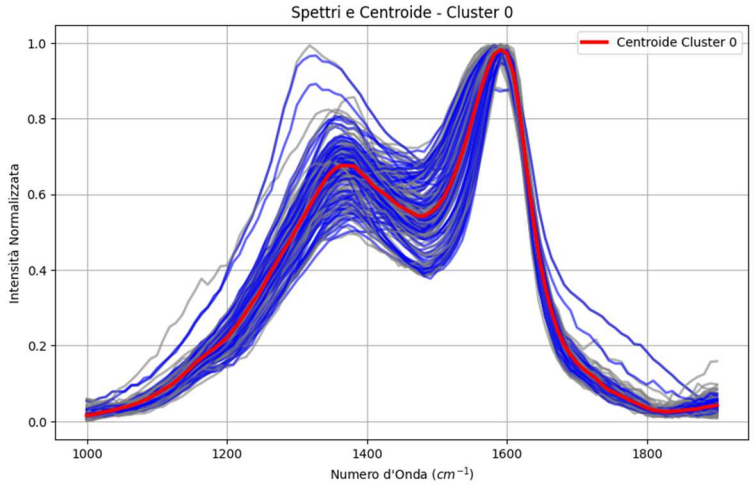
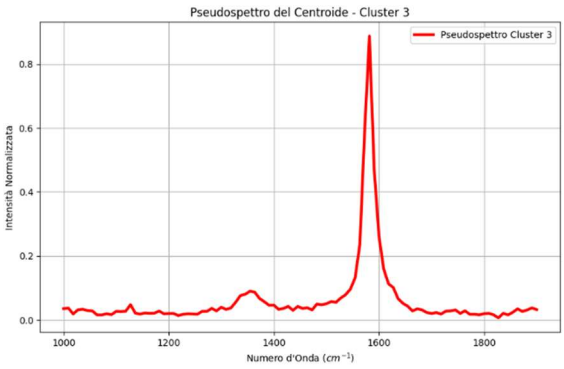


Fig. 4. Disordered CM.

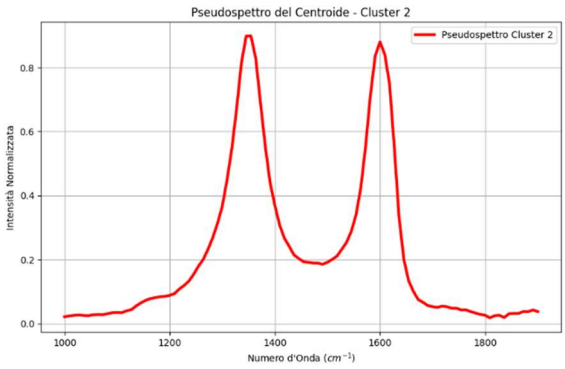
The fourth cluster, depicted in **Figure 4**, represents the significant leap in the structural disorder. In this cluster the relative intensity of the **D-band** at 1350 cm^{-1} has now become dominant, exceeding the role of the **G-band** at 1580 cm^{-1}. Both peaks are notably broad, a clear indication of high structural amorphization. The dominance of the D-band, with its increased peak width, provides a signature of a carbonaceous material that is heavily fragmented and contains a high concentration of defects. The material in this cluster has lost most of its long-range graphitic order, approaching a state of amorphous carbon. This transition is a key marker in the degradation continuum of carbonaceous materials, and the autoencoder's ability to isolate this specific signature further validates its precision as a classification tool.

This finding reinforces the conclusion that the model is capable of distinguishing nuanced stages of material degradation, moving from highly crystalline structures to increasingly disordered ones.

