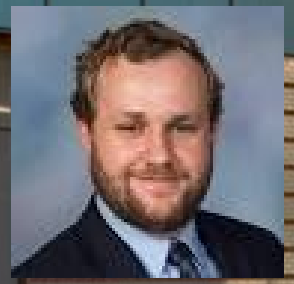
Merging Surface-Enhanced Spectroscopies with Machine Learning for Detection of Environmental Contaminants


Melissa Suter


Ankit Patel

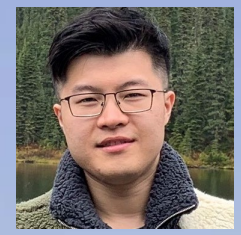

Bhagavatula Moorthy

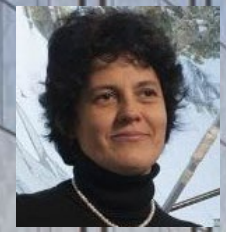

P. Nordlander

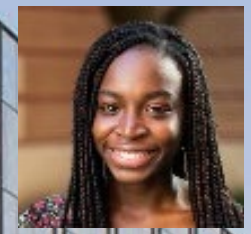

Thomas Senftle

N. J. Halas, Rice University


Jingyi Zhou


Yilong Ju


Oara Neumann


Mary Bajomo


Andres Sanchez


Lexie Chen

Seminar, Stuttgart, Germany: October 17, 2024

Gas Chromatography-Mass Spectrometry: large, costly, time-consuming

PROJECT VISION:

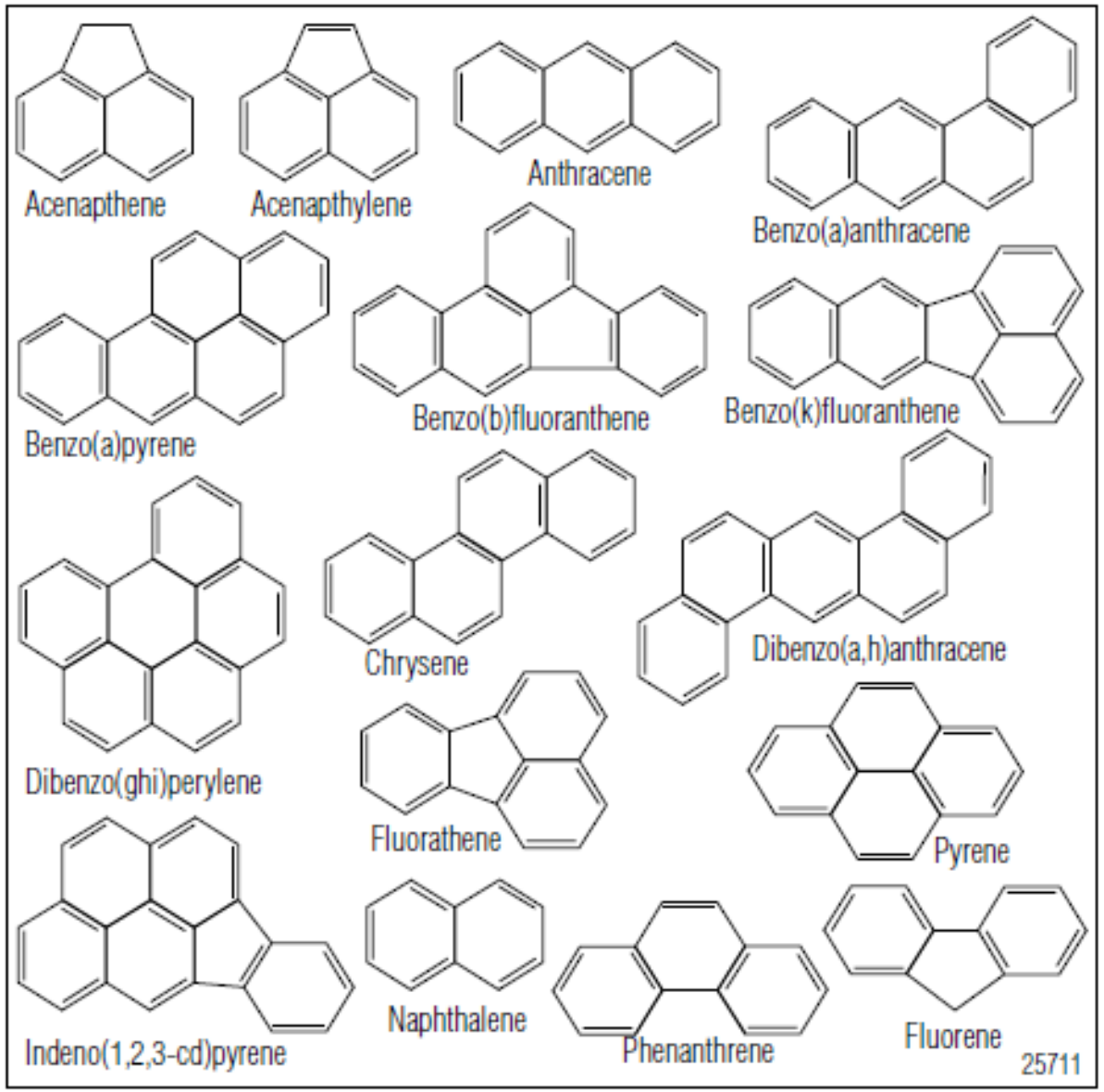
Can we

–rapidly and inexpensively- detect and identify molecules from complex environmental samples by combining Surface-enhanced Spectroscopy with Machine Learning strategies?

Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

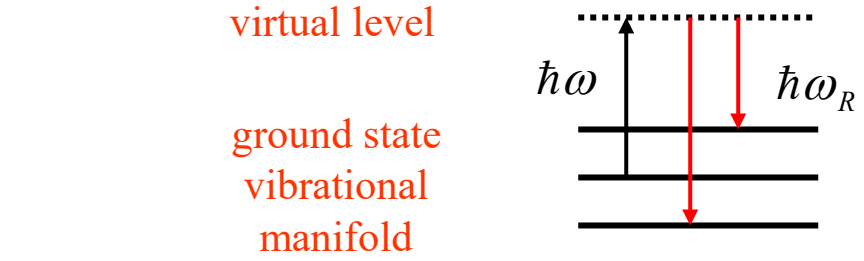
Polycyclic Aromatic Hydrocarbons (PAHs)



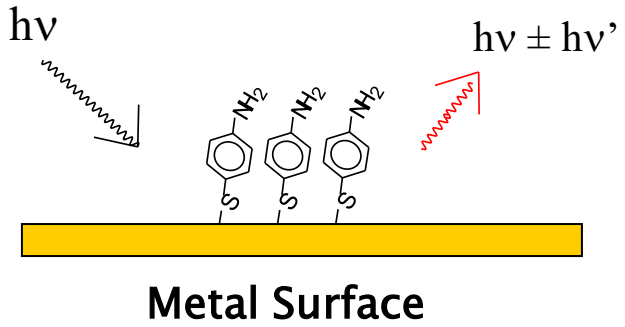
- 16 Priority pollutants (EPA): toxic, carcinogenic, mutagenic
- Found in atmosphere, in soil and in virtually all water sources: “legacy pollutants”
- Many societal sources: incomplete combustion, diesel fuel, sidestream tobacco smoke, barbecue, burnt toast, drinking water
- Many cancers (skin, liver, testicular, bladder)
- Adult onset diseases (Parkinson’s?)
- Poor fetal development: increases in premature birth rates
- **Always found in complex mixtures**

Surface Enhanced Raman Spectroscopy (SERS)

Molecular adsorbates on metal surfaces respond to the local electromagnetic field, emitting inelastically scattered light (Van Duyne, 1974):



Stokes, anti-Stokes emission



spectrum enhanced by near field when molecule in the vicinity of a plasmonic substrate/metasurface

Plasmonic substrate requires intense near field at pump, Stokes frequencies:

$$\underbrace{\langle |E_{NS}(\omega_L)|^2 \cdot |E_{NS}(\omega_s)|^2 \rangle}_{} \approx \underbrace{\langle |E|^4 \rangle}_{}$$

Surface Enhanced Infrared Absorption Spectroscopy (SEIRA)

- IR absorption, $A \propto \left| \mathbf{E} \bullet \left(\frac{\partial \boldsymbol{\mu}}{\partial Q} \right) \right|^2$

