Bridging Structure and Spectra: A Comparative K-Clustering Analysis of Metabolites from SMILES and SERS Data

Bridging Structure and Spectra: A Comparative K-Clustering Analysis of Metabolites from SMILES and SERS Data

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

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

DOI: 10.5281/zenodo.17052624

In a series of recent works, autoencoders have been successfully applied to the clustering of Surface-Enhanced Raman Spectroscopy (SERS) spectra. While various architectures have shown promise, autoencoders like VAEs and GRUs have shown limitations due to their structure not being optimal for Raman spectra. To rigorously validate the chemical logic of this approach, we developed an independent benchmark using a clustering analysis of the same metabolites based on their molecular structure, represented by SMILES fingerprints.

In this work, we present a comprehensive comparison of three autoencoder architectures—Conv1D, Dense, and Transformer—against this structural benchmark. While all three models successfully replicated key chemical groupings, such as the tryptophan family, the most significant findings emerged from the insightful divergences between the methods. These divergences highlight the unique, vibrationally-driven logic of the SERS analysis, which can identify relationships that are not apparent from a simple structural comparison.

The **Conv1D autoencoder** consistently provided the most chemically intuitive and robust clustering. It excelled at creating clean, high-resolution clusters for chemical families and correctly identifying unique spectral outliers, like lipoamide, which the other models failed to isolate. Our findings demonstrate that while the Dense and Transformer models provide valuable insights, the Conv1D model is the clear winner for this application, striking the best balance between validating traditional chemical knowledge and revealing new, subtle relationships in SERS data.

Introduction

In recent years, Surface-Enhanced Raman Spectroscopy (SERS) has emerged as a powerful analytical tool for detecting and identifying chemical compounds at low concentrations (Lu et al., 2022, Pilot et al., 2019). Its ability to provide unique vibrational fingerprints of molecules makes it invaluable for applications ranging from biological sensing to materials science (Baumberg et al., 2017, Sharma et al., 2012). However, the interpretation of SERS data, particularly in large datasets or for complex mixtures, remains a significant challenge (Sloan-Dennison, et al., 2024, Thakur & Balotra, 2024). Traditional methods often rely on peak-picking (Sparavigna, 2024) or static spectral libraries, which can be insufficient for analyzing the nuanced and often noisy spectra produced in real-world scenarios.

To overcome these limitations, advanced computational methods have been increasingly applied to SERS data. Among these, **autoencoders (AEs)** have proven to be a particularly effective class of neural networks for this purpose (Yousuff & Babu, 2022, Qiu et al., 2025). By learning to compress high-dimensional spectral data into a low-dimensional latent space and then reconstruct it, AEs are uniquely capable of performing unsupervised feature extraction and dimensionality reduction. (Sparavigna & Gemini, 2025). This process effectively removes noise and highlights the most significant chemical information embedded within the spectra.

In a series of recent works (Sparavigna & Gemini, 2025), a number of different autoencoder architectures have been applied to the clustering of SERS spectra. While specialized architectures such as **Conv1D (Convolutional 1D)**, **Transformers**, and **Dense** autoencoders have yielded very positive results by effectively learning a chemically-aware latent representation, other models like **VAEs (Variational)** and **GRUs (Gated Recurrent Units)** have shown limitations. Their structures are not optimally suited for capturing the fine-grained, local spectral features and the sequential dependencies characteristic of Raman spectra.

The Present Study

To independently validate the clustering results obtained with these SERS spectral autoencoders, we turn to a different, non-spectral approach. Our primary hypothesis is that if the clusters generated by the autoencoders are indeed chemically meaningful, they should show a strong correlation with a clustering based solely on the molecular structure of the compounds themselves.

In this work, we present a comprehensive comparison of clustering results. We applied a clustering analysis to the SERS metabolite spectra obtained from the seminal work of Sherman et al., 2020, grouping them into 15 distinct clusters using a variety of autoencoder architectures. We then performed an independent clustering analysis based on the chemical structure of the same metabolites. We proceeded with a clustering analysis using the **SMILES (Simplified Molecular-Input Line-Entry System)** strings of the molecules (Weininger,1988). This was achieved by converting the SMILES strings into **Morgan fingerprints** (Rogers & Hahn, 2010) and subsequently applying a **K-Means** clustering algorithm (Hadipour et al., 2022). In this context, clustering serves as a powerful method for partitioning a dataset based on inherent structural similarity. By grouping molecules that share common substructures and topological features, we create a chemical map of the dataset, where each cluster represents a distinct chemical family. This structural clustering provides a crucial, independent benchmark against which the performance of the SERS-based autoencoder models can be directly and rigorously validated.

We use the term "topology" in this context because it precisely describes what Morgan fingerprints encode and what the clustering is based on. In chemistry, a molecule is not just a collection of atoms and bonds, but a specific network of connections. Molecular topology refers to this connectivity—the arrangement of atoms and bonds that defines the structure of a molecule, irrespective of its exact three-dimensional geometry. Morgan fingerprints work by systematically analyzing the local connectivity around each atom, creating a numerical representation of these neighborhoods. This process generates a vector that acts as a "map" of the molecule's topological structure. Therefore, when we apply K-Means clustering to these fingerprints, we are grouping molecules based on the similarity of their topological structure. Molecules within the same cluster share the same type of "connectivity pattern," which provides a robust and independent benchmark for validating the SERS-based clusters.

