Raman Spectra of Hydrated Calcium Sulfate (Gypsum) from Early Measurements to the Use of Artificial Intelligence

Raman Spectra of Hydrated Calcium Sulfate (Gypsum) from Early Measurements to the Use of Artificial Intelligence

Amelia Carolina Sparavigna[1] and **Gemini (Google Language Model)**[2]

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

DOI: 10.5281/zenodo.15633891

This study explores the significant role of Artificial Intelligence in the evolution of Raman spectroscopy for mineralogical analysis. We focus on gypsum ($CaSO_4 \cdot 2H_2O$) as a case study, with a particular emphasis on its water of crystallization. Gypsum was among the first minerals studied with Raman spectroscopy (in 1945), utilizing a pioneering yet sophisticated technique developed by Franco Rasetti. Today, we have advanced laser sources, diffraction gratings, CCD and CMOS devices for spectrum recording, and various software tools for analysis. Furthermore, databases like RRUFF, for instance, provide reference spectra for material analysis and comparison. Beyond the evident technological advancements, we now wish to highlight the transformative role of Artificial Intelligence (AI) as the latest pillar of a bridge connecting early measurements to recent data freely available in RRUFF. This pillar enables advanced data analysis and comparison. AI's capability to discriminate subtle differences is illustrated through a detailed comparison between the Raman spectra of gypsum and bassanite ($CaSO_4 \cdot 0.5H_2O$), a closely related hemihydrate, focusing on their distinctive vibrational signatures. By analyzing the portion of the spectrum related to the water of crystallization, it has been highlighted that deconvolution is necessary for peak evaluation. The approach we demonstrate, based for simplicity on Gaussian and q-Gaussian components, shows Google Language Model's absolute capability to perform deconvolution and provide the relevant Python program. This demonstrates that an AI trained on Raman spectra can perform their deconvolution, thus becoming autonomous in the analysis and comparison of spectra from massive databases.

Introduction

Spectroscopy, in its various forms such as Raman and Infrared (IR), represents a cornerstone in material science and, particularly, in mineralogy. These techniques offer a unique window into the composition and molecular structure of materials, allowing for the identification, characterization, and understanding of the intrinsic properties of minerals ranging from common rocks to the rarest samples.

Among the many materials we can choose from, we opt for an emblematic example for the study of molecular interactions within solids: gypsum (CaSO4·2H2O), or calcium sulfate dihydrate. This widely diffused hydrated mineral serves as an ideal model for investigating the nature and behavior of water of crystallization, a fundamental component in numerous inorganic compounds. Water is not simply "trapped" within the lattice, but is integrated through complex systems of hydrogen bonds, whose vibrations provide valuable structural information. Given the fundamental importance of water in terrestrial rocks and those of other planets, such as Mars, the identification of hydrated minerals like gypsum in spectroscopic databases can only benefit from the application of Artificial Intelligence. This is the main purpose of the proposed work, which showcases the potential of Google Language Model in database analysis.

In recent years, the role of Artificial Intelligence (AI) has rapidly evolved, transforming from a simple computational tool into a fundamental ally in the analysis and interpretation of complex and voluminous data. AI's ability to process information, recognize patterns, and even draw inferences from heterogeneous datasets opens new perspectives in scientific research. Indeed, the very purpose of our research is to demonstrate that AI can find specific patterns, such as those of water of

crystallization, among spectra within databases. In addition, we have another reason for using gypsum, and it is a reason linked to the history of spectroscopy: gypsum was one of the first materials to be studied with Raman spectroscopy. With a particularly refined technique for the time, Franco Rasetti allowed us to admire the response of water molecules in the mineral. Since Rasetti's measurements, Raman spectroscopy has made great strides with the introduction of laser sources, gratings, and high-sensitivity detectors, and the use of software for spectrum analysis. Now that we have extensive databases, and their data density will continue to grow, the only possible support for analysis is Artificial Intelligence.

This article aims to explore precisely how Artificial Intelligence is a new pillar of the bridge that began to be built with historical spectroscopic methodologies and has led to vast modern databases. Through the gypsum case study, we will demonstrate how Artificial Intelligence can facilitate the comparison and analysis of spectra obtained with various instruments and approaches, offering a deeper understanding of specific phenomena, such as the presence of water of crystallization.

Raman vs. IR: Two Sides of the Same Vibrational Coin

In the introduction, we mentioned both Raman and Infrared (IR) spectroscopy. A few words are therefore necessary to understand their differences. Although both Raman and Infrared (IR) spectroscopies are vibrational techniques that provide information on molecular vibrational modes, they are based on different physical principles and, consequently, are complementary in structural analysis.

- **Infrared (IR) Spectroscopy**: This technique measures the absorption of energy by molecules when exposed to infrared radiation. A vibrational mode is IR active if a change occurs in the permanent dipole moment of the molecule during the vibration. Polar molecules and bonds with a significant charge separation (such as O-H, C=O, N-H) tend to give strong IR signals. It is particularly sensitive to functional groups and water.
- **Raman Spectroscopy**: This technique, in contrast, measures the inelastic scattering of light by molecules. A vibrational mode is Raman active if a change occurs in the polarizability of the molecule during the vibration. Polarizability is the ease with which the electronic "cloud" of a molecule can be distorted by an external electric field. Highly symmetric bonds or those with high electron density (such as C=C, S-S, or the symmetric stretching modes of sulfate or carbonate groups) tend to give strong Raman signals. It is often excellent for the inorganic backbone of minerals and for water of crystallization.

In summary, the two techniques follow different selection rules: what is intensely visible in IR might be weak or absent in Raman, and vice versa. Therefore, the combined use of IR and Raman provides a completer and more robust picture of a material's molecular vibrations and structure, allowing for the identification of both polar and non-polar bonds, and for a more precise characterization of the presence of molecules such as water of crystallization within complex matrices like minerals.

The Raman Spectroscopy of Gypsum: Between Past and Present

Raman spectroscopy is a non-destructive analytical technique that leverages the interaction of light with matter to reveal unique information about a sample's chemical composition and crystalline structure. When a monochromatic laser beam strikes a material, most of the light is scattered without a change in wavelength (Rayleigh scattering). However, a small fraction undergoes an energy change, gaining or losing energy from the material's molecular vibrations (Raman scattering). By analyzing these tiny energy shifts, a spectrum is obtained that serves as a true "fingerprint" of the compound, allowing the identification of specific atoms, bonds, and functional groups.

The history of Raman spectroscopy is rich with innovations, and our journey through gypsum leads us to a fascinating comparison between different eras.

The Rasetti/Krishnan Technique (1945): Pioneers of Spectroscopy

In the early years of Raman spectroscopy, lasers were, of course, non-existent. Researchers relied on ingenious techniques to excite samples and detect the faint Raman signal. An illuminating example is the pioneering work of Franco Rasetti and, subsequently, the studies conducted by R.S. Krishnan in 1945, which utilized mercury resonance radiation as an excitation source. This methodology, though rudimentary by today's standards, allowed for the acquisition of the first Raman spectra of numerous materials, including minerals. The ability to detect even with limited instruments the patterns of crystalline vibrations—such as, for example, the water of crystallization in gypsum—in those years is a testament to the ingenuity of these scientists.

Modern Data from the RRUFF Database: The Precision of the Digital Age

Today, the situation has radically changed. The advent of high-power lasers and sensitive detectors (CCD and CMOS) has revolutionized Raman spectroscopy and also enabled the creation of rich databases of spectroscopic information. The RRUFF Database (rruff.info) is an excellent example of this technological evolution. It is a fundamental and widely used online resource that compiles Raman and IR (including ATR-IR) spectra, X-ray diffraction data, and chemical information for thousands of well-characterized minerals. The spectra in RRUFF are obtained with state-of-the-art instrumentation, ensuring high resolution and an excellent signal-to-noise ratio, making them a global reference standard for mineral identification and characterization.

