Implementation Flow: CaPE → *CaPSimAlgorithm*

Note: All page citations refer to Halas et al., ACS Nano 2023, 17, 21251-21261.

Introduction

This document details the step-by-step process for implementing the **Characteristic Peak Similarity (CaPSim)** algorithm. [Page: 21252] This method relies on **Characteristic Peak Extraction (CaPE)** to identify an unknown chemical's SERS spectrum by matching it against a reference Raman library. [Page: 21251]

Experimental Dataset Overview

Goal: To understand the composition of the dataset used for evaluating the algorithm's performance.

- **Dataset Size:** The quantitative evaluation was performed using a reference Raman library consisting of **16 distinct chemicals**. [Page: 21255]
- **Number of Examples:** For each chemical, the dataset included multiple spectra to ensure a robust analysis. The breakdown is as follows:
  - **Reference Set (Raman Spectra):** For each of the 16 chemicals, **25 to 30 Raman spectra** were recorded to form the reference library. [Page: 21255]
  - **Query Set (SERS Spectra):** For each chemical being tested, **25 to 30 SERS spectra** were collected to serve as the "unknown" query examples. [Page: 21253, 21255]
- **Analysis Note:** Using multiple reference spectra allows the algorithm to average the similarity scores from all recordings, leading to a more stable and reliable estimation of the final match. [Page: 21255] This process is detailed in Step 4.

Step 1: Preprocess All Spectra (Query and Library)

Goal: To standardize all spectra by removing non-informative variations (like baseline drift and absolute intensity), ensuring a fair and robust comparison. [Page: 21253, 21257]

- **How it is done:**
  1. **Baseline Correction:** A baseline removal algorithm is applied to eliminate slow-changing trends from the raw spectra. [Page: 21253, 21257] The likely method is Asymmetric Least Squares (AsLS), as the paper cites reference [50]. [Page: 21257]
  2. **l_2-Normalization:** After baseline removal, each spectrum's intensity vector (x) is normalized by its Euclidean (l_2) norm. [Page: 21253, 21257] The formula is:

$$\tilde{x} = \frac{x}{\|x\|_2} = \frac{x}{\sqrt{\sum_i x_i^2}}$$ [Page: 21257]

Step 2: Apply CaPE to Reference Raman Spectra

Goal: To identify the most stable and representative "Characteristic Peak" (CP) locations for each chemical in the reference Raman library. [Page: 21254] These locations serve as the basis for comparison. [Page: 21258]

- **How it is done:**
  1. **Smoothing:** The CaPE algorithm first smooths the spectrum to reduce high-frequency noise before peak detection. [Page: 21258] This is controlled by the hyperparameter K_{smooth}. [Page: 21258]
  2. **Peak Finding:** Peaks are identified based on their **intensity values**. [Page: 21258]
  3. **Top Peak Selection:** The algorithm retains the top N_{peak} peaks with the highest intensity. [Page: 21258]
  4. **Define CP Regions:** A window of width w_{max} is created around each of the top peak locations. [Page: 21258]
- **Hyperparameters:** The paper specifies the values used in their experiments: [Page: 21258]
  - Smoothing kernel size (K_{smooth}): **5** [Page: 21258]
  - Number of top peaks to keep (N_{peak}): **10** [Page: 21258]
  - Maximum width of CP locations (w_{max}): **36** [Page: 21258]

Step 3: Extract CP Feature Vectors via Max-Pooling

Goal: To represent both the query and reference spectra as compact, fixed-length feature vectors based on the CP locations learned in Step 2. [Page: 21258]

- **How it is done:**
  1. **Max-Pooling:** For each of the N_{peak} (i.e., 10) CP regions, the maximum intensity value within that window is extracted. [Page: 21258] This creates a 10-dimensional compressed vector. [Page: 21258]
  2. **Min-Max Normalization:** The resulting 10-dimensional vector is normalized to a range of $[0, 1]$ for a consistent scale across different examples. [Page: 21258]

Step 4: Compute the CaPSim Similarity Score

Goal: To calculate a final, robust similarity score between the query spectrum's feature vector and each reference chemical. [Page: 21254, 21255]

- **How it is done:** To account for the multiple reference spectra (as outlined in the Dataset Overview), the final score is the **mean** of the individual similarity scores calculated against each reference example. [Page: 21258]

$$S_{CaPSim}(q, R_j) = \frac{1}{n_j} \sum_{i=1}^{n_j} \tilde{q}_j^T \tilde{r}_{i,j} \quad \text{[Page: 21258]}$$

Step 5: Identify the Query Spectrum

Goal: To predict the identity of the unknown chemical by finding the best match in the library. [Page: 21254]

- **How it is done:** The CaPSim scores calculated in Step 4 for all 16 reference chemicals are ranked, and the chemical that produces the **highest CaPSim score** is selected as the predicted identity. [Page: 21258]

Clarification: S_{CaPSim} vs. $Attr_{CaPSim}$

The paper uses two related terms to analyze the similarity score. [Page: 21255, 21258]

- S_{CaPSim} **(The Similarity Score):** This is the **final, single scalar value** representing the overall similarity, calculated by summing the contributions from all characteristic peaks. [Page: 21258]
- $Attr_{CaPSim}$ **(The Attribution Vector):** This is a vector showing how much each individual CP contributes to the final score *before* the summation. [Page: 21258] It is the element-wise (Hadamard) product of the two compressed feature vectors. [Page: 21258]

In short, S_{CaPSim} is the sum of the elements in the $Attr_{CaPSim}$ vector, and this vector is used for interpretability to visualize which specific peaks were most important for a given match. [Page: 21255, 21258]