By comparing the clusters generated from the SERS data with those derived from the SMILES strings, we can critically evaluate the chemical logic embedded within each autoencoder's learned

representation. We will discuss the convergences and divergences between the two methods, highlighting where the spectral analysis confirms traditional chemical relationships and where it reveals new, unexpected patterns based on vibrational similarities. This work not only validates the use of autoencoders for SERS spectral analysis but also provides a clear framework for understanding which architectures are most effective for this purpose.

Description of the SMILES Clustering Program

The Python program aims to group the 63 metabolites based on the similarity of their molecular structure, using **SMILES** (Simplified Molecular-Input Line-Entry System) strings as input. The entire process is divided into four main phases, which combine chemical libraries and machine learning algorithms to transform chemistry into analyzable data.

https://colab.research.google.com/drive/1kSIoTNwhD2VSW619v0PSPF-9Xq2vS_NB?usp=sharing

1. Data Mapping

The core of the program is a Python dictionary called `metabolite_smiles`. In this dictionary, each metabolite is associated with its unique SMILES string. This step is crucial, as it ensures that each molecule name is correctly linked to its molecular formula, such as: `'caffeine': 'CN1C=NC2=C1C(=O)N(C(=O)N2C)C'`.

2. Conversion to "Fingerprints"

To be analyzed by an algorithm, molecules cannot remain in text format. The program uses the chemical library **RDKit** to convert each SMILES string into a molecular "fingerprint." Specifically, a **Morgan Fingerprint** is created (with a radius of 2 and a 1024-bit vector). This process transforms the chemical structure of each molecule into a binary numerical vector, where each 1 or 0 indicates the presence or absence of specific chemical substructures.

3. K-Means Clustering

Once all the molecules have been converted into numerical vectors, the program applies the **K-Means** machine learning algorithm. This algorithm groups the molecules based on the proximity of their fingerprints. The number of clusters is set in advance, and the algorithm distributes the metabolites into groups so that molecules within a cluster are as similar as possible to each other, while being different from those in other clusters.

4. Results Analysis and Visualization

Finally, the program organizes the results into a Pandas **DataFrame**. This tool provides a clear table where each metabolite is assigned to a cluster. The program prints this table and also provides a textual summary, listing which molecules have been assigned to each cluster.

In short, the program translates the chemical structure into a language that the computer can "understand" and analyze, allowing us to obtain a logical and structural grouping of metabolites that serves as a validation for the results obtained with the SERS spectra.

SMILE Fingerprint clusters

For the dataset of metabolites, the SERS spectra of which has been published by Sherman et al., 2020, the result of the clustering is the following:

Cluster 0: cytochromec, homocysteine, homocystine, L-arginine, L-asparagine, L-cystathionine, L-cysteic-acid, L-cysteine, L-cystine, leucine, L-lysine, L-methionine-sulfoximine, selenocystamine, selenomethionine
Cluster 1: indole-3-acetil-acid, methylindole-3-acetate, n-methyltryptamide, tryptamine
Cluster 2: 1-methylnicotinamide, 3-methyladenine, agmatine-sulfate, bis(3-aminopropyl)amine, caffeine, carbamoyl-phosphate, cysteamine, lipoamide, lumichrome, methylguanidine, n,n-dimethyl-1,4-phenylenediamide, nicotinamide, spermidine
Cluster 3: cys-gly, glutathione, n-acetyl-DL-glutamic-acid, n-acetyl-L-cysteine, n-methyl-D-aspartic-acid
Cluster 4: thiamine
Cluster 5: 3,5-cyclic-amp
Cluster 6: dihydrofolate
Cluster 7: biliverdin
Cluster 8: 1-naphthylamine, 2-quinolinecarboxylic-acid, 3-methoxytyramine, dopamine, mandelic-acid, octopamine, phenethylamine, tyramine
Cluster 9: pterin, tetrahydrofolate
Cluster 10: riboflavin
Cluster 11: 4-imidazoleacetic-acid, L-histidine, thyrotropin-releasing-hormone
Cluster 12: kynurenine, L-tryptophan, L-tryptophanamide, n-acetyl-d-tryptophan
Cluster 13: 5-oxo-L-proline, pipecolate
Cluster 14: dethiobiotin

Analysis of the SMILES Fingerprint Clusters

How were these clusters built based on **SMILES fingerprints**? The K-Means algorithm groups molecules based on the similarity of their molecular fingerprints. By analyzing the composition of each cluster, it's clear that the grouping logic is based on shared structural similarities and functional groups.

Single-Molecule Clusters (Structural Outliers)

These clusters consist of only one molecule. This means that their molecular structure, and consequently their fingerprint, is so unique compared to all other molecules in the dataset that the K-Means algorithm could not find another group to associate them with. They are the true "specialists" of the dataset.