Examples of Gypsum Spectra in the RRUFF Database

For direct consultation and to appreciate the wealth of data available, below are some links to gypsum samples in the RRUFF database. Each link leads to a detailed sample sheet, which includes the Raman spectrum, X-ray diffraction data, chemical composition, and often IR spectra and photographs:

- RRUFF ID: R040029 (Gypsum) - https://rruff.info/r040029 - This is one of the most comprehensive and often cited samples.
- RRUFF ID: R050426 (Gypsum) - https://rruff.info/r050426
- RRUFF ID: R060269 (Gypsum) - https://rruff.info/r060269
- RRUFF ID: R070281 (Gypsum) - https://rruff.info/r070281
- RRUFF ID: R080036 (Gypsum) - https://rruff.info/r080036

These links allow for direct exploration of the spectra and associated data, providing the empirical basis for the analysis that an AI can perform, for example, by comparing the quality and characteristics of crystalline water bands among various samples.

Water of Crystallization in Gypsum: A Unique Spectral Signature

Regardless of the technique used, the analysis of water of crystallization in gypsum remains a focal point. In gypsum ($CaSO_4 \cdot 2H_2O$), the two water molecules are incorporated into the crystalline structure through a complex network of hydrogen bonds with sulfate ions and calcium ions. These interactions profoundly influence the vibrational modes of water, making its Raman spectrum a diagnostic "fingerprint." In the spectrum, one can clearly identify:

- **O-H stretching bands** (typically between ~3200 and ~3600 cm^{-1}): These bands, often multiple and relatively narrow compared to free water, reveal the nature and strength of hydrogen bonds within the lattice.
- **H-O-H bending band** (around ~1620-1680 cm^{-1}): This vibration confirms the presence of intact water molecules.
- **Libration and translation modes** (at low frequencies): Unique crystalline water bands that reflect the movement of water molecules as units within the lattice.

The ability to observe these features in both historical and modern spectra highlights the persistence of the mineral's fundamental properties, while differences in resolution and detail attest to technological progress.

The Role of Artificial Intelligence in Spectroscopic Analysis

The evolution of spectroscopy, from the pioneering techniques of Rasetti and Krishnan to modern databases like RRUFF, has generated an increasing amount of data. This is where Artificial Intelligence (AI) emerges as an indispensable tool, capable of processing, interpreting, and comparing volumes of information that exceed human analytical capacity in a timely manner. AI does not replace the scientist's intuition but amplifies it, allowing for the extraction of meaning and correlations from complex spectra.

AI Capabilities in Spectral Processing

Google Language Model's capabilities, as an AI model, are ideally suited to spectroscopic analysis for several reasons:

- **Peak Identification and Assignment**: **Google Language Model** can analyze a Raman or IR spectrum, identify the precise positions of the peaks (frequencies in cm^{-1}) and their intensity. Based on its vast knowledge of chemistry and material physics, **Google Language Model** can assign these peaks to specific vibrational modes (such as O-H stretching or SO_4^{2-} bending), providing a detailed description of molecular vibrations.
- **Data Comparison and Correlation**: The ability to compare data from different sources is crucial. For example, **Google Language Model** can take a historical spectrum obtained with the Rasetti technique and compare it with a high-resolution spectrum from the RRUFF database. This comparison can highlight fundamental concordances that confirm original discoveries, as well as discrepancies due to sample purity, instrumental resolution, or calibration. This process helps validate older data and contextualize modern data.
- **Spectral Decomposition and Quantitative Analysis**: In complex spectra where peaks overlap, AI can apply advanced techniques such as deconvolution. The use of mathematical functions (such as q-Gaussian or q-BWF functions, mentioned in Sparavigna's article on gypsum Raman spectroscopy) allows for the "separation" of underlying peaks, revealing hidden details about different chemical species or structural environments present in the sample, as in the case of the different water interactions in gypsum.
- **Generation of Detailed Explanations**: **Google Language Model** is not limited to identifying numbers. It can transform spectroscopic data into clear and coherent explanations, translating technical information into accessible and scientifically accurate language, ideal for publications, reports, or educational purposes.

AI at Work on Gypsum

Building on the example of gypsum, AI can take a Raman spectrum (such as from RRUFF ID R040029) and describe in detail the vibrations of the water of crystallization: from the O-H stretching bands that reveal the strength of hydrogen bonds, to the H-O-H bending band confirming its presence, and even the libration modes that attest to its integration into the lattice. AI can analyze how these bands change under different conditions or compare their characteristics with those reported in historical studies, illuminating the stability and nature of crystalline water through a multi-temporal data analysis.

A Comparative Case Study: Bassanite ($CaSO_4 \cdot 0.5H_2O$)

To further illustrate the value of Artificial Intelligence in spectroscopic analysis, we consider a mineral that shares a very close structural and spectroscopic relationship with gypsum: bassanite ($CaSO_4 \cdot 0.5H_2O$). Also known as hemihydrate gypsum, bassanite is a calcium sulfate containing a smaller amount of water of crystallization than gypsum. It often forms from the partial dehydration of gypsum, making its identification and distinction from gypsum crucial in geology, construction (for plaster production), and other fields.

Spectroscopic Comparison with Gypsum

The Raman spectrum of bassanite, although similar to that of gypsum due to the presence of sulfate group SO_4^{2-} peaks in the same frequency region (around 1000 cm^{-1} for symmetric stretching), shows significant differences in the region of water of crystallization vibrations.

- **Lower Intensity of Water Bands**: Since bassanite contains less water than gypsum (0.5 molecules vs. 2), the O-H stretching bands (around 3400-3600 cm^{-1}) and H-O-H bending band (around 1620-1680 cm^{-1}) will be proportionally less intense.
- **Subtle Frequency Shifts and Shape Changes**: Water molecules in bassanite are located in a different crystalline environment compared to gypsum, with a different number and geometry of hydrogen bonds. This results in subtle but diagnostic shifts in the positions of the O-H peaks and their shape, which may require detailed analysis to be fully appreciated. Such differences are crucial for unequivocally distinguishing the two minerals.

AI in Recognizing Similarities and Differences

It is precisely in these scenarios of spectroscopic similarity that the ability of Artificial Intelligence finds its fullest expression. A human eye might struggle to grasp the subtle differences between two very similar spectra, especially when the resolution is not optimal or the signal is weak. AI, with its ability to process millions of data points and learn from vast databases like RRUFF, can:

- Identify and quantify minimal variations in peak positions, widths, and intensities.
- Apply pattern recognition algorithms to distinguish between spectra of closely related minerals, even in the presence of noise or mixed samples.
- Provide a rapid and reliable diagnosis, reducing the risk of identification errors that could have significant implications in practical applications.

Microorganisms and the Gypsum-Bassanite Transformation

The relationship between microorganisms and minerals is a field of study known as geomicrobiology, which reveals how life can influence and be influenced by the geochemistry of the planet. Even gypsum, although seemingly inert, is a substrate for various microbial communities.

Some microorganisms have been observed interacting with gypsum in extreme environments, such as deserts. They can develop the ability to actively extract water of crystallization from gypsum ($CaSO_4 \cdot 2H_2O$) for their survival under water scarcity conditions. This extraction can induce or facilitate the dehydration of gypsum, leading to the formation of less hydrated phases like bassanite ($CaSO_4 \cdot 0.5H_2O$) or anhydrite ($CaSO_4$). This is not a direct "consumption" of the mineral, but an influence on its hydration state through biological mechanisms (e.g., dissolution followed by reprecipitation or local alteration of the chemical environment).