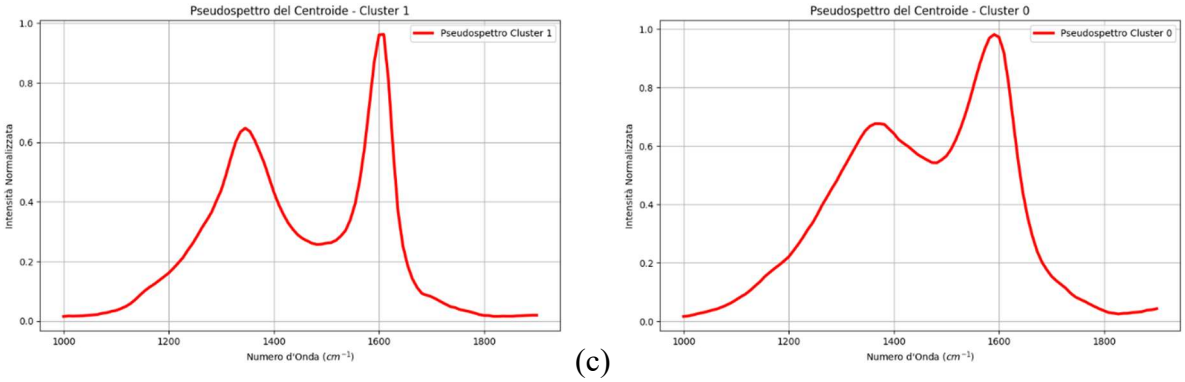
The pseudospectra of the four clusters are given in the following figures. Let us remember that the pseudospectrum is the linear centroid of the cluster.



(a)



(b)



(c) (d)

Fig. 5: Pseudospectra of Highly Graphitized (a), Mildly Graphitized (b), Intermediate (c), and Disordered Carbon Materials.

The Influence of Dataset Size on Model Performance

To further validate our methodology and explore the performance boundaries of the Dense Autoencoder, we tested its classification capability on smaller subsets of the data. This experiment confirmed a crucial principle of machine learning: the size of the dataset is directly proportional to a model's ability to generalize and learn a robust representation of the data.

When the autoencoder was trained on the full dataset of 150 spectra, it had enough examples to learn the fundamental, low-dimensional signatures of the carbonaceous materials. This allowed it to effectively ignore random noise and individual spectral variations. As a result, the reconstructed spectra (**the blue lines**) were remarkably uniform and closely aligned with the ideal pseudospectrum (**the red curve**), demonstrating the model's success in filtering out "unnecessary" information.

However, when the dataset was divided into smaller parts, a different behavior emerged. With fewer examples, the model's ability to generalize was diminished. It began to **overfit** the limited number of spectra, learning not only their core features but also their inherent noise and imperfections. This is precisely why, as observed in our tests, the reconstructed spectra (**the blue lines**) show a slight but noticeable difference from the raw input spectra (**the gray lines**). This expected divergence between the raw and reconstructed data serves as a compelling proof-of-concept for the importance of dataset volume in achieving accurate and generalizable results in unsupervised learning. It reinforces that a large, high-quality dataset is a prerequisite for a model to successfully perform as an autonomous discovery engine.

Experiment B

In this framework, after subdividing the dataset in two parts, we submitted the first part of it (90 spectra) to the Dense autoencoder. The samples in the clusters are given in Appendix B. Here are the plots of the clusters.