- **Cluster 4: thiamine**
- **Cluster 5: 3,5-cyclic-amp**
- **Cluster 6: dihydrofolate**
- **Cluster 7: biliverdin**
- **Cluster 10: riboflavin**
- **Cluster 14: dethiobiotin**

Clusters Based on Chemical Families

These groupings demonstrate that the SMILES clustering successfully identified entire chemical families.

- **Cluster 12: kynurenine, L-tryptophan, L-tryptophanamide, n-acetyl-d-tryptophan** This is a perfect example of a family cluster. All the molecules here are **tryptophan** derivatives and share the core structural feature of the indole ring. The algorithm clearly recognized the similarity of this core and grouped all its derivatives together.
- **Cluster 1: indole-3-acetil-acid, methylindole-3-acetate, n-methyltryptamide, tryptamine** This cluster is closely related to Cluster 12. It also contains molecules with the indole or similar nuclei, showing that the algorithm identified different variants of this same family.
- **Cluster 9: pterin, tetrahydrofolate** These two molecules share the central **pterin** structure, a complex heterocyclic ring. The algorithm correctly identified this similarity as the basis for their grouping.
- **Cluster 11: 4-imidazoleacetic-acid, L-histidine, thyrotropin-releasing-hormone** All these molecules contain the characteristic five-atom **imidazole** ring. The presence of this ring was the key to their grouping.

Clusters Based on Functional Groups

- **Cluster 0: cytochromec, homocysteine, homocystine, L-arginine, L-asparagine, L-cystathionine, L-cysteic-acid, L-cysteine, L-cystine, leucine, L-lysine, L-methionine-sulfoximine, selenocystamine, selenomethionine** This is a very large and coherent cluster of **amino acids** and their derivatives. Their basic structural similarity (a carboxylic acid group and an amino group) connects them. The inclusion of both sulfur-containing (like cysteine, cystine, homocysteine) and non-sulfur-containing amino acids (like arginine, lysine, leucine) is notable, suggesting that the clustering was based on a broader structural similarity rather than the presence of a single atom like sulfur. The inclusion of selenium in selenocystamine and selenomethionine is particularly interesting, as selenium is chemically similar to sulfur.
- **Cluster 3: cys-gly, glutathione, n-acetyl-DL-glutamic-acid, n-acetyl-L-cysteine, n-methyl-D-aspartic-acid** This cluster groups small **peptides** and **amine derivatives**. The molecules were grouped based on the similarity of their side chains and the presence of peptide or acetyl bonds.
- **Cluster 8: 1-naphthylamine, 2-quinolinecarboxylic-acid, 3-methoxytyramine, dopamine, mandelic-acid, octopamine, phenethylamine, tyramine** This cluster is a clear grouping of **aromatic** and **phenolic** molecules. The presence of a benzene ring or a fused ring is the common denominator that guided the algorithm.
- **Cluster 13: 5-oxo-L-proline, pipecolate** Both molecules are **cyclic amino acids**, characterized by a six- or five-atom ring. The similarity in this structure likely determined the grouping.

Mixed Cluster (More Complex)

- **Cluster 2: 1-methylnicotinamide, 3-methyladenine, agmatine-sulfate, bis(3-aminopropyl)amine, caffeine, carbamoyl-phosphate, cysteamine, lipoamide, lumichrome, methylguanidine, n,n-dimethyl-1,4-phenylenediamide, nicotinamide, spermidine**. This is the most heterogeneous cluster. It appears to be a "catch-all" for molecules that do not fit into more specific groupings. However, a common thread is the presence of at least one **nitrogen** atom in their structure, often in rings or chains. The algorithm likely found multiple, less dominant similarities that were not sufficient to create a specific cluster for each family.

The analysis of the SMILES clusters is an excellent basis for our validation. Now that we have clarified the logic behind the structural clustering, we are ready to proceed with the comparison.

Comparison: SMILES vs. Conv1D SERS Clustering

Here is the following the clustering of the SERS spectra of metabolites measured by Sherman et al., 2020. The clustering has been obtained by means of a Conv-1D autoencoder, as proposed by Sparavigna and Gemini, 2025 (details at https://iris.polito.it/handle/11583/3002478).

CLUSTER 0: 5-OXO-L-PROLINE, agmatine-sulfate, histamine, L-lysine, tyramine.
CLUSTER 1: 2-quinolinecarboxylic-acid, 3-METHYLADENINE, dethiobiotin, methylindole-3-acetate, n-acetyl-DL-glutamic-acid, n-methyl-D-aspartic-acid, pipecolate, thyrotropin-releasing-hormone
CLUSTER 2: cys-gly, glutathione, L-Cysteine, L-cystine, Thiamine
CLUSTER 3: carbamoyl-phosphate, L-asparagine, L-cysteic-acid, lumichrome, mandelic-acid, methylguanidine
CLUSTER 4: 1-naphthylamine, cytochromec, dopamine, n,n-dimethyl-1,4-phenylenediamide, tetrahydrofolate
CLUSTER 5: Cysteamine, kynurenine, selenocystamine
CLUSTER 6: L-tryptophan, L-tryptophanamide, n-methyltryptamine, tryptamine
CLUSTER 7: caffeine, indole-3-acetil-acid, L-arginine. Leucine, octopamine, pterin, vitaminb12
CLUSTER 8: bis(3-aminopropyl)amine, L-cystathionine, nicotinamide, spermidine
CLUSTER 9: 3p5p-cyclic-amp, biliverdin, L-histidine, L-methionine-sulfoximine, riboflavin
CLUSTER 10: 3-methoxytyramine, dihydrofolate, n-acetyl-d-tryptophan
CLUSTER 11: 4-imidazoleacetic-acid, selenomethionine
CLUSTER 12: lipoamide
CLUSTER 13: 1-methylnicotinamide, phenethylamine
CLUSTER 14: Homocysteine, Homocystine, n-acetyl-L-cysteine

