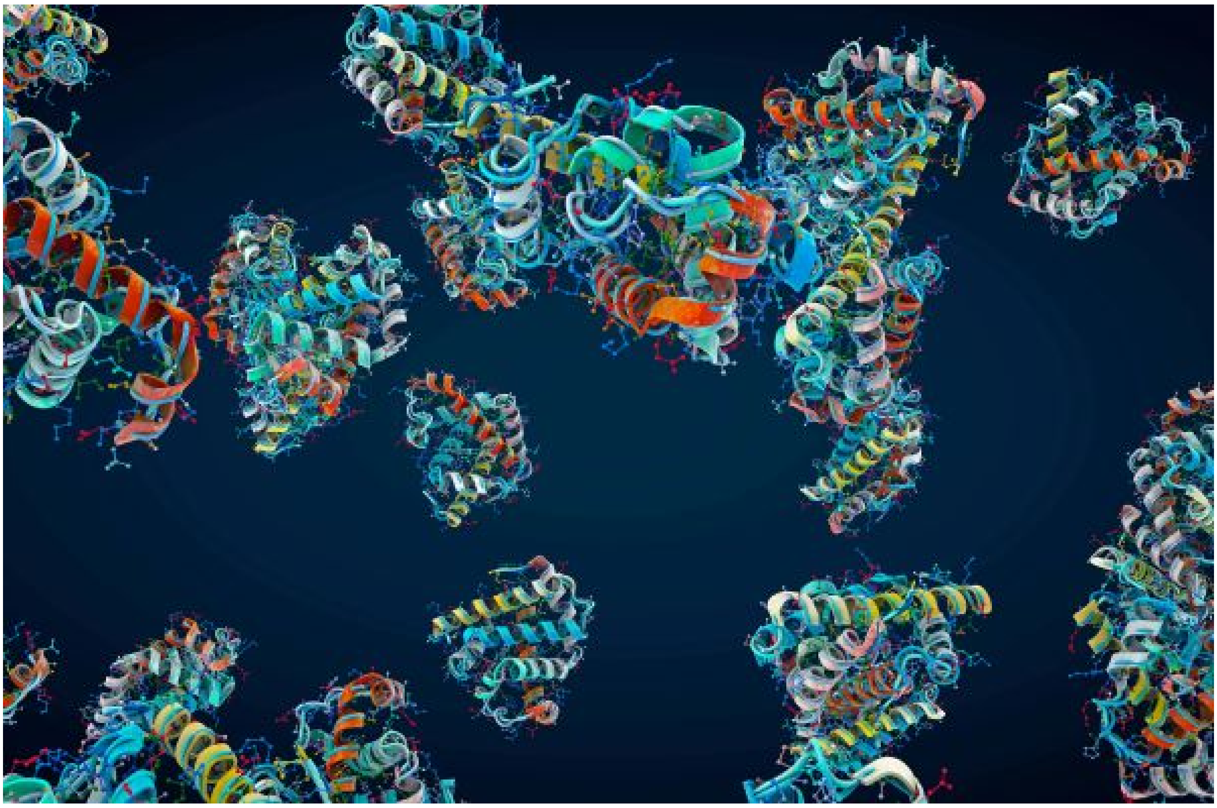

Pioneering open library to identify biomolecules

UOC and ICFO present a key resource for progress in research into diseases such as cancer. The project, based on the Raman spectroscopy analytical technique, is an open database to identify biomolecules that aims to grow with contributions made by the scientific community. The results are published in Chemometrics and Intelligent Laboratory Systems.

November 27, 2025

Researchers at the Universitat Oberta de Catalunya (UOC) and the Institute of Photonic Sciences (ICFO) have created a Raman spectral database that is accessible and open to the scientific community with 140 of the main types of biomolecules, including nucleic acids, proteins, lipids and carbohydrates. Raman spectroscopy is a technique that makes it possible to analyse the chemical composition and molecular structure of materials through the interaction of light with matter -specifically through the phenomenon of Raman scattering, which was discovered by the physicist Chandrasekhara Venkata Raman in 1928.

The study, Open Raman spectral library for biomolecule identification, published as open access in the journal Chemometrics and Intelligent Laboratory Systems, was led by UOC researchers, in collaboration with ICFO researchers **Jose Javier Ruiz** and **Dr. Pablo Loza-Alvarez**.

"One of the limitations of the potential of Raman spectroscopy in biomedical applications to date has been the lack of open spectral data for biomolecules. That is why we set out to create an accessible, standardized and useful library for the scientific community, which will act as the basis for future research and clinical applications," said Marcelo Teran, first author of the article.

In the project, the researchers implemented two search algorithms that proved to be 100% accurate in both top-ten identification of molecules, e.g. collagen, and in the identification of the type of molecule, e.g. protein, in measurements of pure biomolecules when replicating the results of previous studies.

Open biomedical data for progress in medicine

"Raman spectroscopy can be used to analyse the chemical composition of samples in a non-invasive way, which is very valuable in the field of medicine. This database can facilitate the precise identification of biomolecules and, in the future, it will contribute to studying how their presence varies in biological processes such as cancer," said Teran. "The availability of high-quality biomedical data is essential for progress in the development of AI-based solutions. This need was the starting point for the research."

The researchers collected data from Raman spectra of biomolecules from the leading articles published in the field, and developed an algorithm using classical computer vision techniques to extract the data automatically. One of the challenges in this project was the limited amount of spectral data published in open-access format, which they overcame using experimental validations. "Our work provides a tool that can help identify molecular composition based on its Raman spectrum in an objective, fast and standardized way. This identification is currently carried out by visual analysis of the main peaks in the spectra, and is compared with the references in the literature. Our tool can streamline this process while providing a standard solution that reduces human bias during analysis," said Teran.

A database destined to grow with contributions from the community

Looking ahead, the researchers hope that the scientific community will contribute to expanding the database, so that it becomes a leading collaborative Raman spectral library of biomolecules.

"It is still unusual for scientific articles to share data openly, especially in the field of Raman spectroscopy. This lack of access to data limits biomedical research considerably. If AI is to be successfully applied, it needs large volumes of reliable and accessible data, and this is where open science projects play a key role," said Teran.

The aim is that as the database expands, it will boost the training of artificial intelligence models in the field of molecular analysis of biological samples. This will create opportunities for new applications in the diagnosis and monitoring of diseases.

Reference:

Teran, M., Ruiz, J. J., Loza-Alvarez, P., Masip, D. y Merino, D. Open Raman spectral library for biomolecule identification. En: Chemometrics and Intelligent Laboratory Systems. 2025. Vol. 264, 105476. DOI: https://doi.org/10.1016/j.chemolab.2025.105476

Acknowledgements:

The authors acknowledge funding from Fundacio CELLEX; Ministerio de Economia y Competitividad - Severo Ochoa programme for Centres of Excellence in R&D (CEX2019-000910-S); CERCA programme (999619436); Laserlab-Europe (871124); Ministerio de Ciencia e Innovacion PID2021-122807OB-C31 and PID2022-138721NBI00 projects funded by MCIN/AEI/10.13039/501100011033/FEDER, UE; CARET project. The SLN facility corresponds to a i¿½Grup reconeguti¿½ 2021 SGR 01456 Departament de Recer
a i Universitats de la Generalitat de Catalunya.