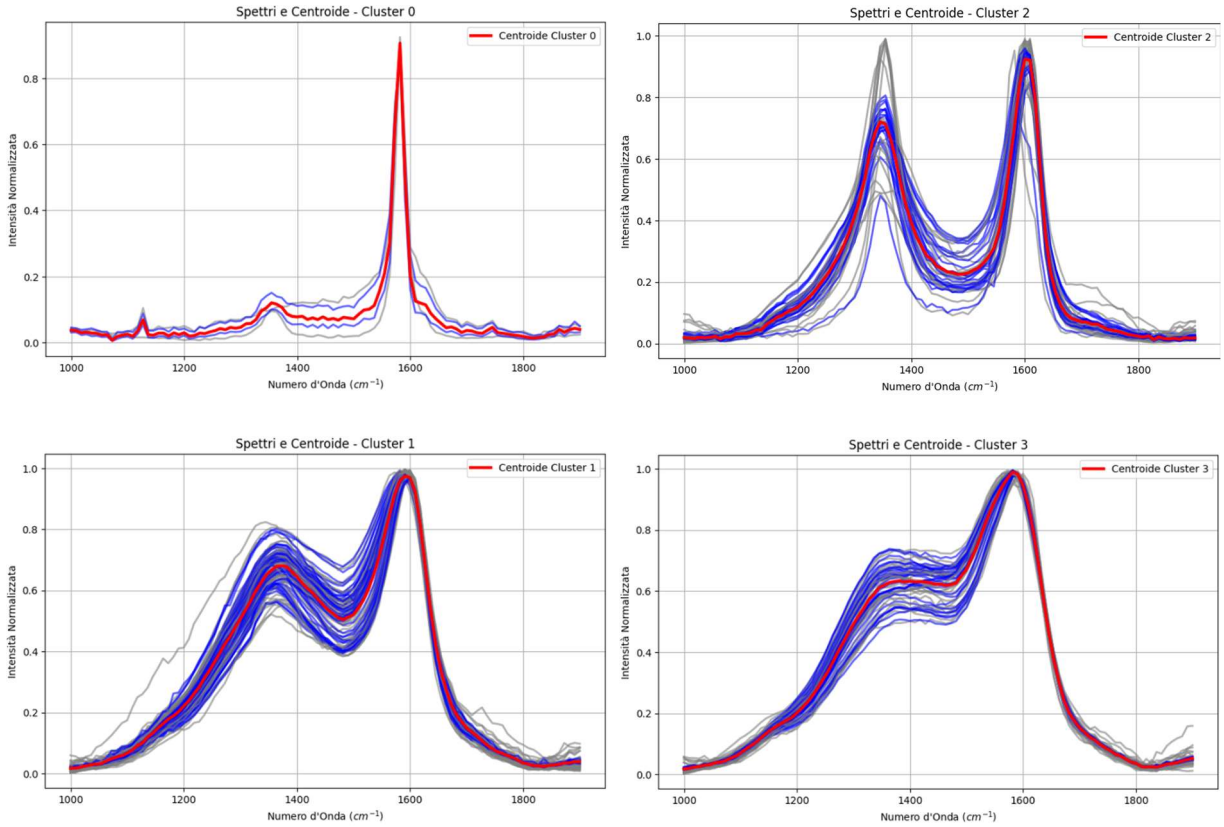


Fig.6: Clusters and pseudospectra (red lines) of the reduced dataset (90 spectra of CM). Note that the pseudospectra of disordered and intermediate CM are different from those in Fig. 3 and 4.