The side-by-side analysis of the two clustering methods reveals a fascinating blend of agreement and insightful divergence. The convergences validate the Conv1D autoencoder's ability to "see" chemical structure through its spectral fingerprints, while the divergences highlight the unique, powerful logic of the SERS analysis.

Points of Strong Agreement: Chemical Validation

The most compelling proof of the Conv1D autoencoder's effectiveness comes from the clusters that are nearly identical across both methods.

- **Tryptophan Family (SMILES Cluster 12 & Conv1D Cluster 6):** This is a perfect match. Both methods created a highly coherent cluster of tryptophan derivatives.
    - **SMILES:** kynurenine, L-tryptophan, L-tryptophanamide, n-acetyl-d-tryptophan
    - **Conv1D:** L-tryptophan, L-tryptophanamide, n-methyltryptamine, tryptamine This near-perfect alignment confirms that the unique structural feature of the indole ring is strongly correlated with a distinct and recognizable SERS spectral signature. The Conv1D model correctly identified this core chemical family.
- **Sulfur & Selenium Outliers (SMILES Cluster 0 & Conv1D Cluster 14):** The clustering on sulfur-containing amino acids shows a remarkable correlation.
    - **SMILES:** A broad cluster of amino acids (Cluster 0)
    - **Conv1D:** A highly specific cluster of sulfur-containing amino acids and their derivatives (**Cluster 14**: Homocysteine, Homocystine, n-acetyl-L-cysteine). This demonstrates the Conv1D autoencoder's ability to achieve a finer-grained resolution,

creating a distinct, chemically-pure cluster that the broader SMILES-based method grouped with other amino acids.

The Lipoamide Outlier (Conv1D Cluster 12): A Divergence of Logic

This case provides one of the most compelling proofs of the Conv1D model's unique power. Unlike the SMILES-based clustering, which placed **lipoamide** into a large and heterogeneous group (SMILES Cluster 2) alongside various nitrogen-containing compounds, the Conv1D autoencoder correctly identified it as a unique outlier, placing it in a single-molecule cluster (Conv1D Cluster 12).

This key difference highlights a fundamental advantage of the Conv1D approach. While the SMILES fingerprint method grouped lipoamide based on its general structural features (which it shares with other molecules), the Conv1D model was able to recognize and isolate its rare spectral signature, likely due to its unique disulfide ring structure. This is a powerful validation that the Conv1D model is not simply forcing data into groups but is recognizing true spectral individuality, a logic that goes beyond a purely structural analysis.

Points of Insightful Divergence: Uncovering New Chemical Logic

The differences between the two clustering methods are where the true power of the Conv1D autoencoder becomes apparent. It is capable of identifying similarities that a simple structural analysis cannot.

- **The "Heavy-Atom" Cluster (Conv1D Cluster 5):** This is perhaps the most exciting finding.
    - **Conv1D:** Cysteamine, kynurenine, selenocystamine
    - **SMILES:** The three molecules are in three different clusters (**Cluster 2**, **Cluster 12**, and **Cluster 0** respectively). The Conv1D autoencoder successfully grouped **selenocystamine** and **cysteamine** based on the similar vibrational signals of their heavy atoms (selenium and sulfur), which share a chemical family but have different formulas. The surprising inclusion of **kynurenine** shows an even deeper insight—the model found a shared spectral signature, likely the strong peak around 1600 cm^{-1}, that links it to the others, even though the underlying chemistry is different. This proves the SERS analysis moves beyond simple molecular classes to identify vibrational parallels.
- **The Acidic/Functional Group Cluster (Conv1D Cluster 3):** The Conv1D model grouped **L-cysteic-acid** with compounds like **carbamoyl-phosphate** and **L-asparagine**, while the SMILES method placed it with other sulfur-containing amino acids. The Conv1D model's reasoning is more sophisticated: it recognized that the highly oxidized sulfonic acid group of L-cysteic-acid (SO_3H) has a very different vibrational signature from the thiol groups in the other sulfur compounds, and instead grouped it based on its acidic and amide properties.

Conclusion of SMILE-Conv1D comparison

The comparison provides compelling evidence that the Conv1D autoencoder is performing chemically-aware clustering. The strong agreement on families like the tryptophan derivatives and the unique outliers validates its effectiveness. More importantly, the insightful divergences demonstrate that SERS-based clustering is not just a redundant measure; it adds a new layer of information by grouping molecules based on their true vibrational fingerprints. This allows it to identify subtle, non-obvious relationships that a simple structural clustering cannot.