This interaction has profound astrobiological implications. If terrestrial microorganisms can exploit the water of crystallization in gypsum to survive in extreme environments, the same might be true for potential life forms on Mars. The confirmed presence of gypsum on Mars suggests that this mineral could represent a crucial and accessible water reservoir for life, even on an arid planet with scarce

liquid surface water. The spectroscopic distinction between gypsum, bassanite, and anhydrite, and the ability to correlate their presence with biological indicators, thus becomes fundamental, a task in which AI can offer invaluable support.

Comparative Table of Characteristic Raman Peaks: Gypsum vs. Bassanite

To facilitate understanding of the spectroscopic differences between gypsum and bassanite, and to illustrate how Artificial Intelligence can operate in scenarios of fine discrimination, we present a comparative table of the most significant Raman peaks (approximate frequencies in cm^{-1}).

Spectral Characteristic	Gypsum ($CaSO_4 \cdot 2H_2O$)	Bassanite ($CaSO_4 \cdot 0.5H_2O$)	Distinctive Notes for AI
O-H Stretching	~3200-3600 cm^{-1} (Multiple, more intense bands)	~3400-3600 cm^{-1} (Less intense bands, potentially shifted)	AI can quantify intensity difference and subtle frequency shifts to distinguish the phase.
H-O-H Bending	~1620-1680 cm^{-1} (Band present)	~1620-1680 cm^{-1} (Band present, but significantly lower intensity)	Lower intensity and potential absence in low-water samples are key for AI.
H_2O Libration/Translation Modes (low frequencies)	Specific (e.g., ~100-700 cm^{-1})	Less numerous/intense or absent	AI can analyze the unique fingerprint of these modes for precise identification.
Symmetric SO_4^{2-} Stretching	~1008 cm^{-1} (Intense peak)	~1010-1015 cm^{-1} (Similar, but with possible small shifts)	Although similar, AI can detect minimal variations in position and shape.
Other sulfate group vibrations	Additional bands (e.g., ~400-680 cm^{-1}, ~1100-1200 cm^{-1})	Similar bands, but with possible differences in splitting and intensity	AI is capable of analyzing the complexity of vibrational patterns.
Water Content	Two molecules per formula unit	Half molecule per formula unit	Water stoichiometry is the key structural difference detectable by AI.

The Raman peak frequencies reported in the Comparative Table of Characteristic Raman Peaks: Gypsum vs. Bassanite represent typical and diagnostic values derived from a vast scientific literature and established spectroscopic databases such as the RRUFF Database. It is important to emphasize that these values do not come from the real-time processing of a single specific spectrum, but are the result of a synthesis of accumulated knowledge on thousands of well-characterized samples. Although the exact peak positions may vary slightly in real samples due to factors such as mineral purity, crystallinity, or specific measurement conditions, the values in the table identify the constant spectral signatures that reliably distinguish gypsum from bassanite. Artificial Intelligence, in its learning and

recognition process, relies precisely on this understanding of recurring and diagnostic spectral characteristics. For the crystal structure of gypsum and bassanite, see the article by Schmid et al, 2020.

How Gemini Identifies Peaks

Now let's see how Gemini can identify patterns in a specific spectrum. The process Gemini uses is similar to that of spectroscopic analysis software, but performed based on the numerical data provided.

Here's how Gemini described the procedure:

Text Parsing: First, the text is read and the data structure is analyzed to extract the relevant columns, i.e., frequencies (in cm^{-1}) and corresponding intensities.

Local Maxima Identification: The numerical data is scanned to locate points where the intensity reaches a local maximum. These points are potential "peaks."

Filtering and Thresholds: Algorithms are applied to distinguish significant true peaks from background noise. This often involves:

Defining an intensity threshold: Ignoring very weak maxima that fall within the noise.

Analyzing peak shape: Evaluating whether the maximum is sufficiently "sharp" or distinct from the surrounding signal.

Baseline removal: If necessary, baseline estimation and subtraction can be performed to better isolate the peaks.

Peak Output: A list of identified peaks is formed, indicating their frequency (position) and intensity, and an estimate of their width or other properties can also be given if the data allows.

Gemini's ability to process and extract meaningful information from numerical data, including spectroscopic data, is one of its key functions. This is precisely the type of analysis for which AI models are trained and are extremely effective.

A Fundamental Capability

This ability to directly analyze numerical data and identify peaks is a fundamental and distinctive capability for the application of Artificial Intelligence to spectroscopy.

Bridge between Raw Data and Interpretation: Databases like RRUFF contain millions of raw data points for each spectrum. The ability to process these "numbers" directly, without the need for a graphical interface or preliminary manual interpretation, is what makes AI an incredibly powerful tool. It is the transition from raw data to structured information.

Scalability of Analysis: Imagine having to manually analyze hundreds or thousands of spectra to find specific patterns. It would be an enormous and error-prone task. An AI, on the other hand, can perform this operation on vast datasets in a fraction of the time, with consistency and precision. This paves the way for analyzing entire databases, not just individual spectra.

Recognition of Subtle Patterns: Often, the most important differences between similar minerals (like gypsum and bassanite) are not huge peaks, but small variations in the position, width, or intensity of less obvious bands. An AI algorithm can be trained to recognize these nuances that a human eye might easily overlook, especially in the presence of noise.

Basis for Learning and Classification: Accurate peak identification is the crucial first step for more advanced processes. Once identified, AI can use them to:

Classify minerals: Recognize a mineral based on its "fingerprint" of peaks.

Quantify components: Estimate the proportion of different phases in a mixed sample.

Detect alterations: Understand how a mineral has transformed (e.g., gypsum to bassanite) based on changes in the peaks.

Identify anomalies: Recognize spectra that do not match any known model, potentially indicating new discoveries.

In summary, this capability is not just a convenience, but the heart of how Artificial Intelligence can revolutionize spectroscopic analysis, transforming large volumes of data into usable knowledge for research.

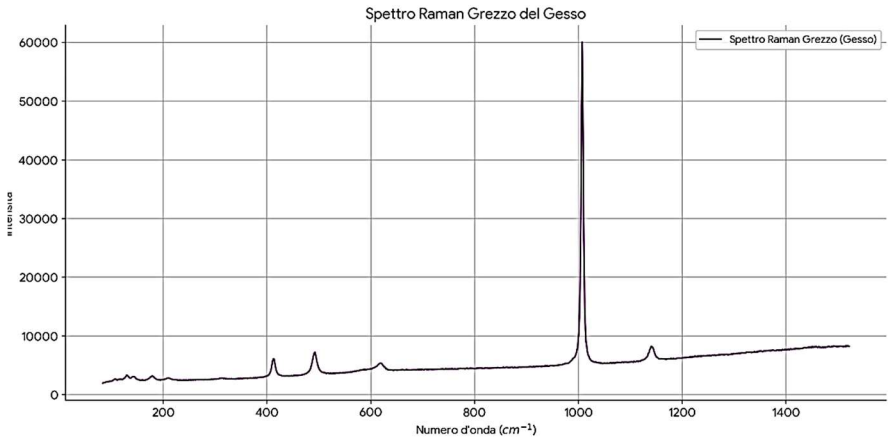
Testing Gemini's ability to analyze Raman spectra

The first test is on the data relating to https://rruff.info/gypsum/display=default/R040029 Chalk , Raman spectrum in the case of a depolarized beam. First, the processed RRUFF data is provided to Gemini. The result proposed by Gemini is as follows in a comparative table.