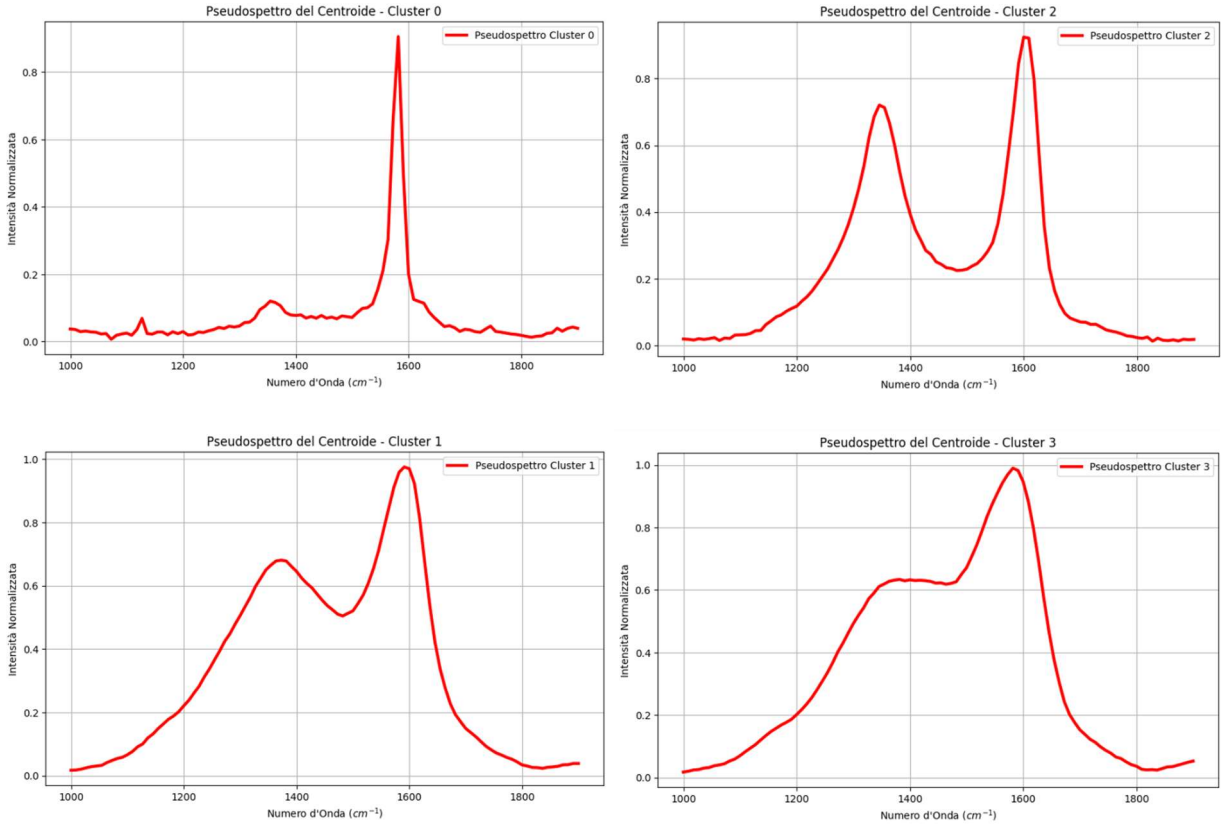


Fig.7: Pseudospectra of dataset with samples given in Appendix B.

While overfitting is a common concern in machine learning, particularly with smaller datasets, our experiment with the 90-element subset provides a crucial insight. Even with a reduced number of spectra, which can lead to a less generalized model, the Dense Autoencoder still produced meaningful and distinct pseudospectra for the intermediate and disordered carbon materials.

This demonstrates a key strength of our approach: the model's ability to extract and retain core, useful information even when faced with data scarcity. While the reconstructed spectra may not be as "perfect" as those from the larger dataset, the fundamental classification is not lost. The pseudospectra still accurately capture the essential differences in the amplitude and width of the G and D bands, allowing for a clear distinction between the clusters. This confirms that the Dense Autoencoder is a robust tool that can provide valuable information even under less-than-ideal training conditions.

Experiment C

We also applied to the dataset of experiment B the "specialist" **1D-Convolutional Autoencoder**, a model that has proven highly effective for the analysis of complex SERS spectra (Sparavigna & Gemini, 2025), asking to determine four clusters. The computation time is much longer than that of the Dense encoder. Here are the results.