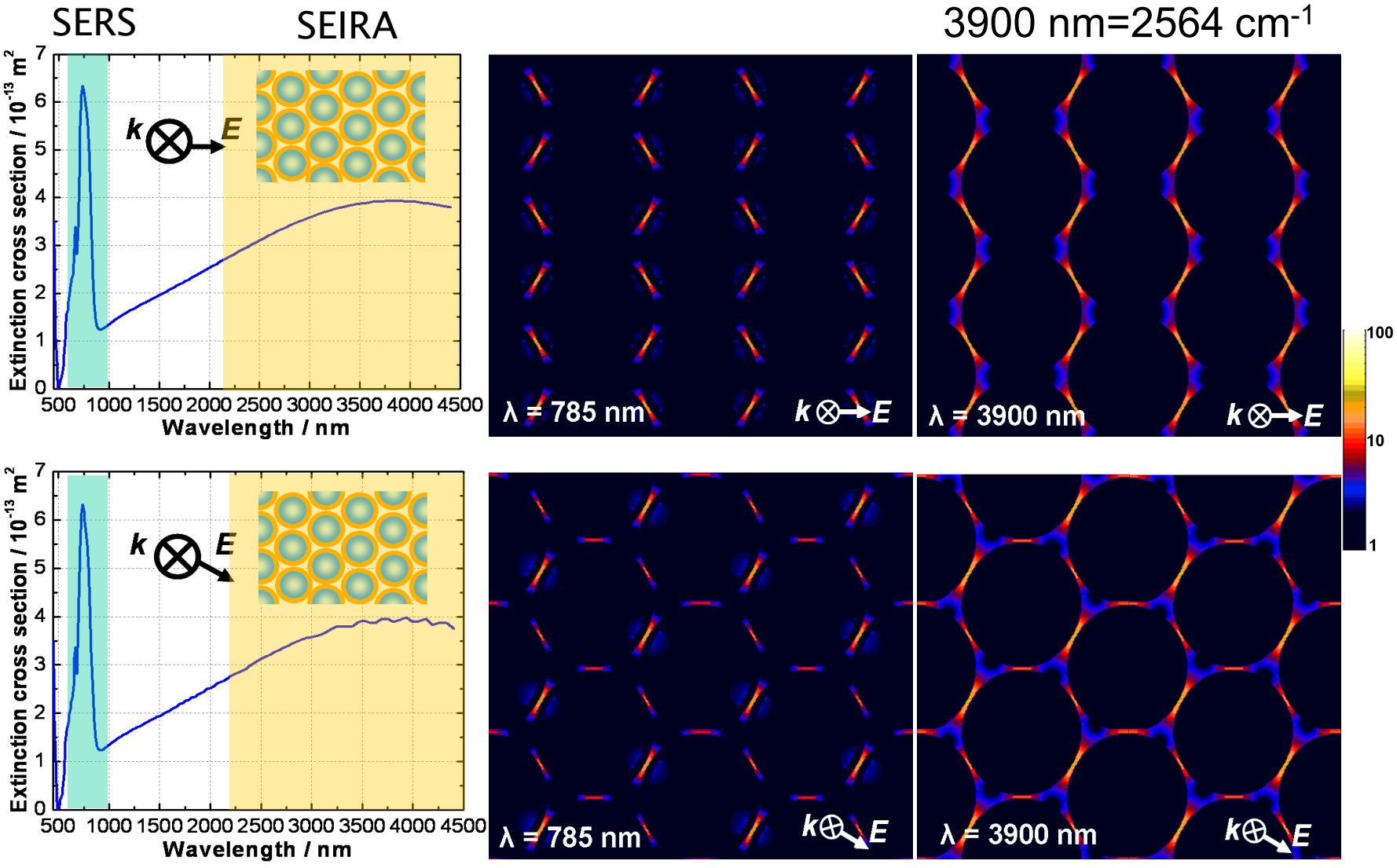
E = Electric field
μ = Dipole moment
Q = Normal coordinate

- SEIRA ~ $|E|^2$, not $|E|^4$ like SERS
- spans near-IR to far-IR regime
- IR cross sections ~10^{10} larger than Raman cross sections!

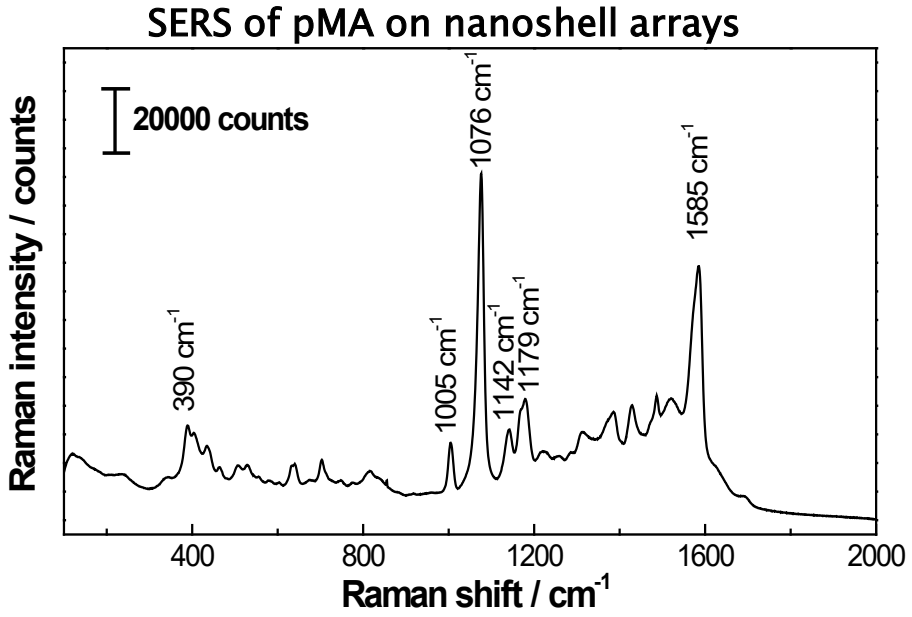
SERS and SEIRA combined on the same substrate

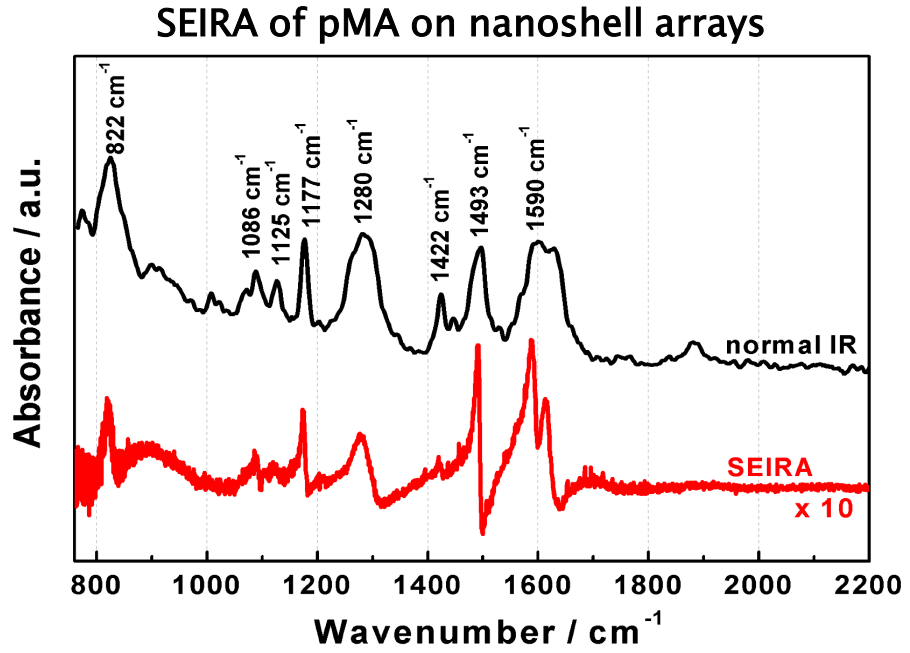
H. Wang et al., *Angewandte Chemie International Edition* **46**, 9040-9044 (2007).

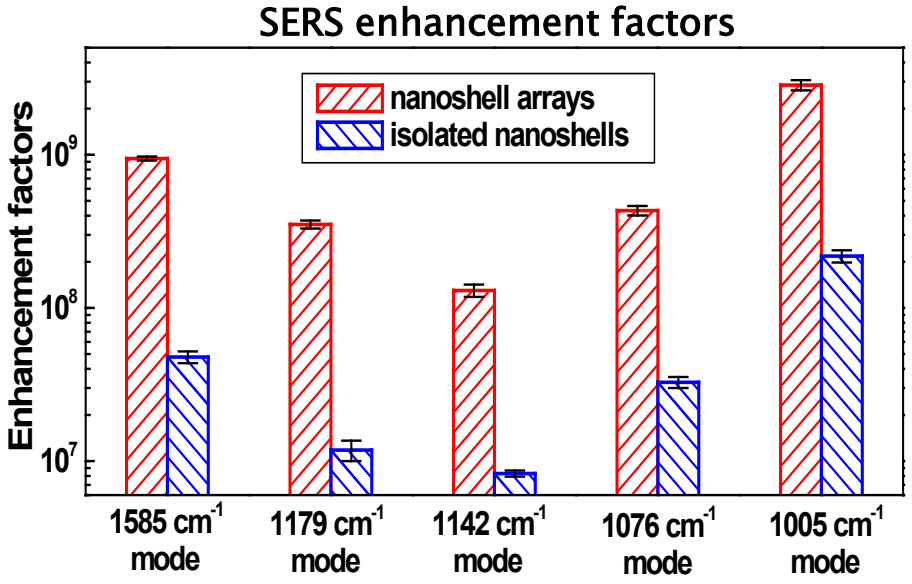
F. Le et al., Metallic nanoparticle arrays: a common substrate for both SERS and SEIRA, *ACS Nano* **2,** 707-718 (2008)

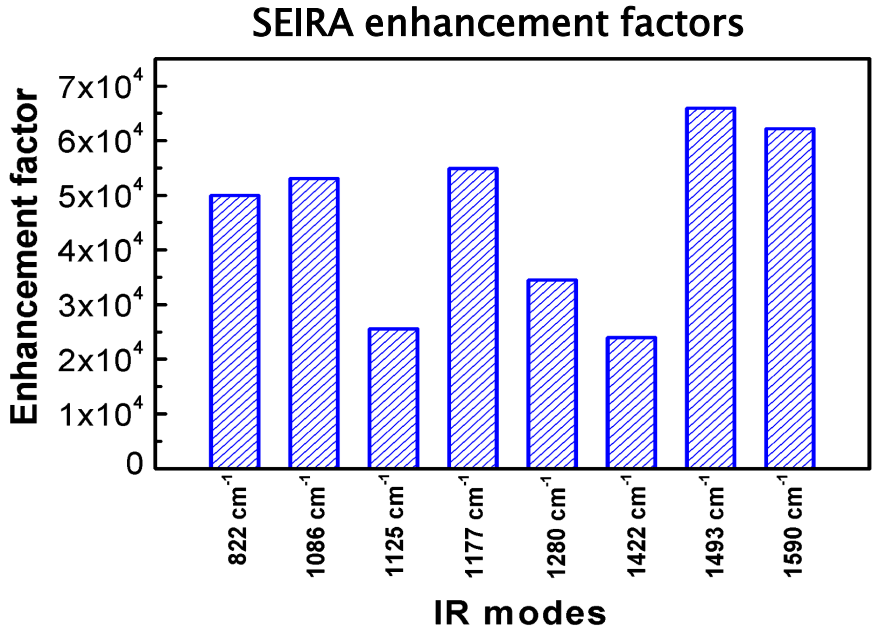


Combined SERS and SEIRA on the same substrate









Outline:

- **Computational Chromatography: identifying unknown molecules in mixtures without separations**

(M. Bajomo, Y. Ju et al., PNAS 119, e2211406119 (2022))

- **Identifying SERS spectra with a Raman library using Machine Learning**

(Y. Ju et al., ACS Nano 17, 21251-21261 (2023))

- **Machine Learning-enhanced SERS + SEIRA Detection of Polycyclic Aromatic Hydrocarbons in Human Placenta**

(O. Neumann et al., TBP)

Computational Chromatography: a Machine Learning strategy for demixing individual chemical components in complex mixtures

M. Bajomo, Y. Ju, et al., PNAS 119 (2022) e2211406119

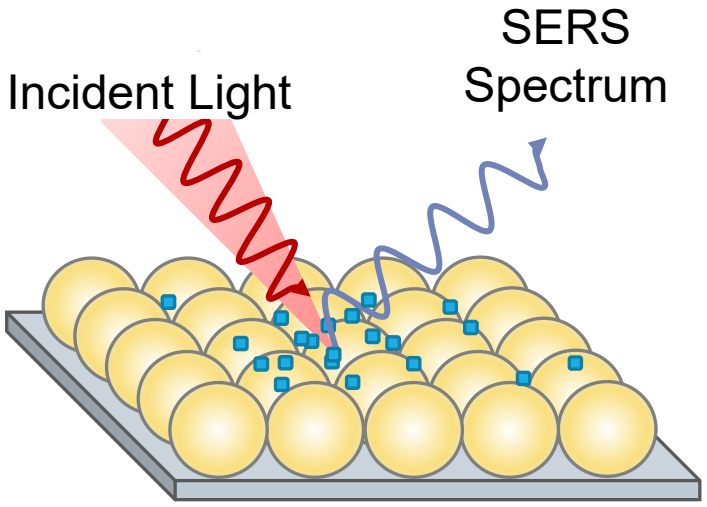
Can we identify the chemical components in a mixture without physically separating them?

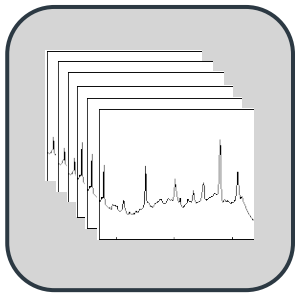
Surface Enhanced Raman Spectroscopy (SERS)

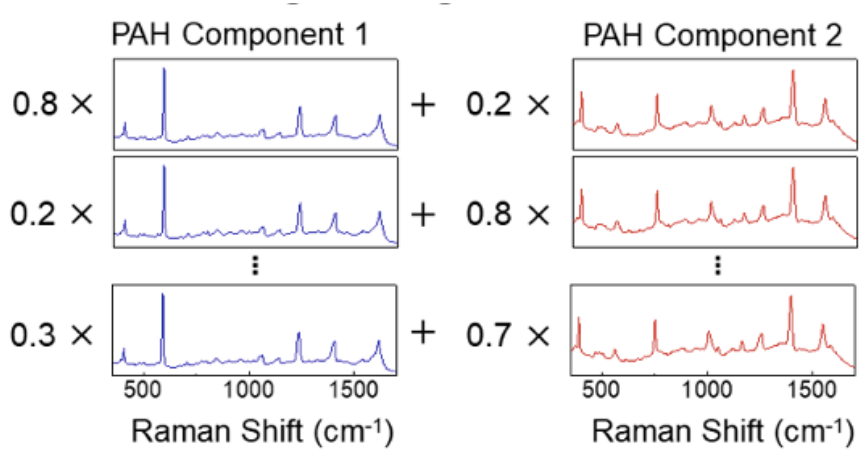
&

Unsupervised ML Demixing Algorithm = NO LIBRARIES

Rapid, ultrasensitive detection and identification of PAHs from "unknown" PAH mixture







Machine Learning (ML) algorithms separate out vibrational spectra of mixture components

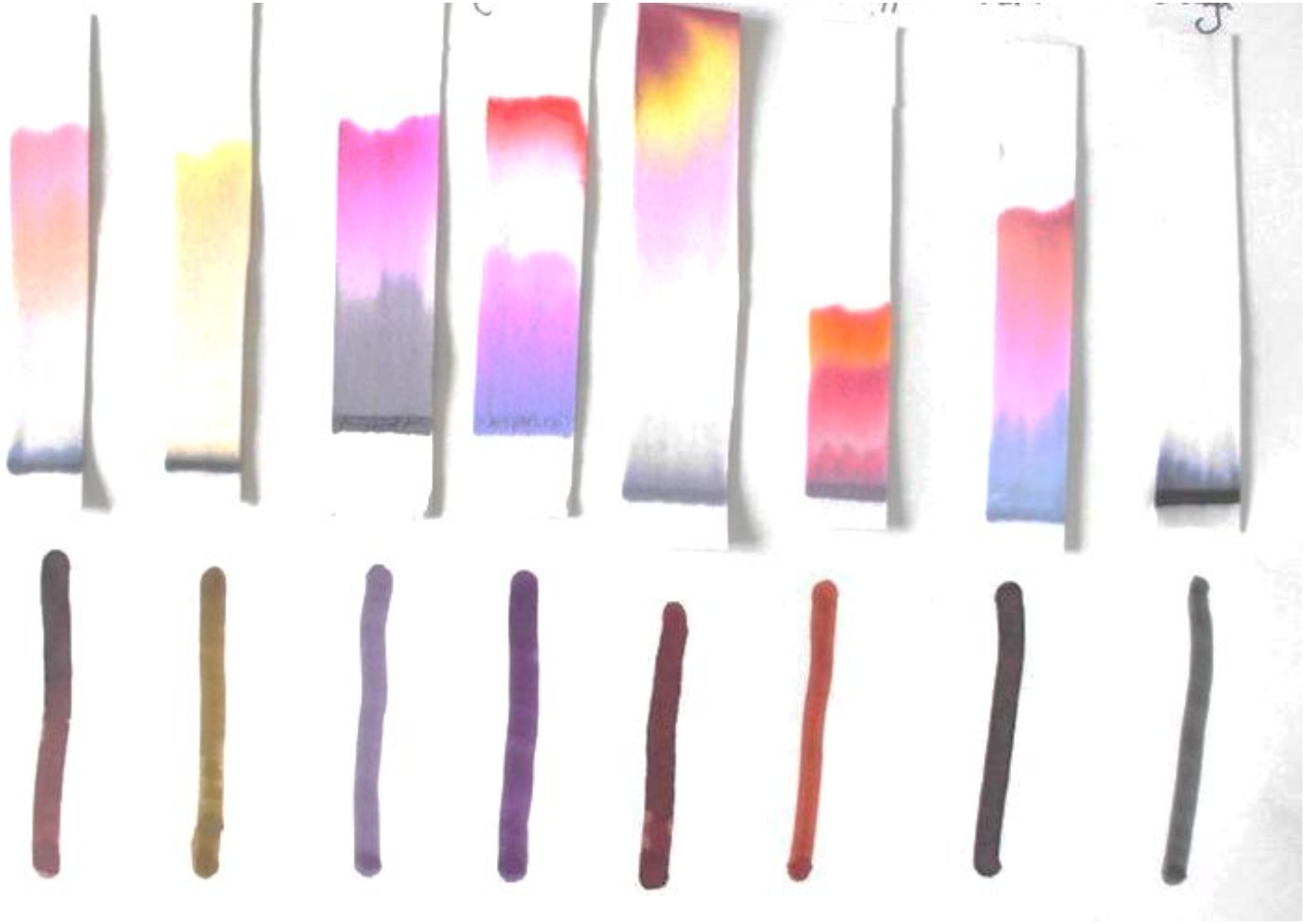
Yilong Ju

Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

Computational Chromatography: a Machine Learning strategy for demixing individual chemical components in complex mixtures

M. Bajomo, Y. Ju, et al., PNAS 119 (2022) e2211406119



Can we identify the chemical components in a mixture without physically separating them?

The Cocktail Party Problem

(aka "Blind Source Separation")

Independent Component Analysis:

Key Assumptions about the Data-Generating Process

Independent component analysis (ICA) is a ML method for separating a multivariate signal into additive subcomponents.