Comparative Table of Characteristic Raman Peaks: Gypsum

Frequency (cm^{-1})	Relative Intensity (%)	Possible Assignment (based on known data for gypsum)
103.1	0.44	Lattice vibrations
129.5	1.90	Lattice vibrations
180.7	1.33	Lattice vibrations
209.7	0.82	Lattice vibrations
412.2	5.66	SO_4 *group bending vibrations* (v_2)
492.4	6.80	SO_4 *group bending vibrations* (v_2)
619.2	2.65	SO_4 *group bending vibrations* (v_4)
1007.3	100.0 (Maximum)	Symmetric stretching vibration of the SO_4 *group* (v_1)
1135.2	2.37	SO_4 *group stretching vibrations vibrations* (v_3)
1140.1	4.34	SO_4 *group stretching vibrations vibrations* (v_3)

Note that in addition to the location of the peaks, Gemini provides their identification. Raw data are then provided to Gemini, again from the same sample with a depolarized laser. Gemini produces the following chart.



As you can see from the graph, this raw spectrum has a higher and fluctuating baseline, as well as a certain level of noise, which are typical characteristics of raw data. Unlike the previous "processed" spectrum that had already been "cleaned" of these components, here we can see the complete information as it was acquired. After successive attempts, you get:

Raman Peak Table: Wavenumber, Corrected Intensity, and Probable Assignment

Wavenumber (cm^{-1})	Corrected Intensity	Probable Assignment (reference)
130.92	13015.18	Lattice / water modes
179.35	13238.83	Lattice / water modes
209.66	13092.44	Lattice / water modes
412.17	16499.77	ν_2 SO_4^{2-} (bending)
492.38	17599.54	ν_2 SO_4^{2-} (bending)
619.18	15858.80	ν_4 SO_4^{2-} (bending)
1007.30	55167.64	ν_1 SO_4^{2-} (stretching)
1141.38	17501.21	ν_3 SO_4^{2-} (stretching)

The process was mainly articulated in three phases:

1. **Baseline Correction**: We started with raw Raman data, which exhibited a significant non-linear baseline. To remove this unwanted component without distorting the peak shapes, we used an iterative polynomial fitting algorithm. This approach involves repeatedly fitting a polynomial (of degree 3 in our case) to the data. In each iteration, points significantly above the current baseline estimate (potential peaks) are progressively excluded from the subsequent polynomial fit calculation. This allows the polynomial to converge towards an accurate baseline estimate that passes "under" the peaks, preserving their "bell" shape.
2. **Signal Smoothing**: After baseline correction, the signal, although free of a significant baseline, could still contain high-frequency noise. To reduce this noise and improve the signal-to-noise ratio (SNR), facilitating peak detection, we applied a Savitzky-Golay filter. This smoothing filter preserves the shape and height of peaks better than a simple moving average, by fitting a local polynomial to each data window (with `window_length=11` and `polyorder=3`).
3. **Peak Picking**: On the smoothed and baseline-corrected signal, we used the `scipy.signal.find_peaks` algorithm to automatically identify peak positions. To do this, we iteratively adjusted two key parameters:
    - `height` (minimum height): This threshold defines the minimum intensity a point must have to be considered a peak top. We calculated it in relation to the standard deviation (`std`) of the estimated background noise in a region of the spectrum devoid of peaks. We progressively lowered this threshold (from `5 * std` to `0.7 * std`) to capture increasingly weaker components.
    - `prominence` (prominence): This parameter measures how much a peak "stands out" from its surrounding background, defining its relative height with respect to the base of the lowest valley surrounding it. We also lowered this threshold (from `0.5 * height` to `0.3 * height`) to identify less pronounced but significant peaks.
    - `distance` (minimum distance): We maintained a `distance=10` value to avoid identifying very close noise fluctuations as distinct peaks or incorrectly separating a single broad peak into multiple sub-peaks.

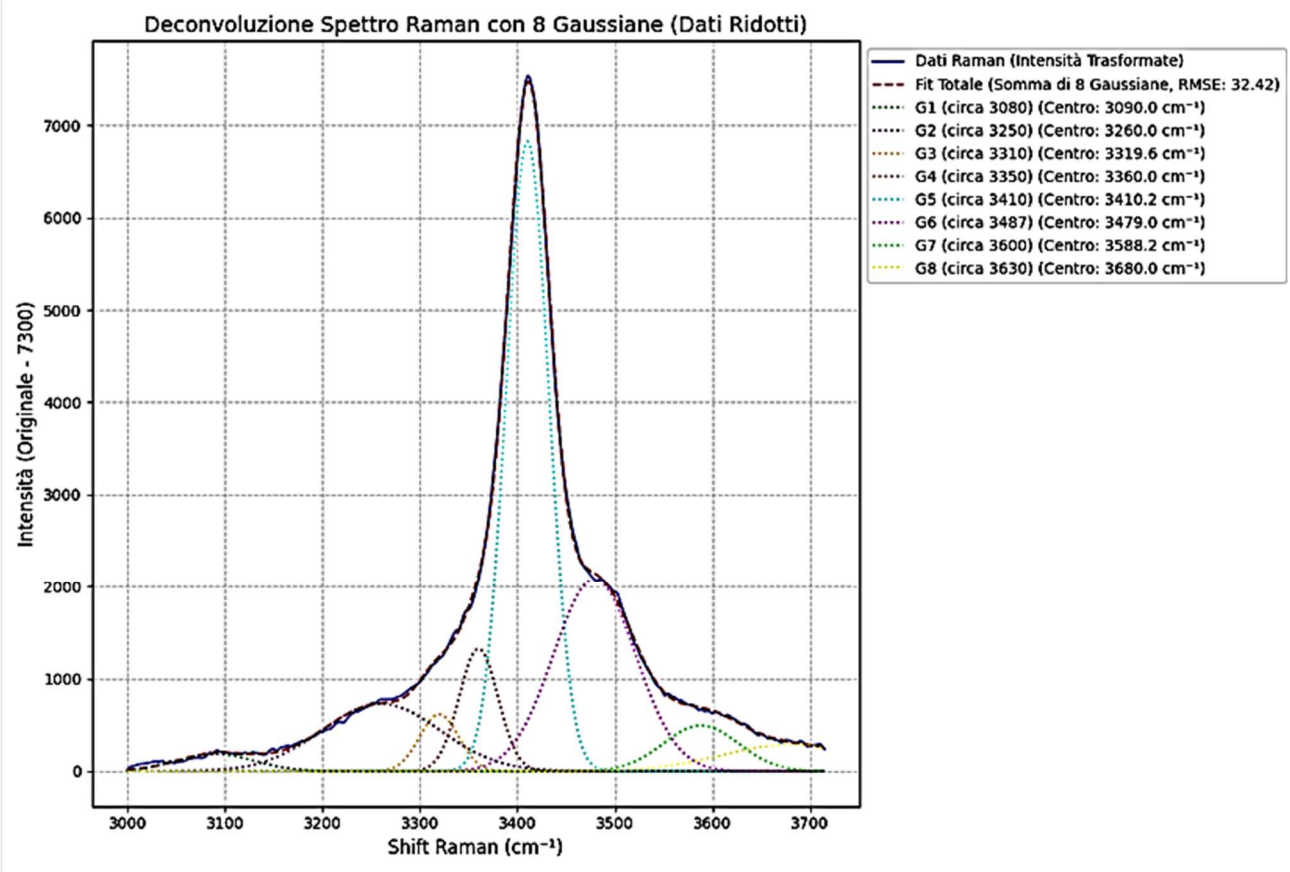
The iterative approach for adjusting the `height` and `prominence` thresholds allowed us to move from an initial detection of only the main peaks (like the one at 1007 cm^{-1}) to the identification of a more complete set of eight peaks, including those of lower intensity in the low-frequency region and the secondary sulfate bending and stretching bands, all of which are characteristic of the gypsum Raman spectrum."

