Classification of high frequency SERS signal with convolutional neural networks

Vincent Thibault, Félix Lussier, **Jean-François Masson**
Département de chimie de l'Université de Montréal, Canada
vincent.thibault@umontreal.ca

Université de Montréal

The problem

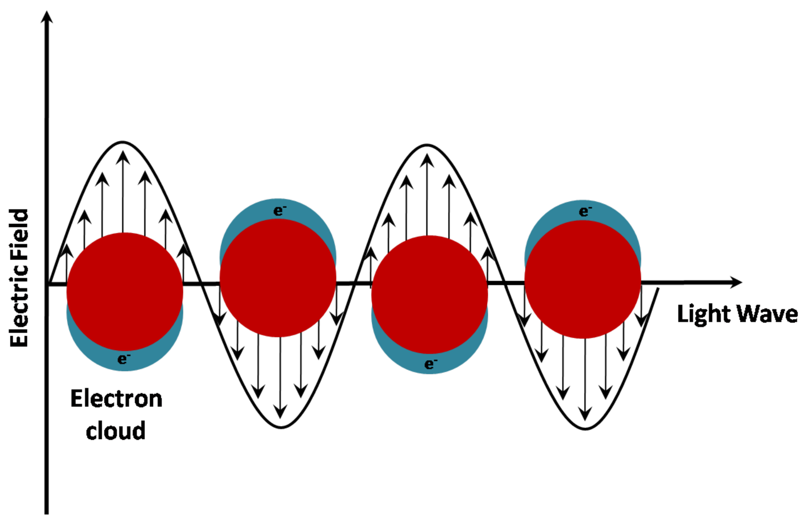
Surface enhanced Raman spectroscopy (SERS) is a subfield of Raman spectroscopy where trace compounds can be detected with the help of plasmonic materials, such as gold nanoparticles. SERS signal varies heavily with the local size of the nanoparticles and the orientation and distance of the molecules in respect of the above. At low integration time, it is the more obvious.
Classical algorithms are not efficients at sorting wildly different signals for a same class. As such, an alternative is needed.

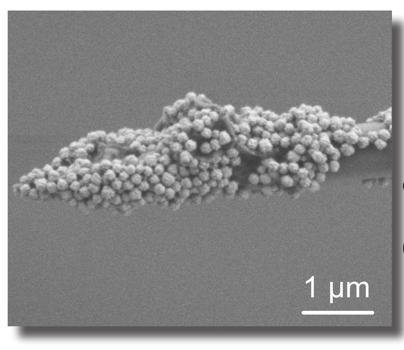
A classification model

based on a convolutional neural network (**CNN**) is used to demonstrate its classification and multiplexing capabilities with SERS spectra. A data pipeline was also made to automate the processing of raw spectra, with the goal of doing real-time classification.

Theory behind SERS

The oscillations of the surface electrons of the nanoparticles are called localized surface plasmons (LSP). At the right light wavelenght, they can absorb it and increase their own electrical field, enhancing the Raman signal from molecules in proximity.





Here, glass rods are stretched to sub-micron thickness and functionalized with a gold nanoparticle monolayer. This so-called nanopipette is able to reach the single molecule sensing domain, where a single, well positione molecule can be detected by the apparatus. With many useages, this nanopipette is the core of multiple articles of this research group.

Classification results

True label \ Predicted label	5HT	ACh	ADP	ATP	DA	GABA	Glu	Saline
5HT	0.94	0.01	0.01	0.02	0.01	0	0	0
ACh	0	0.94	0.02	0	0	0.01	0.03	0
ADP	0.01	0.01	0.97	0.01	0	0	0	0
ATP	0.01	0.01	0.02	0.93	0.03	0	0	0
DA	0	0	0	0.03	0.96	0	0	0
GABA	0	0.01	0.01	0	0	0.95	0.02	0.01
Glu	0	0	0	0	0	0.02	0.97	0
Saline	0	0	0	0	0	0	0	0.99

A normalized confusion matrix showing the accuracy of a multiclass CNN. All analytes are in 1µM aqueous solutions.Based on the test dataset's results, with about 3000 spectra.