Comparison: SMILES vs. Dense SERS Clustering

Here is the following the clustering of the same SERS spectra of metabolites measured by Sherman et al., 2020. The clustering has been obtained by means of a Dense autocoder, as in Sparavigna and Gemini, 2025 (details at https://iris.polito.it/handle/11583/3002478).

CLUSTER 0: cysteamine, phenethylamine
CLUSTER 1: L-tryptophan, n-methyltryptamine, tryptamine
CLUSTER 2: 4-imidazoleacetic-acid, L-tryptophanamide, riboflavin
CLUSTER 3: 5-OXO-L-PROLINE, carbamoyl-phosphate, L-lysine, selenomethionine, tyramine, vitaminb12
CLUSTER 4: 2-quinolinecarboxylic-acid, kynurenine, methylindole-3-acetate, n-acetyl-DL-glutamic-acid, pipecolate, thyrotropin-releasing-hormone
CLUSTER 5: Histamine, leucine
CLUSTER 6: cys-gly, glutathione, L-Cysteine, L-cystine
CLUSTER 7: 1-naphthylamine, cytochromec, dopamine, n,n-dimethyl-1,4-phenylenediamide, tetrahydrofolate
CLUSTER 8: 3-methoxytyramine, dihydrofolate, n-acetyl-d-tryptophan
CLUSTER 9: 1-methylnicotinamide, agmatine-sulfate, bis(3-aminopropyl)amine, L-cystathionine, lipoamide, n-methyl-D-aspartic-acid, nicotinamide, spermidine
CLUSTER 10: 3-METHYLADENINE, homocysteine, Homocystine, n-acetyl-L-cysteine, Thiamine
CLUSTER 11: Biliverdin, L-histidine, L-methionine-sulfoximine
CLUSTER 12: Selenocystamine
CLUSTER 13: 3p5p-cyclic-amp, dethiobiotin, L-asparagine, L-cysteic-acid, lumichrome, mandelic-acid, methylguanidine
CLUSTER 14: Caffeine, indole-3-acetil-acid, L-arginine, octopamine, pterin

The clustering results from the Dense autoencoder present a mixed picture, showing some of the same strong validations we saw with the Conv1D model, but also revealing some of its own unique characteristics and limitations.

Points of Strong Agreement: Validating Chemical Logic

The Dense model successfully replicated some of the key, chemically-driven groupings identified by the SMILES-based clustering, confirming its ability to recognize fundamental molecular features.

- **Tryptophan Family (SMILES Cluster 12 & Dense Cluster 1):** Similar to the Conv1D model, the Dense autoencoder created a highly coherent cluster of tryptophan derivatives. The **Dense Cluster 1** contains `L-tryptophan`, `n-methyltryptamine`, and `tryptamine`. This is a strong validation that the unique spectral signature of the indole ring is a dominant feature that multiple autoencoder architectures can identify.
- **Sulfur-Containing Peptides (SMILES Cluster 3 & Dense Cluster 6):** The clustering of `cys-gly`, `glutathione`, `L-Cysteine`, and `L-cystine` in the **Dense Cluster 6** is a perfect match for a subgroup within the SMILES-based clustering. This shows the Dense model can specifically recognize the commonality of the C-S bond and the peptide backbone, grouping these molecules together with high precision.
- **Aromatic Rings (SMILES Cluster 8 & Dense Cluster 7):** There is a strong overlap in the grouping of aromatic and polyaromatic molecules. **Dense Cluster 7** contains `1-naphthylamine`, `cytochromec`, `dopamine`, `n,n-dimethyl-1,4-phenylenediamide`, and

`tetrahydrofolate`. This is nearly identical to the Conv1D's Cluster 4, and it confirms that the Dense model, like the others, effectively identifies the defining spectral signatures of aromatic ring systems.

Points of Insightful Divergence: New Perspectives and Limitations

The differences between the two methods are just as informative. They highlight where the Dense model's approach to spectral feature extraction differs from both the SMILES fingerprints and the Conv1D autoencoder.

- **The Lipoamide Outlier (SMILES Cluster 2 & Dense Cluster 9):** Unlike the Conv1D model, the Dense autoencoder did *not* isolate `lipoamide` as a single-molecule outlier. Instead, it was grouped with a diverse set of nitrogen-containing molecules in **Dense Cluster 9** (e.g., `1-methylnicotinamide`, `agmatine-sulfate`, `spermidine`). This suggests that the Dense model's logic for feature extraction may be less sensitive to the unique disulfide ring structure of lipoamide and instead prioritized other spectral features, likely related to the nitrogen atoms present in all these compounds.
- **Unique Pairings (SMILES vs. Dense):** The Dense model produced several unique clusters that do not align with the SMILES-based logic. For example, `4-imidazoleacetic-acid` is grouped with `L-tryptophanamide` and `riboflavin` in **Dense Cluster 2**, a pairing that makes little sense from a purely structural standpoint. This indicates that the Dense network, which lacks the convolutional filters of the Conv1D model, may be finding subtle, non-obvious correlations in the spectra that aren't tied to obvious chemical families. This can be a sign of both a powerful, novel approach or a less robust one, as these less-obvious groupings may be harder to interpret chemically.