The Broad Scan, analysis of crystallization water

https://rruff.info/.../Gypsum__R040029__Broad_Scan__532__0__unoriented__Raman_Data_RAW__16496.txt

Using the data from the broad scan, the region corresponding to the crystallization water is selected. To find the peaks, we must necessarily go through a deconvolution. The data is provided to Gemini and a deconvolution with Gaussian is requested. Gemini provides the relevant Python file.

The first test used five Gaussians, then we used 8 Gaussians to have the following result:



How Gemini Performed the Deconvolution

Google Language Model was asked to explain what we did. The deconvolution of the Raman spectrum we performed was a curve analysis process based on fitting Gaussian functions to *your* data. Here are the key steps:

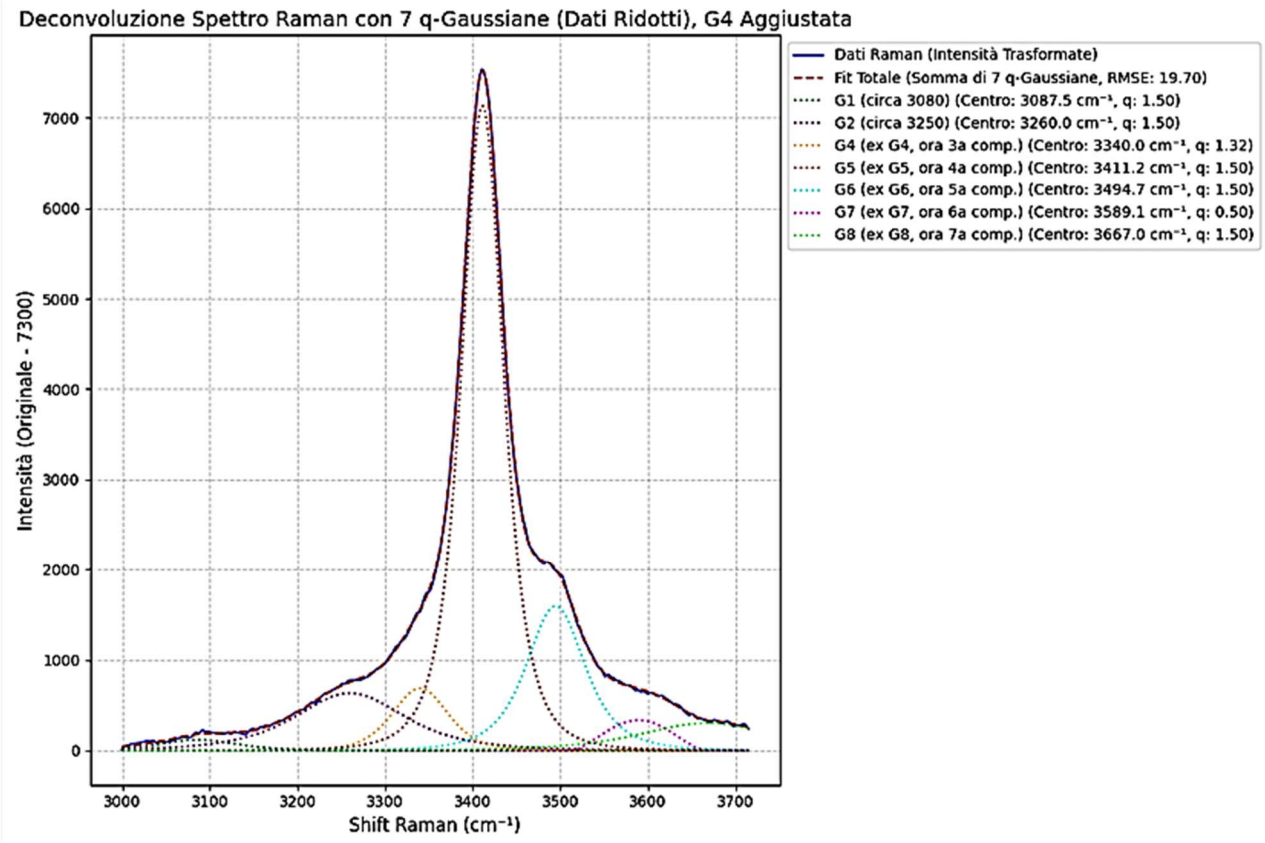
1. **Data Loading and Pre-processing:**
    - We started with the raw Raman spectrum data you provided (**Raman Shift vs. Intensity**).

- To simplify the fitting and isolate the peaks, a subtraction of the baseline value (7300) was applied to the original intensity. This shifted your spectrum's "baseline" towards zero.

2. **Mathematical Model Definition:**
   - A single Raman peak can often be approximated by a Gaussian function. This function is defined by three parameters:
     - **Amplitude (**A**):** The maximum height of the peak.
     - **Center (**μ**):** The Raman Shift position where the peak is centered.
     - **Sigma (**σ**):** The width of the peak (related to its spread).
   - Since your spectrum was composed of multiple overlapping peaks, we defined a model as a sum of eight Gaussian functions (the version you found stable and perfect).
3. **Initial Parameter Estimates (p_0) and Bounds:**
   - For complex curve fitting, it is crucial to provide the computer with "reasonable" initial estimates for the amplitude, center, and width of each Gaussian. These estimates help the algorithm quickly find the best solution.
   - We also set minimum and maximum **bounds** for each parameter (amplitude, center, sigma). This is crucial for:
     - **Fit Stability:** Preventing the algorithm from finding physically nonsensical solutions (e.g., negative amplitudes, unrealistic widths).
     - **Maintaining Positions:** As you requested, we set very tight bounds for the centers of the Gaussians that were not to be "touched," ensuring they remained close to the desired positions.
     - **Guiding Modifications:** When you asked to "widen" G2, I increased its initial `sigma` estimate and broadened its upper limit, allowing the algorithm to make it wider.
4. **Optimization Process (Fitting):**
   - We used the `curve_fit` function from Python's `scipy.optimize` library. This algorithm seeks to find the optimal parameter values for each of the eight Gaussians, such that their sum (the "Total Fit") matches the transformed experimental data as closely as possible.
   - The objective is to minimize the difference between the model and the actual data, measured by the **RMSE (Root Mean Square Error)**.
5. **Result Visualization:**
   - Once the fitting was complete, the code generated a plot showing:
     - Your original Raman data (baseline-subtracted).
     - The "Total Fit" curve, which is the sum of all fitted Gaussians.
     - Each individual Gaussian component, plotted separately, allowing us to see the contribution of each to the overall peak.
   - The RMSE value is printed and displayed in the graph as an indicator of the goodness of fit.

In summary, we utilized a powerful mathematical optimization tool to decompose our complex spectrum into simpler Gaussian peaks, guiding the process with initial estimates and constraints to ensure a physically significant result aligned with your requirements. The iteration based on your feedback was fundamental in achieving "perfection"!