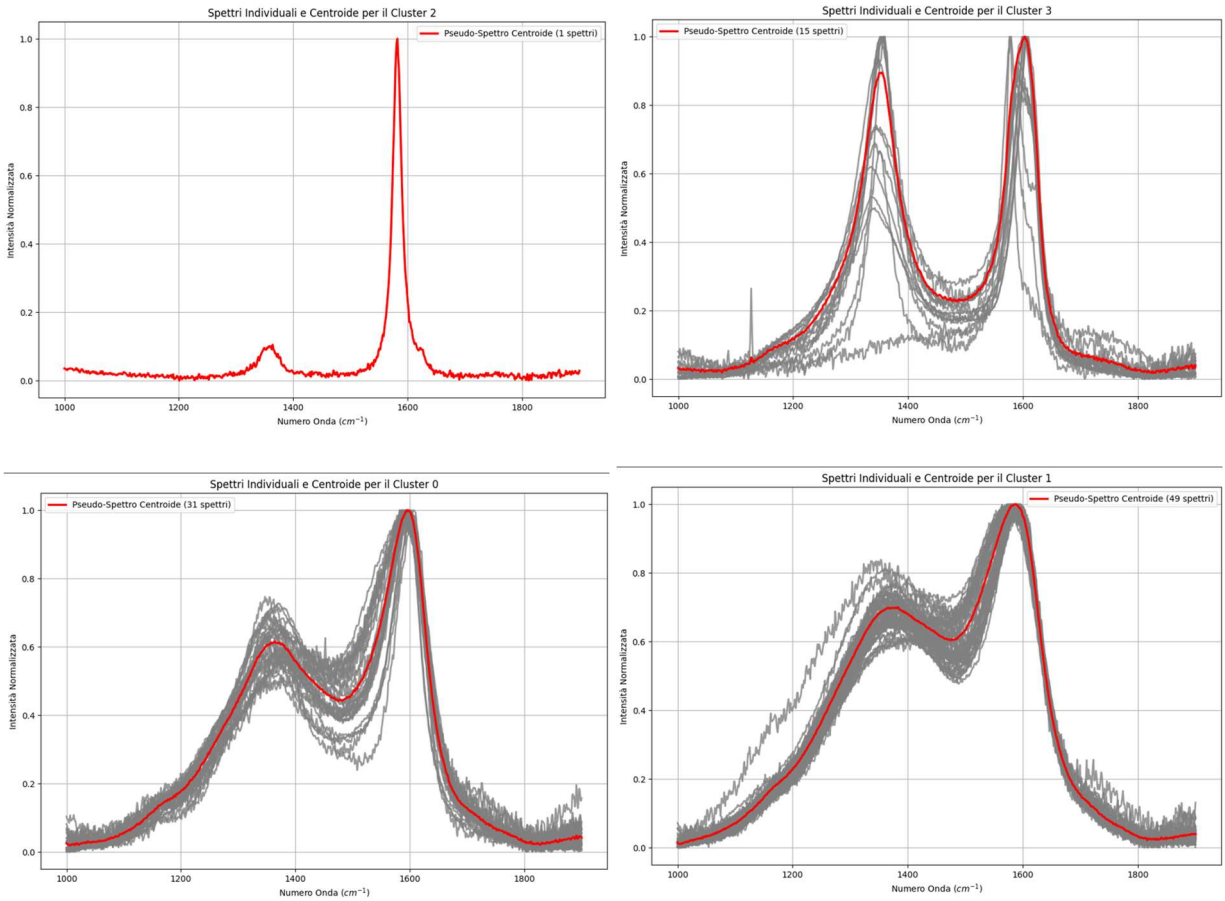


Fig. 8: Clusters obtained by means of the Conv-1D autoencoder.

The difference is that a graphitic spectrum is now in the mildly graphitized cluster. At the same time, a spectrum that in Fig.6 was in the intermediate CM cluster, here is in the disordered carbon material cluster. Note that the pseudospectra of the disordered CM clusters in Fig.6 and Fig.8 are slightly different.

The result is that we can use the Dense autoencoder which is much faster obtaining essentially the same results.

Back to SERS spectra (experiment D)

A final experiment was considered to have further evidence of the behavior of autoencoder Conv-1d. We applied it to the same SERS spectra used in Sparavigna & Gemini, 2025, but with a restricted range (1000–1900 cm^{-1}). In this manner, we proved that the model's effectiveness is not solely dependent on the type of spectrum, but also on its **information density**. The Conv-1D's convolutional architecture requires a high density of features—the numerous and complex peaks characteristic of full-range SERS spectra—to function properly. Without detailed information, the model's specialized filters become less effective, leading to a breakdown in its learning process.

We experimented with the 15 clusters, as in Sparavigna & Gemini, 2025, on the reduced-range SERS dataset (1000-1900 cm^{-1}). By restricting the spectral range, we effectively truncated essential bond vibrations and skeletal modes located below 1000 cm^{-1}, thereby removing the very molecular fingerprint that the Conv-1D Autoencoder was designed to analyze. This provided a critical finding. A detailed chemical analysis of the clusters revealed that the reduced spectral range significantly impacted their coherence. The Conv-1D Autoencoder, built to extract fine features from information-dense spectra, was unable to find robust chemical relationships when starved of the data it relies upon. As a result, it produced several mixed clusters containing chemically disparate molecules, a clear indication that the numerical success of finding 15 clusters does not equate to scientific validity. This result underscores that a model's effectiveness is fundamentally linked to the information density of the data, reinforcing our thesis that the choice of AI architecture is paramount to successful unsupervised analysis.

The outcome is, therefore, a problem of both **range** and **hyperparameter selection**.

Conclusion

In this study, we have demonstrated a **paradigm shift** in the application of artificial intelligence for spectral analysis. We have shown that the effectiveness of an autoencoder is not universal but is fundamentally dependent on the match between its architectural design and the **information density of the data**.

Our "generalist" **Dense Autoencoder** proved to be an unprecedentedly powerful tool for unsupervised classification of carbonaceous materials. It autonomously discovered and extracted the fundamental spectral signatures of distinct material classes, bypassing the need for manual deconvolution. The generated pseudospectra, ideal and noise-free, served as irrefutable proof of its analytical power. Let us also stress that the Dense Autoencoder is a very fast machine learning tool.