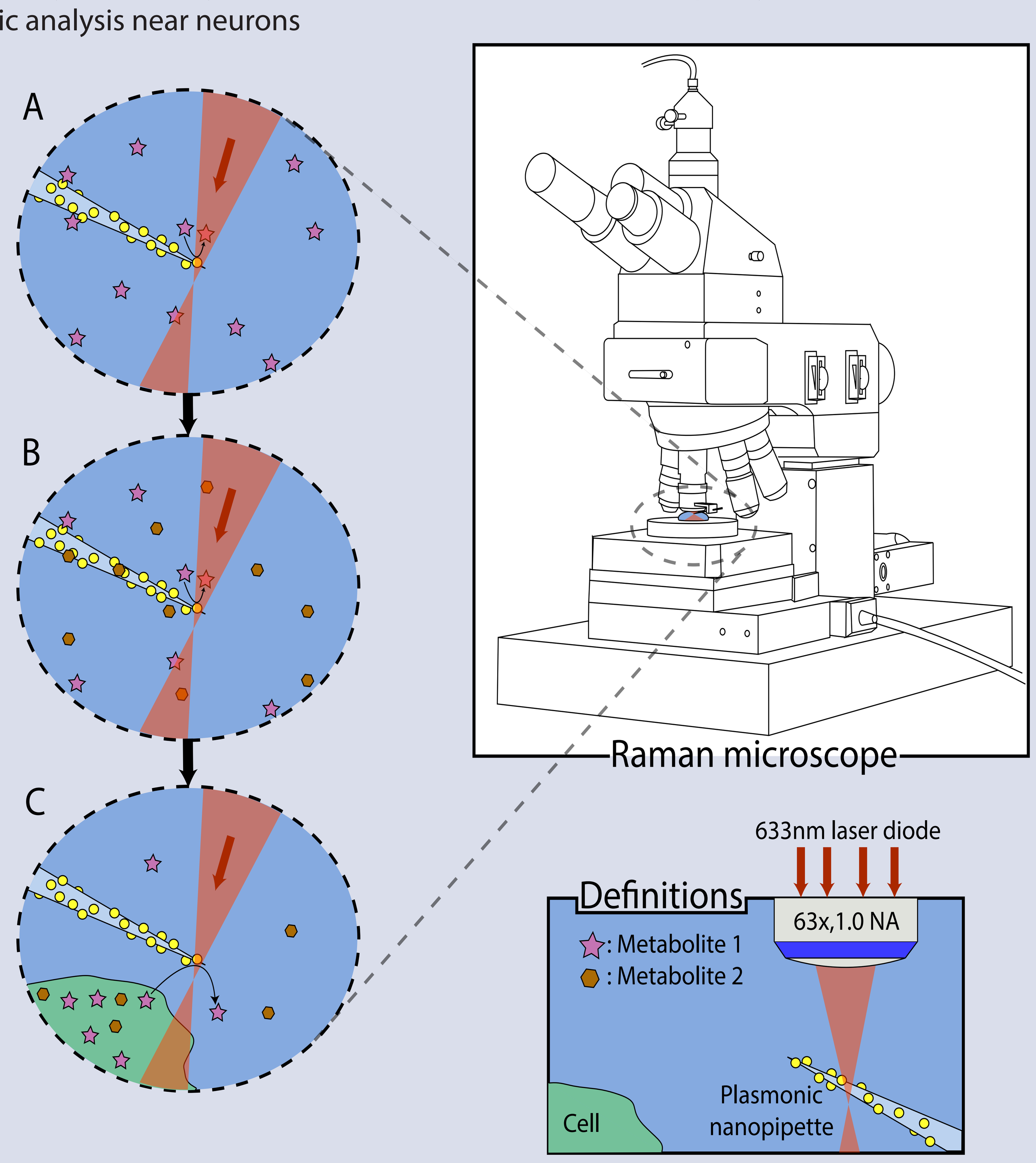
Future works

- Publication of open-source Python toolkit for spectral data processing and CNN model building.
- Unification of the preprocessing and inference for real time classification.
- Increase homogeinity of plasmonic surface to facilitate quantification
- Event-based quantification