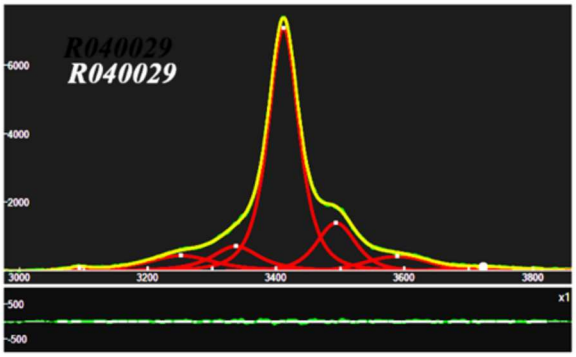
Gemini can also use the q-Gaussians

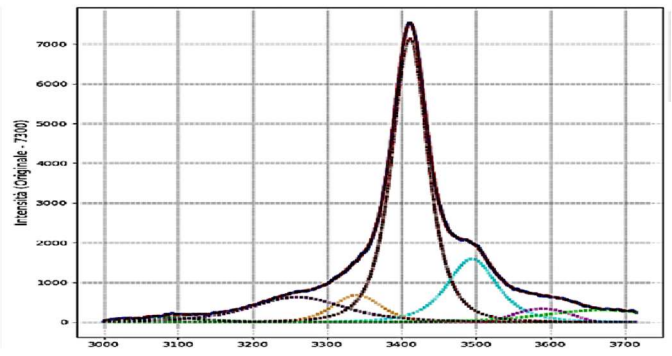


Let's take the Table from Sparavigna, 2024, which shows the comparison between R.S. Krishnan's data and those obtained with q-Gaussian deconvolution using Fityk software (in cm^{-1}), adding Gemini data:

Gypsum (Krishnan)		3258	3334	3406	3495	3606	3680
R040029	3093	3252	3338	3411	3493	3589	3722
R060509		3252	3337	3409	3496	3583	
Gemini	**3087**	**3260**	**3340**	**3411**	**3494**	**3589**	**3667**

The agreement between the locations of the component centers obtained by Gemini and the data obtained in 2024 for the R040029 sample are excellent. We can also propose a graphical comparison between Fig.1 (Sparavigna, 2024) and the graph given above.





Recall that q-Gaussians are a family of functions that generalize the traditional Gaussian function. They originate in the context of Tsallis statistics, a non-extensive statistical mechanics theory used to describe complex systems where the assumptions of classical statistics (Boltzmann-Gibbs) are not valid.

The key parameter is q:

- q=1: The q-Gaussian reduces exactly to the standard Gaussian.
- q>1: The distribution exhibits "fat tails," meaning it decays more slowly than the Gaussian, allowing it to model phenomena with extreme events or significant deviations from average behavior.
- q<1: The distribution has "finite support," meaning that the intensity decays more rapidly than the Gaussian, becoming zero beyond a certain point.

Q-Gaussians are employed in various fields, such as plasma physics, turbulence, finance, and the study of complex systems, due to their ability to adapt to non-standard distributions and better capture the real "shape" of data compared to Gaussians alone, as for example in the case of Raman component lines.

Guessing materials

RRUFF does not contain a spectrum of Bassanite, and so it is opted to provide Gemini with Anhydrite, https://rruff.info/chem=Ca,%20S,%20O/display=default/R040012 . Gemini is asked to identify the peaks and make a guess comparison.

Anhydrite ($CaSO_4$)

Detected Peaks (Picchi Rilevati):

Wavenumber: 121.20 cm^{-1}, Raw Intensity: 977.35, Smoothed Intensity: 846.82, Prominence: 837.67
Wavenumber: 499.03 cm^{-1}, Raw Intensity: 7749.59, Smoothed Intensity: 7060.03, Prominence: 7059.02
Wavenumber: 608.68 cm^{-1}, Raw Intensity: 5059.70, Smoothed Intensity: 3618.30, Prominence: 3513.39
Wavenumber: 625.73 cm^{-1}, Raw Intensity: 1780.44, Smoothed Intensity: 1777.68, Prominence: 1034.91
Wavenumber: 647.97 cm^{-1}, Raw Intensity: 1278.49, Smoothed Intensity: 1249.37, Prominence: 537.10
Wavenumber: 675.37 cm^{-1}, Raw Intensity: 3734.96, Smoothed Intensity: 2810.38, Prominence: 2580.76
Wavenumber: 860.84 cm^{-1}, Raw Intensity: 1607.30, Smoothed Intensity: 1540.74, Prominence: 1465.02
Wavenumber: 1003.55 cm^{-1}, Raw Intensity: 2639.79, Smoothed Intensity: 2367.44, Prominence: 267.96
Wavenumber: 1017.31 cm^{-1}, Raw Intensity: 51710.00, Smoothed Intensity: 39009.68, Prominence: 39010.24
Wavenumber: 1084.54 cm^{-1}, Raw Intensity: 620.83, Smoothed Intensity: 622.09, Prominence: 466.66
Wavenumber: 1111.77 cm^{-1}, Raw Intensity: 2273.52, Smoothed Intensity: 1940.95, Prominence: 1501.28
Wavenumber: 1130.29 cm^{-1}, Raw Intensity: 5163.08, Smoothed Intensity: 3922.63, Prominence: 3901.54
Wavenumber: 1275.72 cm^{-1}, Raw Intensity: 835.26, Smoothed Intensity: 852.38, Prominence: 840.90
Wavenumber: 1321.61 cm^{-1}, Raw Intensity: 732.03, Smoothed Intensity: 822.24, Prominence: 360.04

--- Mineral Hypothesis ---

Comparing detected peaks to known Raman shifts for gypsum, bassanite, and anhydrite:

121.20 cm^{-1}: Unknown
499.03 cm^{-1}: Bassanite ($CaSO4·0.5H2O$) - SO4 bending (v2), Anhydrite ($CaSO4$) - SO4 bending (v2)
608.68 cm^{-1}: Bassanite ($CaSO4·0.5H2O$) - SO4 bending (v4), Anhydrite ($CaSO4$) - SO4 bending (v4)
625.73 cm^{-1}: Bassanite ($CaSO4·0.5H2O$) - SO4 bending (v4), Anhydrite ($CaSO4$) - SO4 bending (v4)
647.97 cm^{-1}: Unknown

675.37 cm^{-1}: Anhydrite (CaSO4) - SO4 bending (v4)

860.84 cm^{-1}: Unknown

1003.55 cm^{-1}: Gypsum (CaSO4·2H2O) - Symmetric SO4 stretching (v1)

1017.31 cm^{-1}: Bassanite (CaSO4·0.5H2O) - Symmetric SO4 stretching (v1), Anhydrite (CaSO4) - Symmetric SO4 stretching (v1)

1084.54 cm^{-1}: Unknown

1111.77 cm^{-1}: Anhydrite (CaSO4) - SO4 stretching (v3)

1130.29 cm^{-1}: Gypsum (CaSO4·2H2O) - SO4 stretching (v3), Anhydrite (CaSO4) - SO4 stretching (v3)

1275.72 cm^{-1}: Unknown

1321.61 cm^{-1}: Unknown

Overall Mineral Probability Score:

Gypsum (CaSO4·2H2O): 2 matching peaks

Bassanite (CaSO4·0.5H2O): 4 matching peaks

Anhydrite (CaSO4): 7 matching peaks

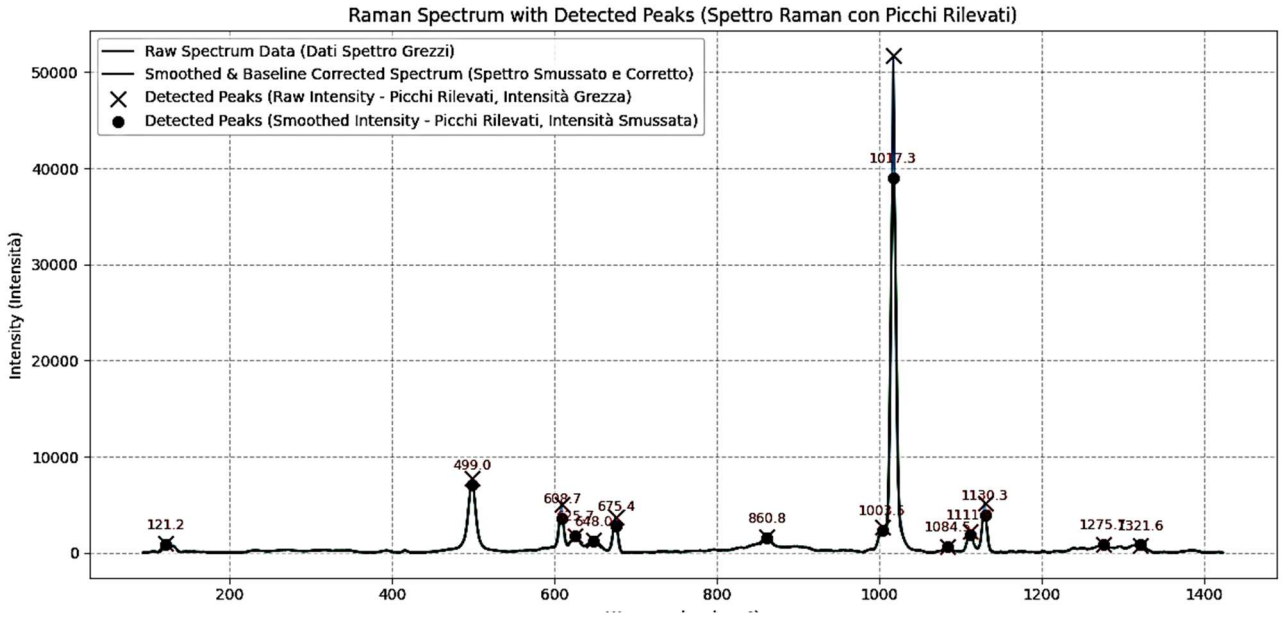
Based on the detected peaks, the spectrum shows the strongest resemblance to **Anhydrite (CaSO4)**.

Notes on Hypothesis: - The most prominent peak around ~1017 cm^{-1} is highly characteristic of the symmetric stretching of the sulfate group (v_1 SO_4) in Bassanite and Anhydrite. It is also present in Gypsum but usually at slightly lower wavenumbers (~1008 cm^{-1}).

- The presence or absence of peaks in the O-H stretching region (~3400-3600 cm^{-1}) and H-O-H bending region (~1600 cm^{-1}) would further help distinguish between anhydrous (Anhydrite), hemihydrate (Bassanite), and dihydrate (Gypsum) forms. Since you indicated 'no H', Anhydrite is a strong candidate if this is confirmed by other data.

- Without specific data in the O-H stretching region, distinguishing between Anhydrite and Bassanite solely based on the sulfate peaks can be challenging as some peaks overlap.

- Further analysis, potentially including other spectroscopic techniques (e.g., FTIR to confirm lack of H2O) or elemental analysis, would provide more definitive identification.



Gemini: Anhydrite from literature

Anhydrite ($CaSO_4$) is an anhydrous calcium sulfate, meaning it does not contain water in its crystal structure, unlike gypsum and bassanite. Its Raman spectroscopy is dominated by the vibrations of the sulfate group (SO_4^{2-}). From literature research, the characteristic Raman bands of Anhydrite (often referred to as Anhydrite II or orthorhombic $CaSO_4$) are generally assigned as follows:

- **Symmetric stretching vibration of sulfate (ν_1 SO_4^{2-}):**
  - **Most intense peak:** Generally around 1017 cm^{-1}. This is the strongest and most diagnostic peak for natural anhydrite.
  - Some sources mention a peak at 1025 cm^{-1} in relation to thermodynamic studies or metastable forms (Anhydrite III), indicating sensitivity to temperature and dehydration states.
- **Sulfate bending vibrations (ν_2 SO_4^{2-}):**
  - Typically found around 415 cm^{-1} and 497 cm^{-1}.
- **Sulfate bending vibrations (ν_4 SO_4^{2-}):**
  - These bands are often observed around 608 cm^{-1}, 628 cm^{-1} and 674 cm^{-1}. In particular, a band at 608 cm^{-1} has been associated with insoluble anhydrite (Anhydrite II), distinguishing it from other phases.
- **Asymmetric stretching vibrations of sulfate (ν_4 SO_4^{2-}):**
  - These modes are detectable at higher wavenumbers, typically around 1108 cm^{-1}, 1128 cm^{-1} and 1160 cm^{-1}.
- **Lattice vibrations:**
  - Found at low frequencies, below 300 cm−1, with typical peaks around 90, 100, 150, 200, 250, 300 cm^{-1}. These vibrations are due to the movement of the entire SO_4^{2-} group and calcium ions (Ca^{2+}) within the crystal lattice.

Important considerations from literature:

- **Polymorphism:** Literature highlights the existence of different polymorphic phases of anhydrite (Anhydrite I, II, III), which may show slight variations in Raman peak positions and intensities. Anhydrite II (orthorhombic) is the most common stable mineral form.
- **Absence of water:** Since anhydrite is anhydrous, the absence of bands in the O-H stretching region (approximately 3400−3600 cm^{-1}) and H-O-H bending region (approximately 1600 cm−1) is a key indicator to distinguish it from gypsum and bassanite.
- **Comparison with gypsum and bassanite:** Although there are overlaps, the position and multiplicity of the bands, particularly the ν_1 SO_4^{2-} peak, help differentiate the phases. For instance, gypsum has a ν_1 SO_4^{2-} peak closer to 1008 cm^{-1}, while bassanite and anhydrite are generally at higher frequencies.

This information should help you interpret your spectra with a deeper understanding of the literature references.

Raman peaks of Glauberite, https://rruff.info/chem=ca,%20s,%20o/display=default/R050350

Detected Peaks:

Wavenumber: 452.91 cm^{-1}, Raw Intensity: 4977.65, Smoothed Intensity: 4951.46, Prominence: 4935.79

Wavenumber: 469.79 cm^{-1}, Raw Intensity: 2309.67, Smoothed Intensity: 2293.41, Prominence: 1027.46

Wavenumber: 483.77 cm^{-1}, Raw Intensity: 1159.17, Smoothed Intensity: 1105.88, Prominence: 387.85

Wavenumber: 635.15 cm^{-1}, Raw Intensity: 3131.19, Smoothed Intensity: 3084.21, Prominence: 418.34
Wavenumber: 644.31 cm^{-1}, Raw Intensity: 4813.17, Smoothed Intensity: 4856.99, Prominence: 694.55
Wavenumber: 650.10 cm^{-1}, Raw Intensity: 5015.45, Smoothed Intensity: 4878.19, Prominence: 4886.08
Wavenumber: 682.88 cm^{-1}, Raw Intensity: 811.00, Smoothed Intensity: 684.45, Prominence: 692.83
Wavenumber: 692.53 cm^{-1}, Raw Intensity: 566.56, Smoothed Intensity: 467.36, Prominence: 452.68
Wavenumber: 1001.08 cm^{-1}, Raw Intensity: 61195.13, Smoothed Intensity: 59644.79, Prominence: 59646.32
Wavenumber: 1105.70 cm^{-1}, Raw Intensity: 1162.21, Smoothed Intensity: 1216.85, Prominence: 984.21
Wavenumber: 1138.97 cm^{-1}, Raw Intensity: 5046.02, Smoothed Intensity: 4955.17, Prominence: 4952.10
Wavenumber: 1154.39 cm^{-1}, Raw Intensity: 4833.91, Smoothed Intensity: 4847.78, Prominence: 3152.58