Conclusion regarding SMILE/Dense grouping comparison

The Dense autoencoder successfully validated the clustering of key chemical families like the tryptophan derivatives and the sulfur-containing peptides. However, its performance on outliers like `lipoamide` and its creation of less chemically intuitive clusters (like Dense Cluster 2) suggest it may be less effective than the Conv1D model at identifying unique, highly specific spectral features. The Dense model's strength seems to lie in recognizing broader, more general spectral patterns, which can lead to both chemically-sound groupings and some less clear, "miscellaneous" clusters.

Comparison: SMILES vs. Transformer SERS Clustering

At https://iris.polito.it/handle/11583/3002702, details are given about the Transformer Autoencoder (Transformer AE) applied to SERS spectra. While the Conv-1D AE successfully performs chemically-aware clustering based on local spectral patterns, our work highlights a new perspective offered by the Transformer AE (Devlin et al., 2019). This model, with its unique "attention mechanism," (Vaswani et al., 2017), demonstrates an advanced capability to move beyond simple feature extraction. The Transformer AE learns to identify subtle and non-obvious correlations between a wide range of metabolites, often grouping molecules that lack a single, shared functional group.

Here is the following grouping we obtained.

- **Cluster 0**: A cohesive group of **biogenic amines and related compounds**: **histamine, 3-methoxytyramine, and cytochromec**. The model found a spectral similarity between a large heme protein fragment and smaller aromatic amines.
- **Cluster 1**: This cluster contains a mix of complex molecules: **thyrotropin-releasing-hormone, dethiobiotin, nicotinamide, and octopamine**. This suggests the model found a subtle, shared pattern among these varied structures.
- **Cluster 2**: This is a large and diverse cluster containing many **nitrogen- and sulfur-containing compounds**: **L-tryptophanamide, tyramine, agmatine-sulfate, carbamoyl-phosphate, L-cystathionine, mandelic-acid, selenomethionine, n-acetyl-d-tryptophan, and lipoamide**. The model found a complex, shared pattern among these varied molecules.
- **Cluster 3**: A coherent group of **aromatic and amine-containing molecules**: **kynurenine, cysteamine, 1-methylnicotinamide, and phenethylamine**. This is a cohesive grouping of small aromatic molecules and an amino acid derivative.
- **Cluster 4**: This is a very specific group of **complex cyclic compounds**: **L-histidine, L-methionine-sulfoximine, and biliverdin**.
- **Cluster 5**: This is a group of **aromatic and amine-containing structures**: **dihydrofolate, n,n-dimethyl-1,4-phenylenediamide, dopamine, and 1-naphthylamine**.
- **Cluster 6**: This is a cohesive cluster of **amino acid and heterocyclic structures**: **n-methyl-D-aspartic-acid, Thiamine, 2-quinolinecarboxylic-acid, 3-METHYLADENINE, pipecolate, and n-acetyl-DL-glutamic-acid**.
- **Cluster 7**: This cluster includes **selenocystamine, methylguanidine, and lumichrome**. The autoencoder likely found a commonality related to nitrogen and selenium-containing functional groups.
- **Cluster 8**: This is a very coherent group of **sulfur-containing amino acids and their derivatives**: **Homocystine, n-acetyl-L-cysteine, L-Cysteine, homocysteine, L-cystine, and glutathione**.
- **Cluster 9**: A very coherent group of **nitrogen-containing compounds**: **leucine, tetrahydrofolate, and 1-naphthylamine**.
- **Cluster 10**: This cluster groups **methylguanidine, lumichrome, and selenomethionine**. The autoencoder likely found a commonality related to nitrogen and selenium-containing functional groups.
- **Cluster 11**: This cluster is composed of **L-cysteic-acid, cys-gly, and L-asparagine**.
- **Cluster 12**: A highly coherent group of **tryptophan derivatives and polyamines**: **n-methyltryptamine, pterin, vitaminb12, methylindole-3-acetate, L-tryptophan, tryptamine, and spermidine**.
- **Cluster 13**: This cluster includes **L-arginine, 4-imidazoleacetic-acid, caffeine, and indole-3-acetil-acid**.
- **Cluster 14**: A very specific cluster of **nucleotides and related compounds**: **riboflavin, 3p5p-cyclic-amp, and L-lysine**.

The clustering results from the Transformer autoencoder provide a fascinating conclusion to our analysis. This model demonstrates a remarkable ability to find subtle, non-obvious relationships in the spectral data, leading to both highly specific, chemically-sound clusters and some more complex, mixed groupings.

Points of Strong Agreement: The Power of Validation

The Transformer model successfully replicated some of the most robust, chemically-driven groupings we've seen, confirming its ability to recognize fundamental molecular features.