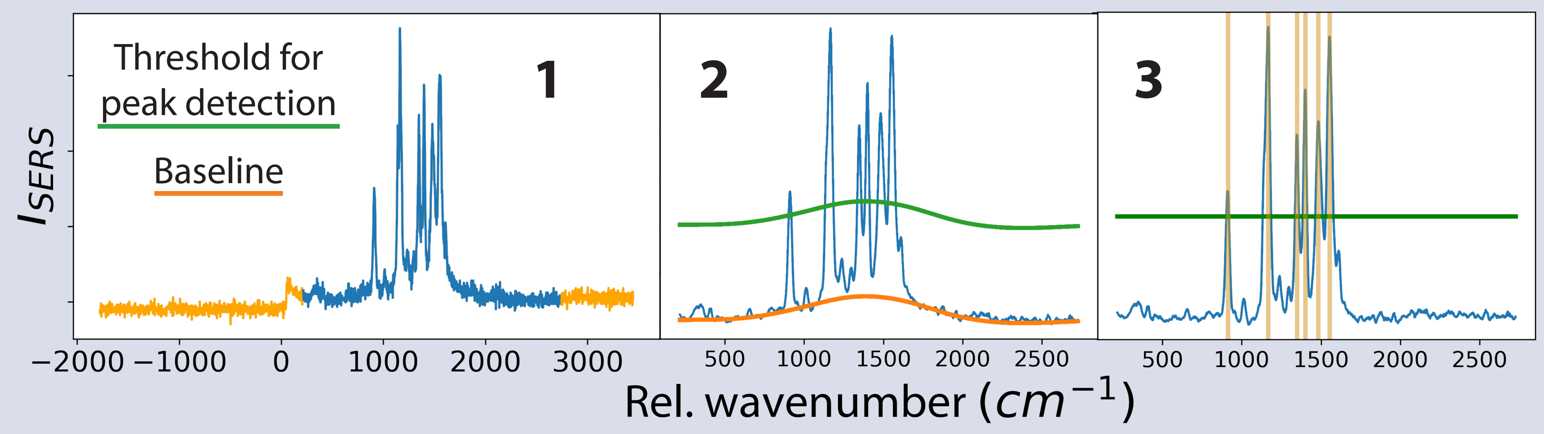
Protocol

1. Data gathering

A Analysis of each metabolite in water to make a database of the various spectra for all species.
B Multivariate analysis of binary mix of 2 metabolites to confirm the quality of the classificator.
C *In vitro* kinetic analysis near neurons



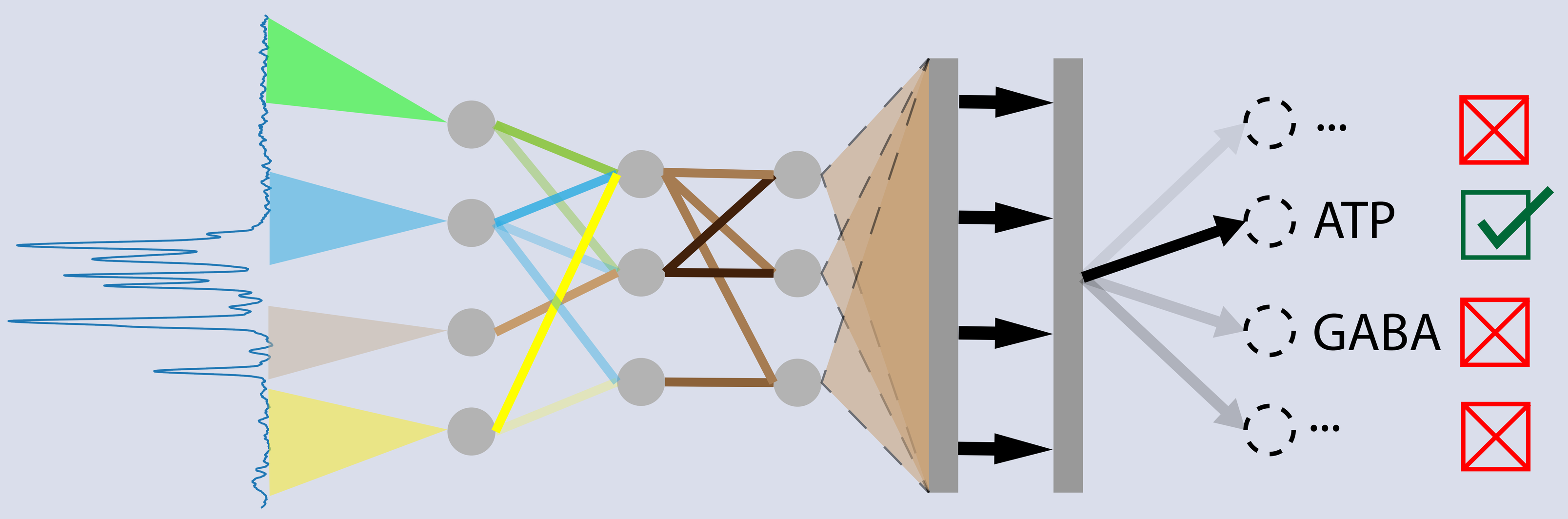
2. Preprocessing the spectra



1 Truncation of raw spectra to keep relevant pixels.
2 Spectra smoothing (Savinsky-Golay filter), baseline and peak threshold calculation (airPLS)
3 Baseline removal, peak finding and active/inactive classification

- The most important part of classification! "***Garbage in, garbage out.***"
- Before analyte classification, spectra must be sorted between **active** and **inactive** with classical algorithms.
- Only active spectra will be used for classification in the CNN.

3. Model building and inference



The CNN found to give the best result consist of 5 convolutional layer with maxpooling and batch normalization inbetween, 2 dense layers and an output layer.

Any type of data that vaguely ressemble an image will be a good candidate to try to use CNNs for its classification.

INNOVATION.CA CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION
Fonds de recherche Nature et technologies Québec
CSACS CRMAA
NSERC CRSNG