To validate this finding, we applied the "specialist" **1D-Convolutional Autoencoder**, a model previously proven highly effective for complex SERS data. This experiment provided that the specialist model produced coherent four pseudospectra from the carbon materials data, proving it was a good tool for the job too. However, the computation time is much longer than that of the Dense autoencoder. Returning to the previously investigated SERS spectra, the same Conv-1D model that performed flawlessly when the spectra were considered on the total available wave-length range, fails when the range is reduced, thereby confirming its inherent design for high-information-density data.

This work serves as a compelling proof-of-concept for the future of analytical chemistry. It is a resounding call for a re-evaluation of how we approach large spectral datasets, proving that AI can move beyond being a mere data processing tool to become an **autonomous discovery engine**. This methodology provides a crucial framework for selecting the right AI model for the right task, promising to accelerate material characterization and unlock new scientific insights with unparalleled efficiency and objectivity.

Appendix: Cluster Assignments and Data Provenance

The following tables (A and B) provide the cluster assignments for each filename in the dataset, allowing for a direct comparison between the original sample provenance and the autoencoder's unsupervised classification.

It is crucial to note that the **autoencoder had no prior knowledge of the filenames, the acronyms they contain, or any other metadata**. The model was trained purely on the spectral data itself. Therefore, the resulting clusters were formed *exclusively* based on the intrinsic similarities of the spectra's G and D bands, as interpreted by the autoencoder. The fact that the model successfully grouped spectra from different original sources (e.g., files beginning with `1132`, `CH-D`, and `KY-D`) into a single cluster is a powerful testament to its ability to identify an underlying chemical signature that transcends simple labeling. This serves as a fundamental validation of our approach.

APPENDIX A

filenamecluster
Cluster 0
1132_Long001.txt 0, 1132_Long012.txt 0, 1132_Long110.txt 0, 1132_Long111.txt 0
1132_Long112.txt 0, 1132_Long113.txt 0, 1132_Long114.txt 0, 1132_Long116.txt 0
1132_Long117.txt 0, 1132_Long118.txt 0, 1132_Long120.txt 0, 1132_Long121.txt 0
1132_Long123.txt 0, 1132_Long124.txt 0, 1132_Long129.txt 0, 1132_Long14.txt 0
1132_Long15.txt 0, 1132_Long16.txt 0, 1132_Long17.txt 0, 1132_Long18.txt 0
1132_Long19.txt 0, 1181_Long01.txt 0, 1181_Long02.txt 0, 1181_Long03.txt 0
1181_Long10.txt 0, 1181_Long11.txt 0, 1181_Long14.txt 0, 1181_Long15.txt 0
1181_Long16.txt 0, 1181_Long17.txt 0, 1181_Long18.txt 0, 1181_Long19.txt 0
1181_Long20.txt 0, 1181_Long21.txt 0, 1181_Long22.txt 0, 1181_Long23.txt 0
1181_Long25.txt 0, 1181_Long26.txt 0, 1181_Long27.txt 0, 1181_Long28.txt 0
1181_Long29.txt 0, 1181_Long30.txt 0, 1181_Long7.txt 0, 1181_Long8.txt 0
1181_Long9.txt 0, 1202_Long02.txt 0, 1202_Long03.txt 0, 1202_Long10.txt 0
1202_Long11.txt 0, 1202_Long14.txt 0, 1202_Long16.txt 0, 1202_Long17.txt 0
1202_Long18.txt 0, 1202_Long20.txt 0, 1202_Long21.txt 0, 1202_Long22.txt 0
1202_Long24.txt 0, 1202_Long28.txt 0, 1202_Long29.txt 0, 1202_Long30.txt 0
1202_Long31.txt 0, 1202_Long32.txt 0, 1202_Long33.txt 0, 1202_Long34.txt 0
1202_Long35.txt 0, 1202_Long36.txt 0, 1202_Long37.txt 0, 1202_Long38.txt 0
1202_Long39.txt 0, 1202_Long40.txt 0, 1202_Long6.txt 0, 1202_Long7.txt 0