Signals must be linearly mixed

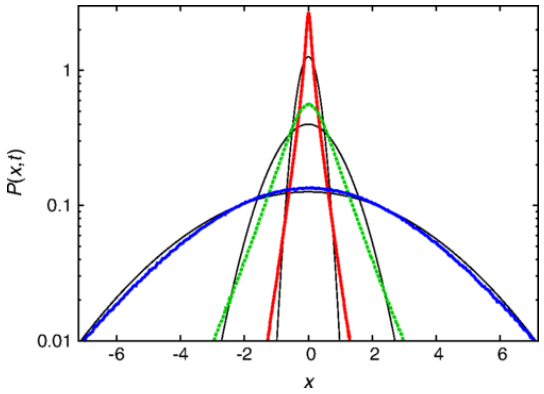
$$x_1 = a_{11}s_1 + a_{12}s_2$$
$$x_2 = a_{21}s_1 + a_{22}s_2$$

$x = detected\ signal$

$s = source\ signal$

$a = mixing\ weight$

Signals must be non-gaussian

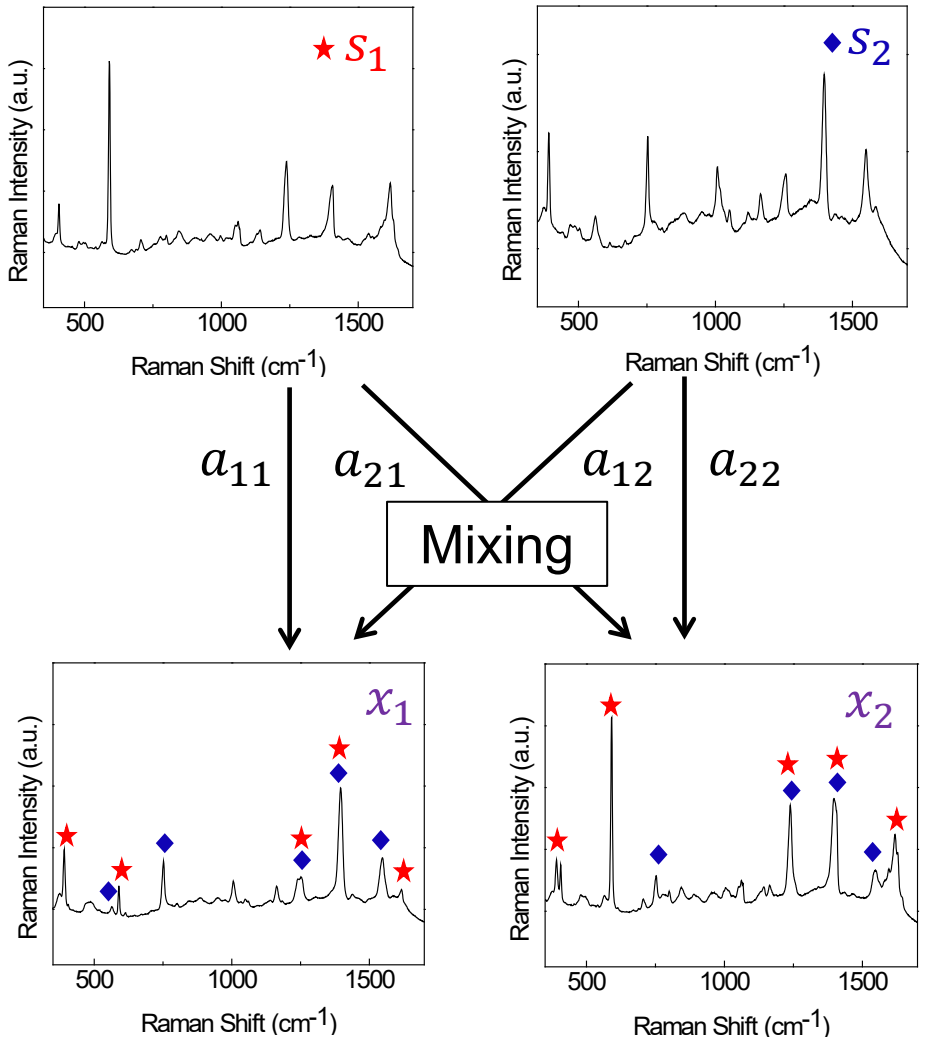


Signals must be independent

$$p(s_1, s_2) = p(s_1)p(s_2)$$

Chechkin, Aleksei V., et al. "Brownian yet non-Gaussian diffusion: from superstatistics to subordination of diffusing diffusivities." *Physical Review X* 7.2 (2017): 021002.

ICA Approximately Demixes to Recover Source Signals (Individual Component Spectra)



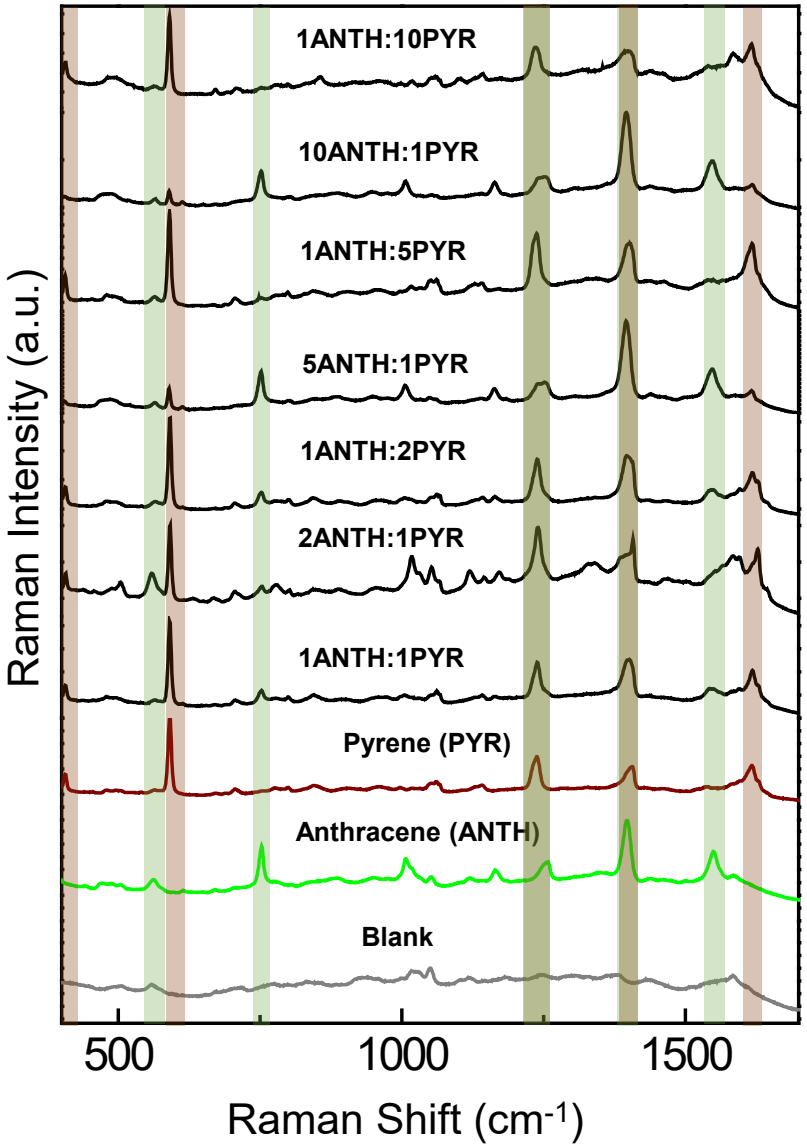
$$x_1 = a_{11}s_1 + a_{12}s_2$$
$$x_2 = a_{21}s_1 + a_{22}s_2$$

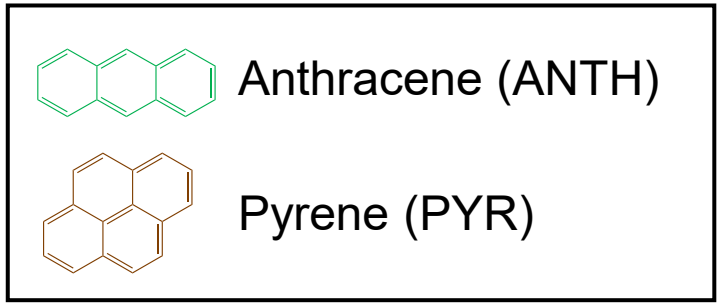
$$x = As$$
$$s = A^{-1}x$$

$$\hat{s} = Wx$$

Goal of ICA: Find *de*mixing matrix *W* (approximation of A^{-1}) so that $\hat{s} \approx s$.

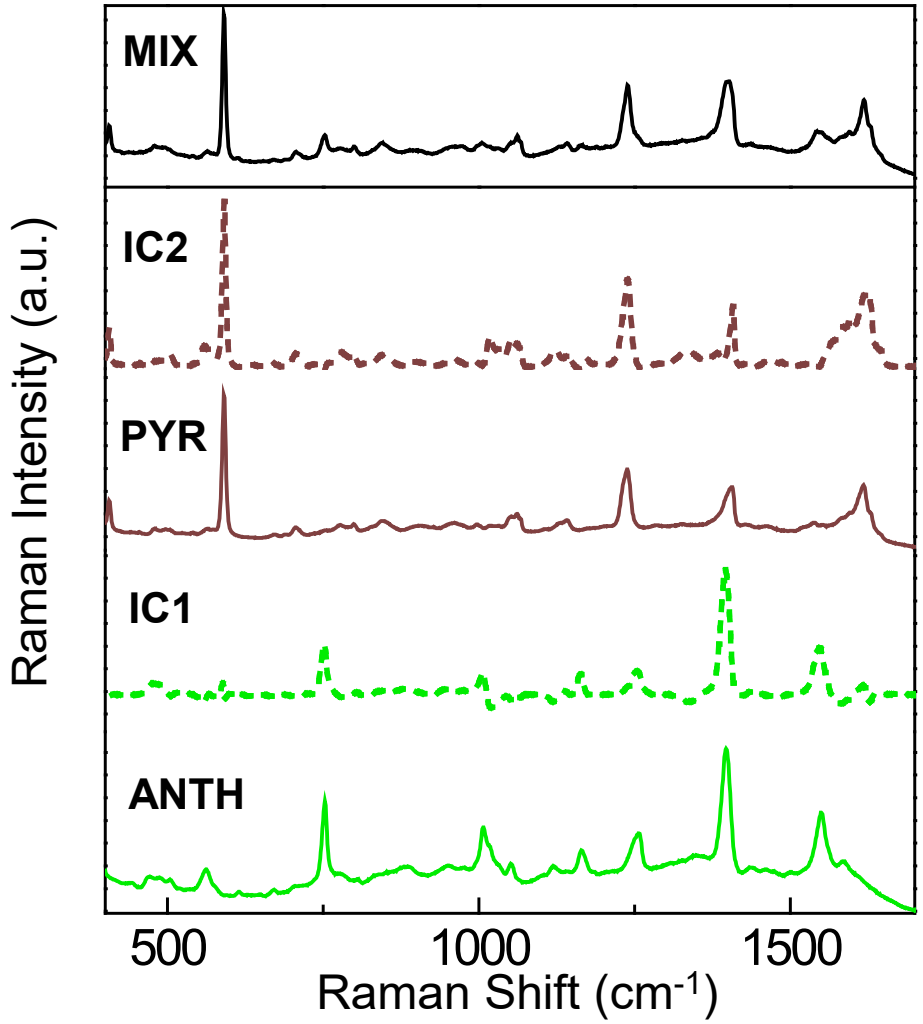
Example: SERS of 2-Component Mixtures of PAHs

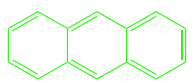




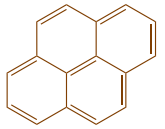
Relative intensities of different peaks in the SERS of ANTH+PYR varies as the concentration ratio of ANTH:PYR varies.