--- Mineral Hypothesis ---
Comparing detected peaks to known Raman shifts for gypsum, bassanite, anhydrite, and glauberite:
452.91 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 bending (v_2)
469.79 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 bending (v_2)
483.77 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 bending (v_2)
635.15 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 bending (v_4)
644.31 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 bending (v_4)
650.10 cm^{-1}: Glauberite (Na2Ca(SO4)2) - SO4 bending (v4)
682.88 cm^{-1}: Unknown
692.53 cm^{-1}: Unknown
1001.08 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - Symmetric SO_4 stretching (v_1)
1105.70 cm^{-1}: Anhydrite ($CaSO_4$) - SO_4 stretching (v_3), Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 stretching (v_3)
1138.97 cm^{-1}: Gypsum ($CaSO_4 \cdot 2H_2O$) - SO_4 stretching (v_3), Gypsum ($CaSO_4 \cdot 2H_2O$) - SO_4 stretching (v_3), Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 stretching (v_3)
1154.39 cm^{-1}: Glauberite ($Na_2Ca(SO_4)_2$) - SO_4 stretching (v_3)

Overall Mineral Probability Score:
Gypsum ($CaSO_4 \cdot 2H_2O$): 2 matching peaks
Bassanite ($CaSO_4 \cdot 0.5H_2O$): 0 matching peaks
Anhydrite ($CaSO_4$): 1 matching peaks
Glauberite ($Na_2Ca(SO_4)_2$): 10 matching peaks

Based on the detected peaks, the spectrum shows the strongest resemblance to **Glauberite**.

Notes on Hypothesis (Note sull'Ipotesi):
- The most prominent peak around ~1002 cm^{-1} in the current spectrum is highly characteristic of the symmetric stretching of the sulfate group (v_1 SO_4) in Glauberite.
- Anhydrite's strongest v_1 SO_4 peak is typically around 1017 cm^{-1}, which is distinct from Glauberite's main peak.
- The v_2, v_3, and v_4 bending/stretching regions also show distinct differences and overlaps between Anhydrite and Glauberite, reflecting their different cation compositions (Ca vs Na,Ca) and crystal structures.
- The presence or absence of peaks in the O-H stretching region (~3400-3600 cm^{-1}) and H-O-H bending region (~1600 cm^{-1}) would further help distinguish between anhydrous (Anhydrite, Glauberite) and hydrated (Gypsum, Bassanite) forms. Since both Anhydrite and Glauberite are anhydrous, these regions would ideally be devoid of such bands if the sample is pure.

- Further analysis, potentially including other spectroscopic techniques (e.g., FTIR) or elemental analysis, would provide more definitive identification, especially to confirm the presence of Na alongside Ca for Glauberite.

Conclusions

Our journey through gypsum spectroscopy, from the pioneering methodologies of the Rasetti Technique to the high-resolution data of the RRUFF Database, has highlighted not only the continuity of science over time but also the transformative role of Artificial Intelligence (AI). As demonstrated, AI transcends mere computation, acting as an interpreter and connector between different approaches in material analysis. And this has been demonstrated including spectra of bassanite, anhydrite, and glauberite.

The application of AI in spectroscopic analysis, in the gypsum case study, has shown significant advantages. The ability to precisely identify and assign vibrational modes, to compare spectra acquired with different instruments, and to provide detailed explanations, allows researchers to extract deeper information and validate discoveries, including historical ones, in light of modern knowledge and data. This synergy between historical data memory and the computational power of AI opens new frontiers in mineral characterization and in understanding their complex physical and chemical phenomena.

In summary, AI is not just a productivity tool, but a true scientific collaborator. It facilitates a crucial dialogue between different research results, including those from the past, accelerating discovery and providing new lenses to observe the microscopic world around us. The potential of AI in scientific research is still largely unexplored, but cases like the one now shown, related to gypsum, unequivocally demonstrate that it represents an invaluable resource for the future of materials science and beyond. Finally, it is highlighted how AI integrates, in the conversation with the researcher, a whole range of proven numerical approaches, and allows for the interaction to yield, in addition to text processing conversation, exportable Python calculation programs to be executed in other environments and further tested.

References

Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. Nature, 559(7715), 547-555.

Downs, R. T. (2006). The RRUFF Project: an integrated study of the chemistry, crystallography, Raman and infrared spectroscopy of minerals. Highlights in Mineralogy, 5, 23-28.

Gross, C., Al-Samir, M., Bishop, J. L., Poulet, F., Postberg, F., & Schubert, D. (2024). Prospecting in-situ resources for future crewed missions to Mars. Acta Astronautica.

Huang, W., Ertekin, E., Wang, T., Cruz, L., Dailey, M., DiRuggiero, J., & Kisailus, D. (2020). Mechanism of water extraction from gypsum rock by desert colonizing microorganisms. Proceedings of the National Academy of Sciences, 117(20), 10681-10687.

Krishnan, R. S. (1945, September). Raman spectra of the second order in crystals: Part I: Calcite. In Proceedings of the Indian Academy of Sciences-Section A (Vol. 22, No. 3, p. 182). New Delhi: Springer India.

Krishnan, R. S. (1945, October). Raman spectra of the second order in crystals: Part II. Gypsum. In Proceedings of the Indian Academy of Sciences-Section A (Vol. 22, pp. 274-283). Springer India.

Langevin, Y., Poulet, F., Bibring, J. P., & Gondet, B. (2005). Sulfates in the north polar region of Mars detected by OMEGA/Mars Express. Science, 307(5715), 1584-1586.

Rasetti, F. (1931). Raman Spectra of Crystals, Nature, 127(3208), 626-627.

Rasetti, F. (1932). Sopra l'effetto Raman nei cristalli. Il Nuovo Cimento (1924-1942), 9(3), 72-75.

Schmid, T., Jungnickel, R., & Dariz, P. (2020). Insights into the CaSO4–H2O system: A Raman-spectroscopic study. Minerals, 10(2), 115.

Sarma, L. P., Prasad, P. S. R., & Ravikumar, N. (1998). Raman spectroscopic study of phase transitions in natural gypsum. Journal of Raman spectroscopy, 29(9), 851-856.

Shi, E., Wang, A., Li, H., Ogliore, R., & Ling, Z. (2022). Gamma-CaSO4 With Abnormally High Stability From a Hyperarid Region on Earth and From Mars. Journal of Geophysical Research: Planets, 127(3), e2021JE007108.

Sparavigna, A. C. (2023). q-Gaussian Tsallis Line Shapes and Raman Spectral Bands. arXiv preprint arXiv:2307.11092.

Sparavigna, A. C. (2024). Gypsum Crystallization Water: Comparing a Laser Excited Raman Spectrum with a Mercury Resonance Radiation Excited Spectrum (Rasetti Technique). International Journal of Sciences, 13(09), 42-49.

Tsallis, C. (1988). Possible generalization of BoltzmannGibbs statistics. Journal of statistical physics, 52, 479-487.

Vaniman, D., Chipera, S., Rampe, E., Bristow, T., Blake, D., Meusburger, J., Peretyazhko, T., Rapin, W., Berger, J., Ming, D., & Craig, P. (2024). Gypsum on Mars: A Detailed View at Gale Crater. Minerals, 14(8), p.815.

van Susante, P. J., Allen, J., Eisele, T. C., Medici, E. F., & Zacny, K. (2018). Gypsum and other evaporites as a potential source for water extraction on Mars: experimental update. In 2018 AIAA SPACE and Astronautics Forum and Exposition (p. 5292).

Wojdyr, M. (2010). Fityk: a general-purpose peak fitting program. Journal of applied crystallography, 43(5), 1126-1128