1202_Long8.txt 0, 1208_Long01.txt 0, 1208_Long03.txt 0, CH-D-Long02.txt 0
CH-D-Long04.txt 0, CH-D-Long0a.txt 0, CH-D-Long13.txt 0, CH-D-Long16.txt 0
CH-D-Long22.txt 0, CH-D-Long4.txt 0, CH-D-Long5.txt 0, KY-D-Long15.txt 0
KY-D-Long26.txt 0
Cluster 1
1132_Long011.txt 1, 1132_Long115.txt 1, 1132_Long119.txt 1, 1132_Long122.txt 1
1132_Long126.txt 1, 1181_Long13.txt 1, 1181_Long5.txt 1, 1202_Long13.txt 1
1202_Long15.txt 1, 1202_Long23.txt 1, 1208_Long02.txt 1, CH-D-Long002.txt 1
CH-D-Long01.txt 1, CH-D-Long03.txt 1, CH-D-Long10.txt 1, CH-D-Long11.txt 1
CH-D-Long12.txt 1, CH-D-Long15.txt 1, CH-D-Long18.txt 1, CH-D-Long19.txt 1
CH-D-Long1a.txt 1, CH-D-Long20.txt 1, CH-D-Long21.txt 1, CH-D-Long23.txt 1
CH-D-Long7.txt 1, CH-D-Long9.txt 1, KY-D-Long00.txt 1, KY-D-Long01.txt 1
KY-D-Long02.txt 1, KY-D-Long12.txt 1, KY-D-Long13.txt 1, KY-D-Long14.txt 1
KY-D-Long16.txt 1, KY-D-Long18.txt 1, KY-D-Long19.txt 1, KY-D-Long20.txt 1
KY-D-Long21.txt 1, KY-D-Long22.txt 1, KY-D-Long23.txt 1, KY-D-Long24.txt 1
KY-D-Long25.txt 1, KY-D-Long28.txt 1, KY-D-Long29.txt 1, KY-D-Long3.txt 1
KY-D-Long4.txt 1, KY-D-Long6.txt 1, KY-D-Long7.txt 1, KY-D-Long8.txt 1
KY-D-Long9.txt 1
Cluster 2
1132_Long013.txt 2 , 1132_Long125.txt 2, 1132_Long127.txt 2, 1181_Long12.txt 2
1181_Long24.txt 2, 1202_Long19.txt 2, 1202_Long5.txt 2, 1208_Long4.txt 2
CH-D-Long05.txt 2, CH-D-Long0b.txt 2, CH-D-Long14.txt 2, CH-D-Long17.txt 2
CH-D-Long3.txt 2, CH-D-Long6.txt 2, KY-D-Long10.txt 2, KY-D-Long17.txt 2
KY-D-Long27.txt 2, KY-D-Long5.txt 2
Cluster 3
1132_Long128.txt 3 , 1181_Long6.txt 3, KY-D-Long11.txt 3

APPENDIX B

filename cluster

Cluster 0
1132_Long128.txt 0, 1181_Long6.txt 0
Cluster 1
1132_Long111.txt 1, 1132_Long112.txt 1, 1132_Long113.txt 1, 1132_Long114.txt 1
1132_Long116.txt 1, 1132_Long117.txt 1, 1132_Long120.txt 1, 1132_Long121.txt 1
1132_Long123.txt 1, 1132_Long14.txt 1, 1132_Long15.txt 1, 1132_Long17.txt 1
1132_Long18.txt 1, 1132_Long19.txt 1, 1181_Long02.txt 1, 1181_Long03.txt 1
1181_Long10.txt 1, 1181_Long11.txt 1, 1181_Long14.txt 1, 1181_Long16.txt 1
1181_Long17.txt 1, 1181_Long18.txt 1, 1181_Long19.txt 1, 1181_Long20.txt 1
1181_Long21.txt 1, 1181_Long22.txt 1, 1181_Long23.txt 1, 1181_Long25.txt 1
1181_Long26.txt 1, 1181_Long27.txt 1, 1181_Long28.txt 1, 1181_Long29.txt 1
1181_Long30.txt 1, 1181_Long7.txt 1, 1181_Long8.txt 1, 1181_Long9.txt 1
1202_Long17.txt 1, 1202_Long18.txt 1, 1202_Long21.txt 1, 1202_Long22.txt 1
1202_Long28.txt 1, 1202_Long34.txt 1, 1202_Long37.txt 1, 1202_Long38.txt 1
1202_Long39.txt 1, 1202_Long40.txt 1, 1202_Long8.txt 1, 1208_Long01.txt 1
Cluster 2
1132_Long011.txt 2, 1132_Long013.txt 2, 1132_Long115.txt 2, 1132_Long119.txt 2
1132_Long122.txt 2, 1132_Long125.txt 2, 1132_Long126.txt 2, 1132_Long127.txt 2
1181_Long12.txt 2, 1181_Long13.txt 2, 1181_Long24.txt 2, 1181_Long5.txt 2
1202_Long13.txt 2, 1202_Long15.txt 2, 1202_Long19.txt 2, 1202_Long23.txt 2
1202_Long5.txt 2, 1208_Long02.txt 2, 1208_Long4.txt 2,
Cluster 3
1132_Long001.txt 3, 1132_Long012.txt 3, 1132_Long110.txt 3, 1132_Long118.txt 3
1132_Long124.txt 3, 1132_Long129.txt 3, 1132_Long16.txt 3, 1181_Long01.txt 3
1181_Long15.txt 3, 1202_Long02.txt 3, 1202_Long03.txt 3, 1202_Long10.txt 3
1202_Long11.txt 3, 1202_Long14.txt 3, 1202_Long16.txt 3, 1202_Long20.txt 3