Estimated Independent Components Match PAH Source Spectra





Anthracene (ANTH)



Pyrene (PYR)

Independent components produced match the SERS of the components of the mixture

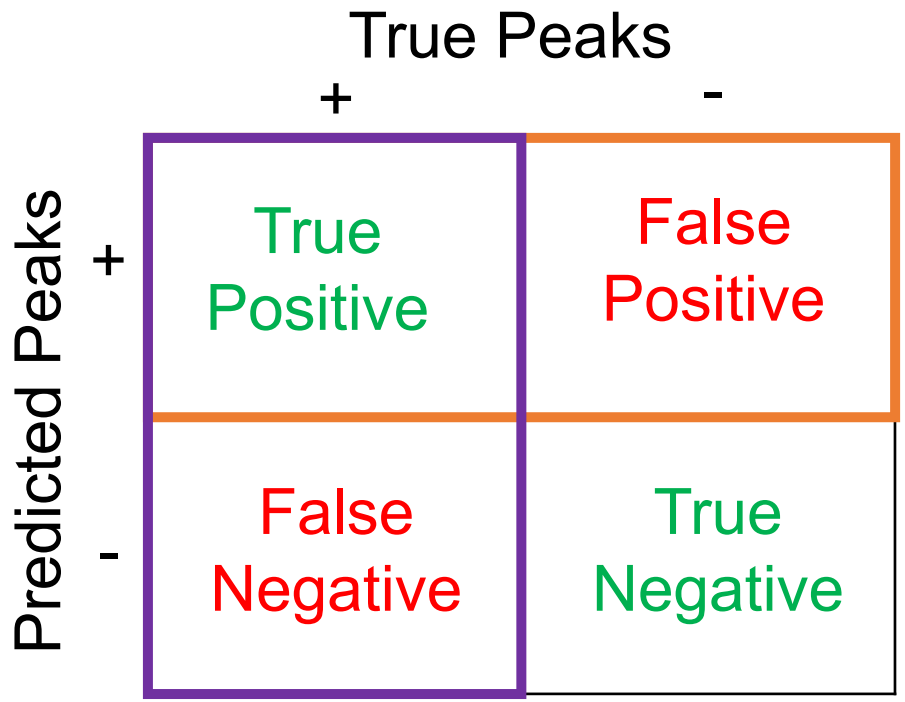
Quantifying Demixing Performance:
Evaluation Metrics

-More peaks detected
-More likely to recover original peaks

Larger detection threshold →

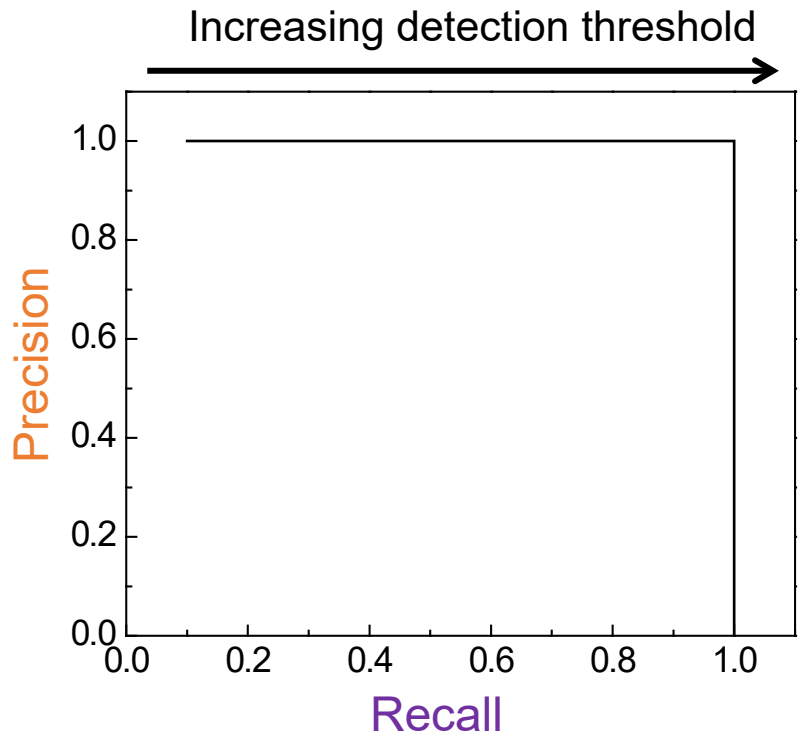
← Smaller detection threshold

-Fewer peaks detected
-Less likely to recover original peaks

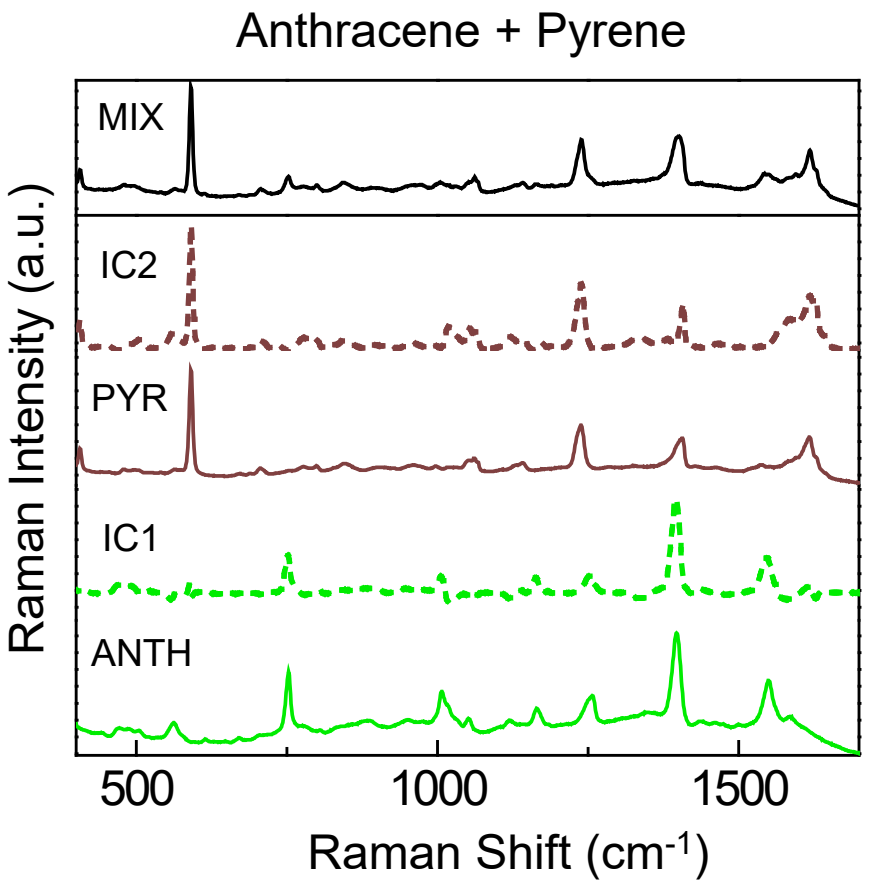


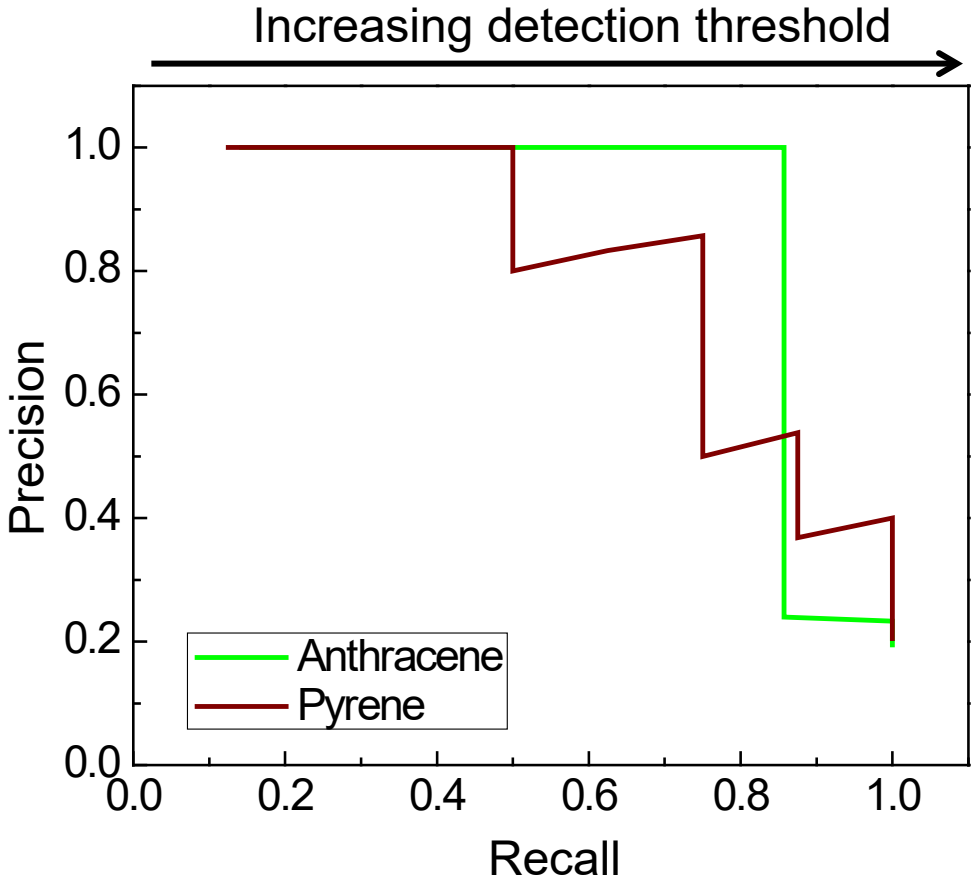
Precision= TP/(TP+FP)