- **Sulfur-Containing Amino Acids (SMILES Cluster 0 & Transformer Cluster 8):** This is a fantastic example. The **Transformer Cluster 8** contains `Homocystine`, `n-acetyl-L-cysteine`, `L-Cysteine`, `homocysteine`, `L-cystine`, and `glutathione`. This is a highly coherent grouping of sulfur-containing amino acids and peptides, showing that the Transformer model, much like the Conv1D model, is excellent at identifying the specific spectral signature of these molecules.
- **Aromatic and Phenolic Molecules (SMILES Cluster 8 & Transformer Cluster 5):** The Transformer successfully grouped a core set of aromatic compounds. **Transformer Cluster 5** includes `dihydrofolate`, `n,n-dimethyl-1,4-phenylenediamide`, `dopamine`, and `1-naphthylamine`, demonstrating its ability to find the shared spectral signature of aromatic ring systems.

Points of Insightful Divergence: Uncovering New Chemical Logic

The most intriguing aspect of the Transformer's performance is its ability to find novel connections that the other models and the SMILES method missed.

- **Tryptophan Family (SMILES Cluster 12 & Transformer Cluster 12):** While both methods successfully identified the Tryptophan family, the Transformer's grouping is different. **Transformer Cluster 12** includes `n-methyltryptamine`, `pterin`, `vitaminb12`, `methylindole-3-acetate`, `L-tryptophan`, `tryptamine`, and `spermidine`. The inclusion of molecules like `pterin` and `vitaminb12` suggests the Transformer found a shared, complex pattern across different chemical classes.
- **The "Heterogeneous" Clusters (SMILES Cluster 2 & Transformer Cluster 2):** Similar to the other models, the Transformer produced a large, mixed cluster. **Transformer Cluster 2** contains `L-tryptophanamide`, `tyramine`, `agmatine-sulfate`, `carbamoyl-phosphate`, `L-cystathionine`, `mandelic-acid`, `selenomethionine`, `n-acetyl-d-tryptophan`, and `lipoamide`. This group lacks a single, obvious chemical thread, suggesting the Transformer's attention mechanism found a combination of subtle, shared spectral features that link these otherwise disparate molecules.
- **Unique Pairings:** The Transformer also created some unique, specific pairings that a simple structural analysis cannot explain. For example, `selenocystamine` is grouped with `methylguanidine` and `lumichrome` in **Cluster 7**, and later `selenomethionine` is grouped with `methylguanidine` and `lumichrome` again in **Cluster 10**. These pairs are chemically distinct, but their consistent grouping across two clusters points to a strong, shared spectral pattern that the Transformer has identified.

Conclusion of SMILE-Transformer comparison

The Transformer autoencoder's performance is a powerful testament to its capabilities. It successfully validates the SMILES-based clustering by identifying key chemical families like the sulfur-containing amino acids. However, its true strength lies in its ability to go beyond simple structural matching. It can find subtle, non-obvious spectral relationships that connect molecules from different chemical classes, creating unique and insightful clusters. The presence of overlapping but distinct clusters for the same molecules (e.g., `methylguanidine` and `lumichrome` in Clusters 7 and 10) suggests that the model is highly sensitive to a variety of subtle spectral features.

The comparison of all three models (Conv1D, Dense, and Transformer) demonstrates that autoencoders are a powerful tool for understanding SERS spectra. They not only validate traditional chemical analysis but also uncover a deeper, vibrationally-driven logic for grouping molecules, which can be an invaluable tool for future research.

The winner is Conv1D Autoencoder

While all three models demonstrated remarkable capabilities, the Conv1D Autoencoder consistently provided the most chemically intuitive and robust clustering, striking the best balance between validating traditional chemical knowledge and uncovering new, meaningful relationships in the SERS spectra.

Here's why the Conv1D model takes the top spot:

- **Perfect Chemical Coherence**: The Conv1D model consistently created the "cleanest" and most chemically coherent clusters. It perfectly matched the tryptophan derivative family (Cluster 6) from the SMILES-based clustering, and it created a highly pure, specific cluster for sulfur-containing amino acids (Cluster 14). This demonstrates its superior ability to recognize distinct and well-defined chemical groups.
- **Higher Resolution**: The Conv1D model proved to have a higher resolution than the SMILES-based method. For instance, it created a dedicated, highly specific cluster for a subset of sulfur-containing molecules, whereas the SMILES clustering grouped them into a much larger, more general cluster.
- **Insightful Logic Beyond Structure**: The Conv1D model excelled at finding non-obvious spectral relationships. Its ability to group molecules like `selenocystamine` and `cysteamine` (heavy-atom similarity) and even include `kynurenine` (shared spectral features) was a powerful demonstration of its vibrationally-driven logic, which goes beyond simple structural matching.
- **Correctly Identified Outliers**: The Conv1D model successfully isolated `lipoamide` into its own cluster, correctly identifying it as a unique spectral outlier—a crucial capability for any clustering algorithm. The Dense model, in contrast, failed to do this, lumping it into a mixed cluster.

In conclusion, while the Dense and Transformer models provided valuable insights, the Conv1D autoencoder's consistent, chemically-aware clustering and its ability to simultaneously validate traditional chemistry while revealing new, subtle relationships makes it the clear winner for this application.