1202_Long24.txt	3, 1202_Long29.txt	3, 1202_Long30.txt	3, 1202_Long31.txt	3
1202_Long32.txt	3, 1202_Long33.txt	3, 1202_Long35.txt	3, 1202_Long36.txt	3
1202_Long6.txt	3, 1202_Long7.txt	3, 1208_Long03.txt	3	

References

Beyssac, O., Goffé, B., Chopin, C., & Rouzaud, J. N. (2002). Raman spectra of carbonaceous material in metasediments: a new geothermometer. Journal of metamorphic Geology, 20(9), 859-871.

Fan, X. G., Zeng, Y., Zhi, Y. L., Nie, T., Xu, Y. J., & Wang, X. (2021). Signal-to-noise ratio enhancement for Raman spectra based on optimized Raman spectrometer and convolutional denoising autoencoder. Journal of Raman Spectroscopy, 52(4), 890-900.

Loc, I., Kecoglu, I., Unlu, M. B., & Parlatan, U. (2022). Denoising Raman spectra using fully convolutional encoder–decoder network. Journal of Raman Spectroscopy, 53(8), 1445-1452.

MacQueen, J. (1967). Multivariate observations. In Proceedings ofthe 5th Berkeley Symposium on Mathematical Statisticsand Probability (Vol. 1, pp. 281-297).

Morga, R., & Bielowicz, B. (2022). Raman Spectroscopy of Lignite Gasification Char Morphotypes. Energies, 15(16), 6057.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. IEEE access, 8, 80716-80727.

Sousa, D. V. D., Guimarães, L. M., Felix, J. F., Ker, J. C., Schaefer, C. E. R., & Rodet, M. J. (2020). Dynamic of the structural alteration of biochar in ancient Anthrosol over a long timescale by Raman spectroscopy. PloS one, 15(3), e0229447.

Sparavigna, A. C. (2023). q-Gaussian Tsallis Line Shapes for Raman Spectroscopy. SSRN, n. 4445044. DOI http://dx.doi.org/10.2139/ssrn.4445044

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). Unsupervised CNN e K-means applicati a spettri Raman dei minerali. Zenodo. https://doi.org/10.5281/zenodo.16743376

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). Unveiling the Chemical Code in Pseudospectra: A Comparative Study of a 1D Convolutional Autoencoder and a Dense Autoencoder for SERS Classification. Zenodo. https://doi.org/10.5281/zenodo.16912956

Sparkes, R. B., Maher, M., Blewett, J., Doğrul Selver, A., Gustafsson, Ö., Semiletov, I. P., & Van Dongen, B. E. (2018). Carbonaceous material export from Siberian permafrost tracked across the Arctic Shelf using Raman spectroscopy. The Cryosphere, 12(10), 3293-3309.

Sparkes, R. B., Maher, M., & Blewett, J. (2018) Raw data for paper title "Carbonaceous material export from Siberian permafrost tracked across the Arctic Shelf using Raman spectroscopy". [Dataset]. https://doi.org/10.23634/MMUDR.00620205