Recall=TP/(TP+FN)



Performance of ICA on 2-Component Mixtures





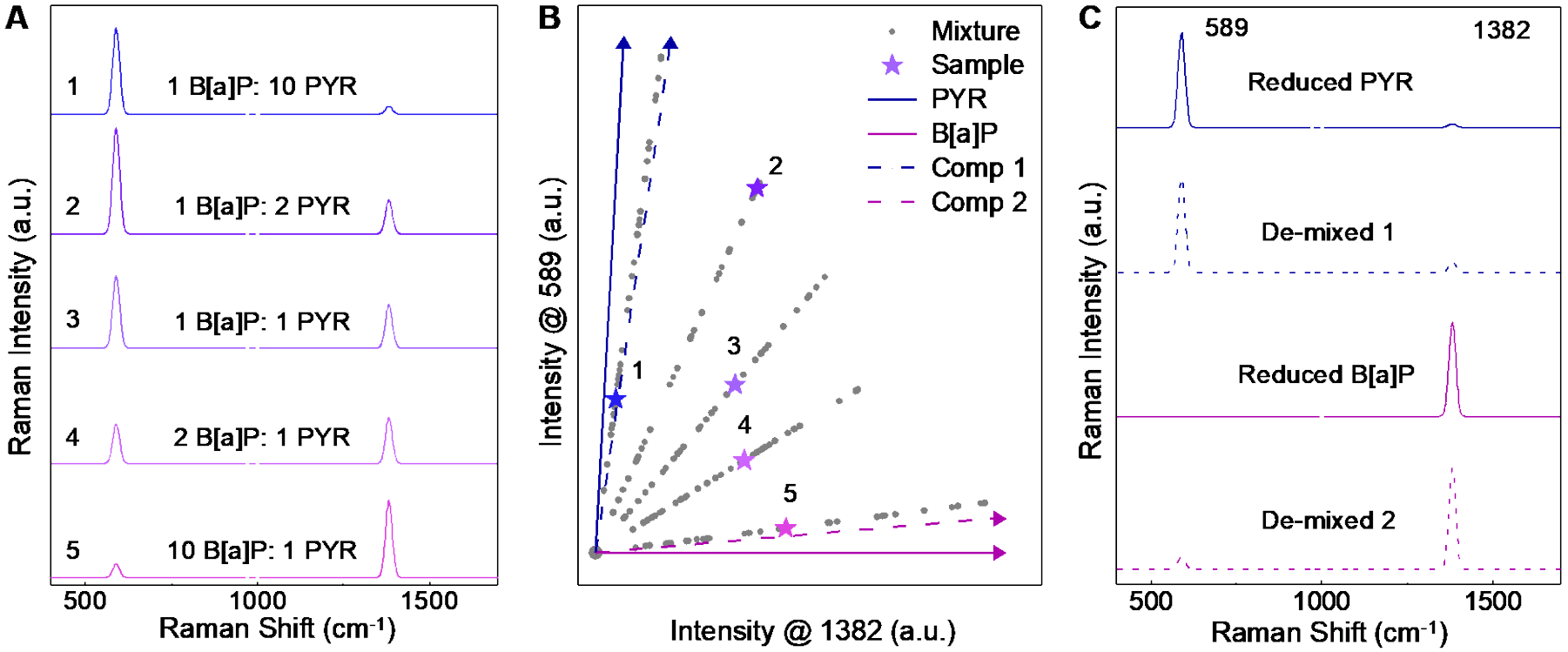
How Demixing works “under the hood”:

Multiple mixture spectra with varying concentration ratios are required

Synthetic mixture spectra of varying ratios

Spectra in a 2D space using the intensity at the two peaks as x and y-coordinates.

Original and demixed components



High-throughput sampling with naturally-occurring concentration variations makes this straightforward

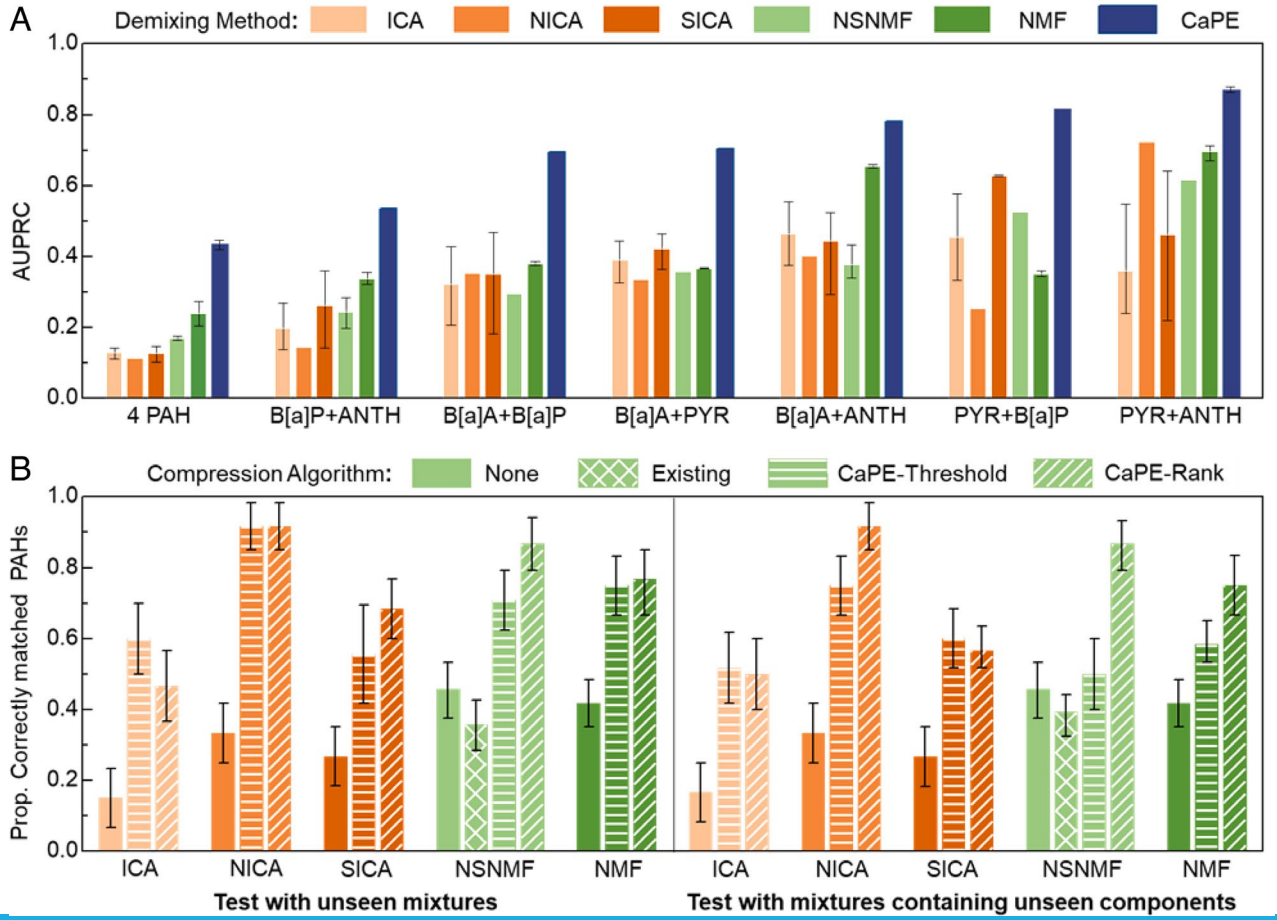
Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

Comparing Different ML-based Demixing Algorithms:
Informing based on Prior Domain Knowledge is Valuable

Demixing Algos with various assumptions/constraints imposed (aka *inductive biases*):

- Independent Components Analysis (ICA)
- Nonnegative ICA (NICA)
- Sparse ICA (SICA)
- Nonnegative Matrix Factorization (NMF)
- Near-Separable NMF (NSNMF)
- Characteristic Peak Extraction-based NMF (CaPE)



Extracting Characteristic Peaks in a shift-tolerant manner significantly improves performance