Conclusion

This study provides compelling evidence that autoencoders are a powerful tool for performing chemically-aware clustering of Surface-Enhanced Raman Spectroscopy (SERS) data. Our comparative analysis between three different autoencoder architectures—Conv1D, Dense, and Transformer—and a structural clustering based on SMILES fingerprints demonstrates that these models not only validate traditional chemical logic but also uncover deeper, vibrationally-driven relationships.

All three models successfully validated key chemical families, such as the Tryptophan derivatives and sulfur-containing amino acids, by producing clusters that showed strong agreement with the SMILES-based structural analysis. This convergence is a crucial testament to the autoencoders' ability to "see" chemical structure through their spectral fingerprints.

However, the most significant findings emerged from the **divergences** between the two methods. The Conv1D and Transformer models, in particular, proved to be highly effective at identifying subtle,

non-obvious relationships that a simple structural analysis could not. For instance, the Conv1D model created a "heavy-atom" cluster that grouped selenocystamine, cysteamine, and kynurenine based on their shared spectral features, despite their distinct chemical structures. Similarly, the Transformer model was able to connect molecules from different classes, like pterin and vitaminB12, within the Tryptophan family cluster.

Ultimately, the **Conv1D autoencoder** emerged as the clear winner. Its consistent, chemically coherent clustering and its ability to simultaneously validate traditional chemistry while revealing new, subtle relationships makes it the optimal choice for this application. It demonstrated higher resolution than the SMILES clustering by creating dedicated, highly specific clusters for sulfur-containing molecules. Furthermore, it correctly identified unique spectral outliers, such as lipoamide, which the Dense model failed to do.

In conclusion, our work shows that SERS-based clustering is not a redundant measure. It adds a new and invaluable layer of information by grouping molecules based on their true vibrational fingerprints, providing a powerful new lens for chemical analysis and discovery

References

Baumberg, J., Bell, S., Bonifacio, A., Chikkaraddy, R., Chisanga, M., Corsetti, S., Delfino, I., Eremina, O., Fasolato, C., Faulds, K., & Fleming, H. (2017). SERS in biology/biomedical SERS: general discussion. Faraday Discussions, 205, pp.429-456.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers) (pp. 4171-4186).

Hadipour, H., Liu, C., Davis, R., Cardona, S. T., & Hu, P. (2022). Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means. BMC bioinformatics, 23(Suppl 4), 132.

Lu, Y., Lin, L., & Ye, J. (2022). Human metabolite detection by surface-enhanced Raman spectroscopy. Materials Today Bio, 13, 100205.

Pilot, R., Signorini, R., Durante, C., Orian, L., Bhamidipati, M., & Fabris, L. (2019). A review on surface-enhanced Raman scattering. Biosensors, 9(2), 57.

Qiu, T.Y., Ding, Y., Huang, Y.Y., Zeng, M., Li, C., Zha, X.M., Zhou, W.B., Wang, N.N., Pian, C., Chen, F., & Cao, Y. (2025). Intraoperative Assessment of Parathyroidectomy Outcomes via Autoencoder–Support-Vector-Machine-Assisted Label-Free Differential SERS Spectroscopy. Nano Letters.

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. Journal of chemical information and modeling, 50(5), 742-754.

Sharma, B., Frontiera, R. R., Henry, A. I., Ringe, E., & Van Duyne, R. P. (2012). SERS: Materials, applications, and the future. Materials today, 15(1-2), 16-25.

Sherman, L. M., Petrov, A. P., Karger, L. F., Tetrick, M. G., Dovichi, N. J., & Camden, J. P. (2020). A surface-enhanced Raman spectroscopy database of 63 metabolites. Talanta, 210, 120645.

Sloan-Dennison, S., Wallace, G. Q., Hassanain, W. A., Laing, S., Faulds, K., & Graham, D. (2024). Advancing SERS as a quantitative technique: challenges, considerations, and correlative approaches to aid validation. Nano Convergence, 11(1), 33.

Sparavigna, A. C. (2024). Atlas of Metabolite SERS Fingerprints obtained by means of q-Gaussian deconvolutions and Fityk Software. ChemRxiv. doi:10.26434/chemrxiv-2024-85l19-v2

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). Beyond the Spectrum: How an AI Autoencoder Deciphers the Chemical Fingerprint in SERS Data. Zenodo. https://doi.org/10.5281/zenodo.16895315

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

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). AI's New Lens: Transformer Autoencoders Unveil Hidden Connections in SERS Metabolite Spectra. Zenodo. https://doi.org/10.5281/zenodo.17021372

Thakur, S., & Balotra (2024). Challenges and limitations of SERS. Surface-Enhanced Raman Spectroscopy for Water Quality Monitoring, 267.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Yousuff, M., & Babu, R. (2022). Deep autoencoder based hybrid dimensionality reduction approach for classification of SERS for melanoma cancer diagnostics. Journal of Intelligent & Fuzzy Systems, 43(6), 7647-7661.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. 1988, 28, 1, 31–36