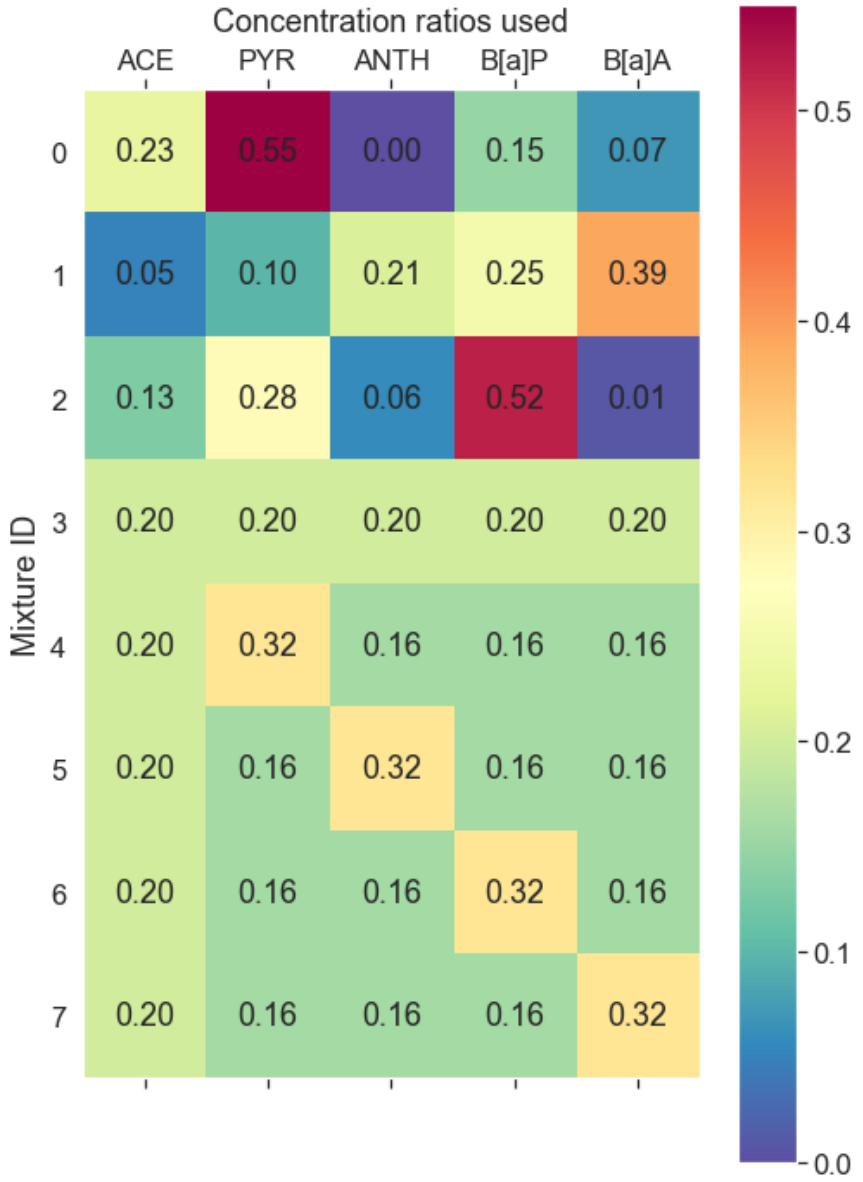
Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

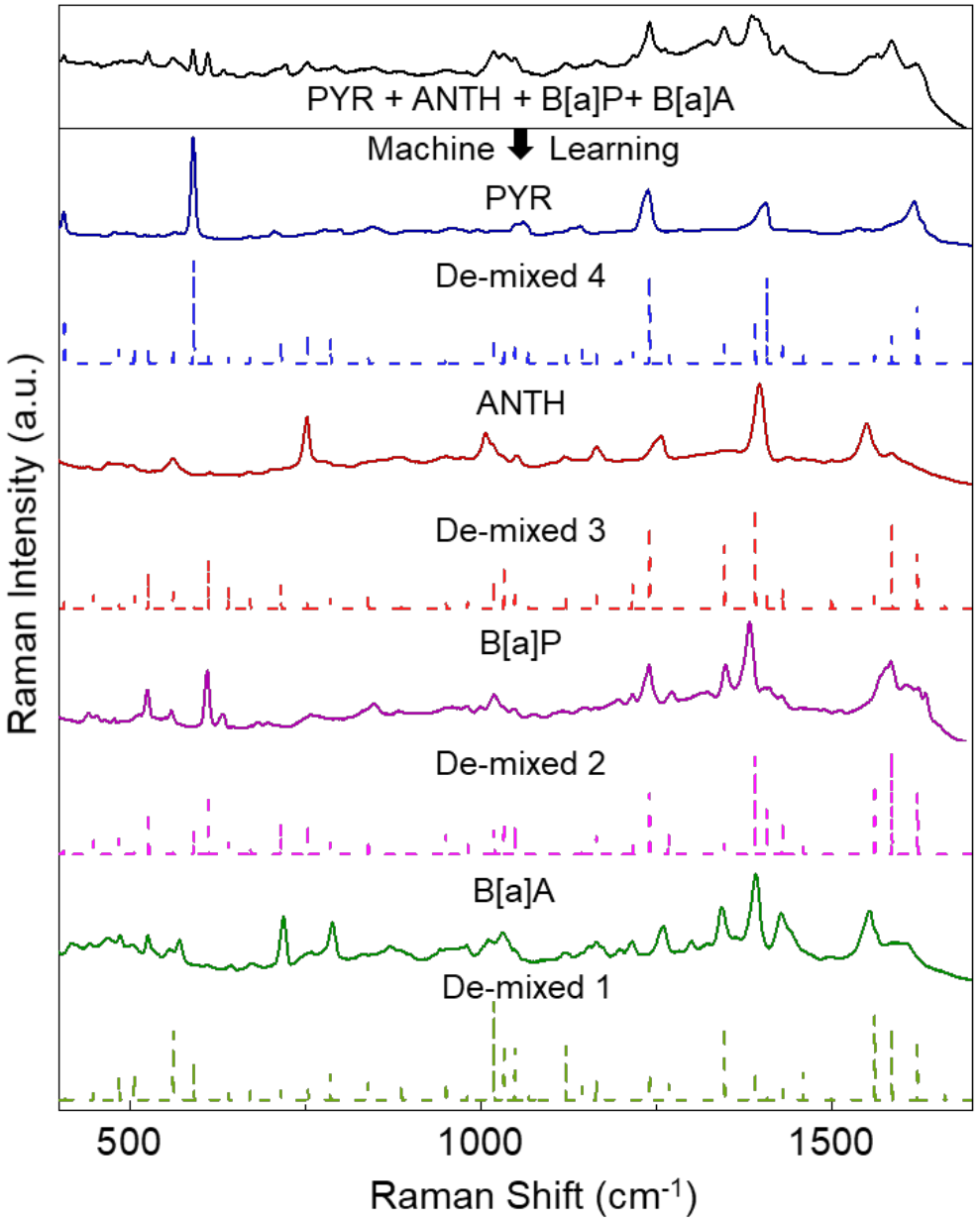
Can we extend this to multiple components?

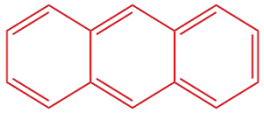
How few samples do we need with varied concentrations?

A synthesized 4-component mixture:
8 concentration ratios

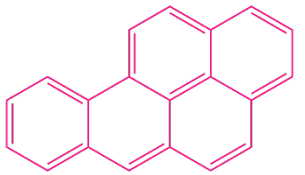


Four Component PAH Demixing

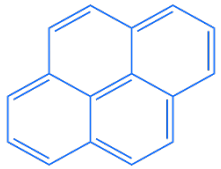




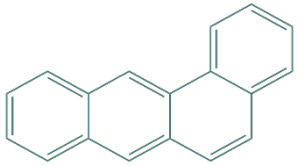
Anthracene (ANTH)



Benzo[a]pyrene)B[a]P)



Pyrene (PYR)



Benz[a]anthracene (B[a]A)

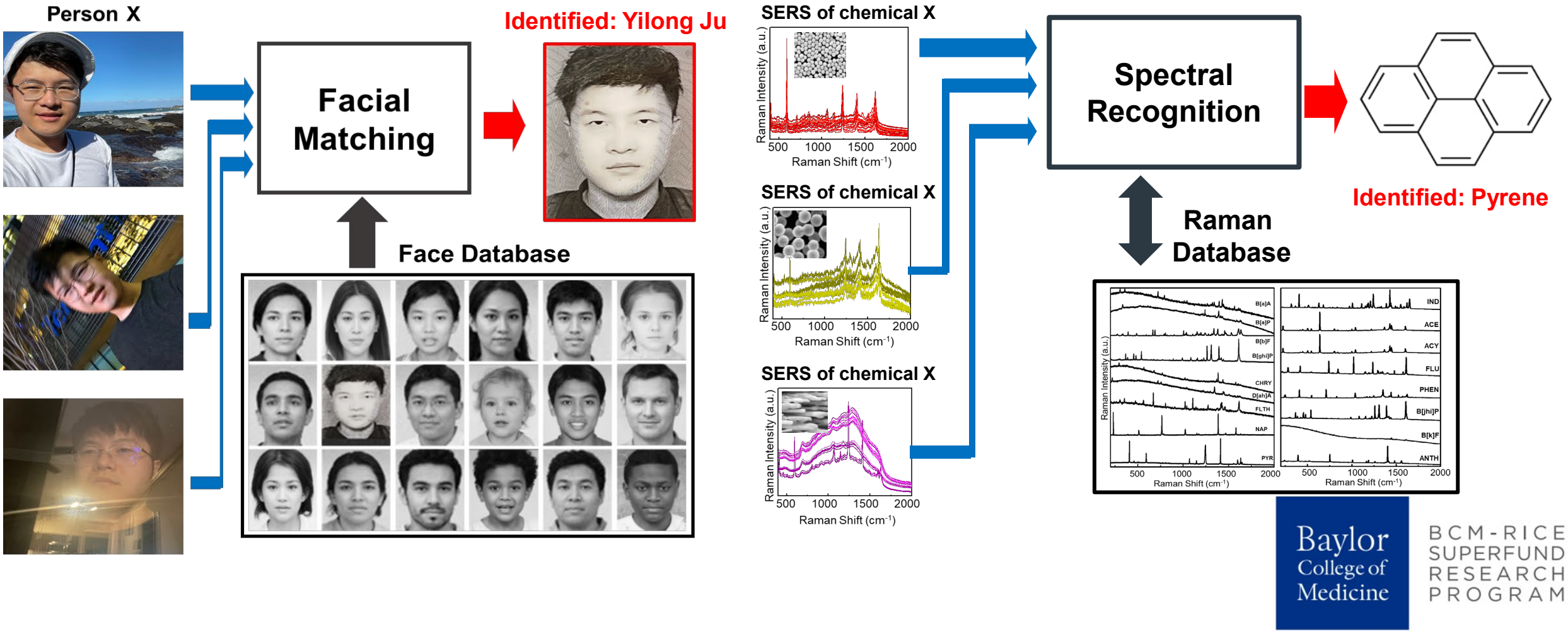
- Despite more peak overlap, there is acceptable agreement between demixed spectra and PAH component spectra

Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

Can we use Facial Recognition Approaches to identify SERS spectra of molecules using a Raman Database?

Yilong Ju et al., ACS Nano 17, 21251-21261 (2023).



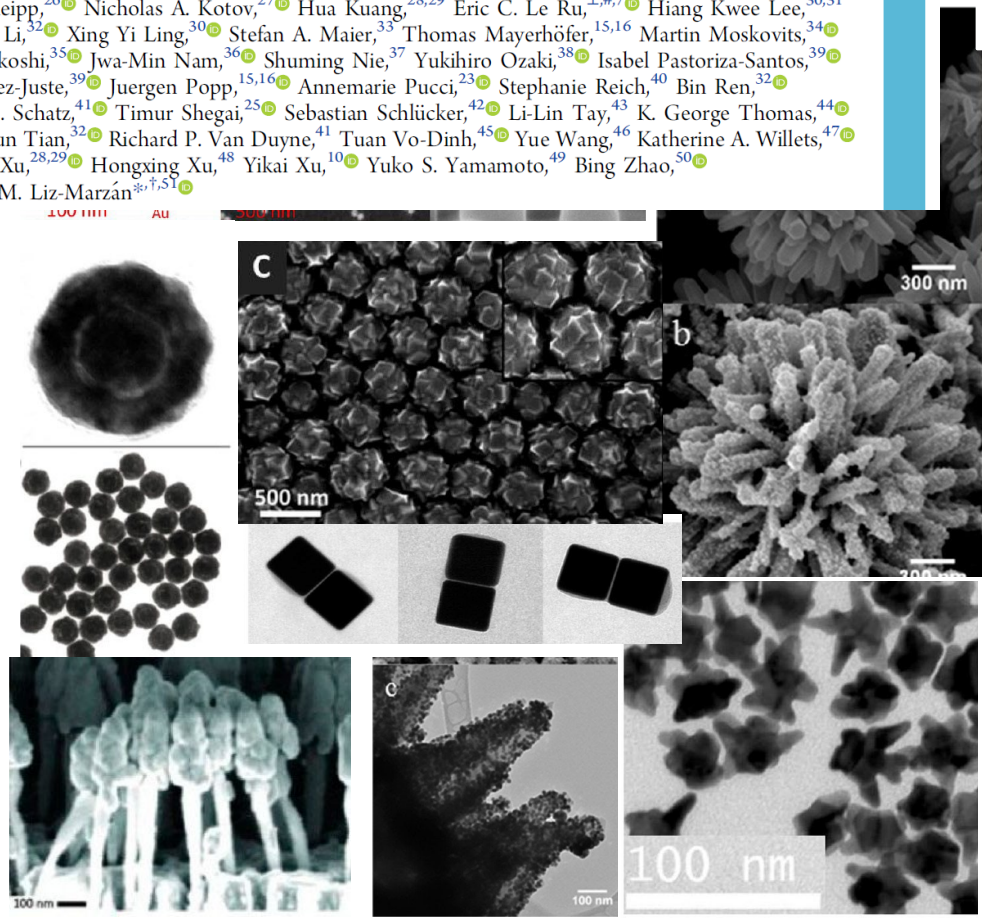
Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

ACS NANO
Cite This: ACS Nano 2020, 14, 28−117
www.acsnano.org
REVIEW

Present and Future of Surface-Enhanced Raman Scattering

Judith Langer,[†] Dorleta Jimenez de Aberasturi,[†] Javier Aizpurua,[‡] Ramon A. Alvarez-Puebla,[§,∥] Baptiste Auguié,[⊥,#,7] Jeremy J. Baumberg,[8] Guillermo C. Bazan,[9] Steven E. J. Bell,[10] Anja Boisen,[11] Alexandre G. Brolo,[12,13] Jaebum Choo,[14] Dana Cialla-May,[15,16] Volker Deckert,[15,16] Laura Fabris,[17] Karen Faulds,[18] F. Javier García de Abajo,[∥,19] Royston Goodacre,[20] Duncan Graham,[18] Amanda J. Haes,[21] Christy L. Haynes,[22] Christian Huck,[23] Tamitake Itoh,[24] Mikael Käll,[25] Janina Kneipp,[26] Nicholas A. Kotov,[27] Hua Kuang,[28,29] Eric C. Le Ru,[⊥,#,7] Hiang Kwee Lee,[30,31] Jian-Feng Li,[32] Xing Yi Ling,[30] Stefan A. Maier,[33] Thomas Mayerhöfer,[15,16] Martin Moskovits,[34] Kei Murakoshi,[35] Jwa-Min Nam,[36] Shuming Nie,[37] Yukihiro Ozaki,[38] Isabel Pastoriza-Santos,[39] Jorge Perez-Juste,[39] Juergen Popp,[15,16] Annemarie Pucci,[23] Stephanie Reich,[40] Bin Ren,[32] George C. Schatz,[41] Timur Shegai,[25] Sebastian Schlücker,[42] Li-Lin Tay,[43] K. George Thomas,[44] Zhong-Qun Tian,[32] Richard P. Van Duyne,[41] Tuan Vo-Dinh,[45] Yue Wang,[46] Katherine A. Willets,[47] Chuanlai Xu,[28,29] Hongxing Xu,[48] Yikai Xu,[10] Yuko S. Yamamoto,[49] Bing Zhao,[50] and Luis M. Liz-Marzán[*,†,51]



The BIG Problem with SERS:

- SERS spectra are different than Raman spectra (in database)!
- SERS spectra are different on each SERS substrate!

Solution Proposed by others:
create a SERS database for each type of SERS substrate…(not realistic!)

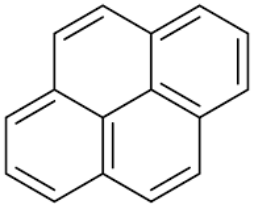
GOAL: Can we develop ML algorithms so we can use a Raman library (database)to identify chemicals through their SERS spectrum?

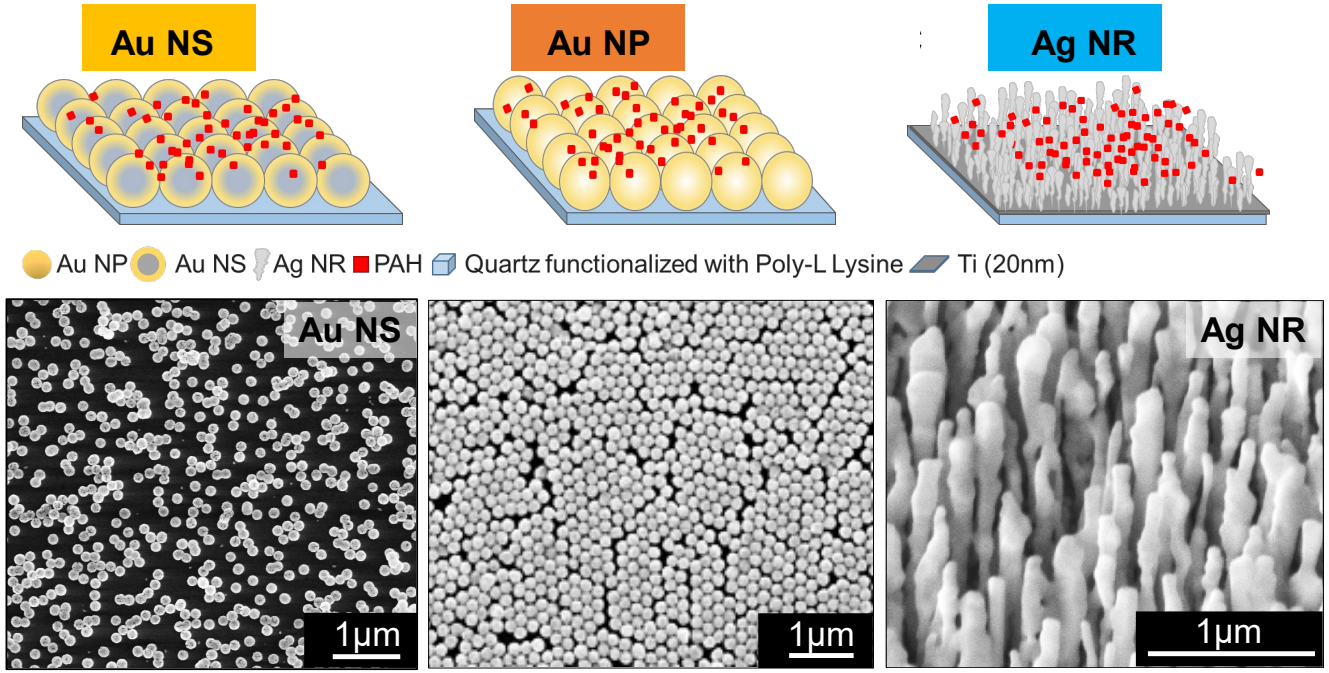
Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

Different SERS substrates produce different SERS spectra of the same molecule

PYRENE





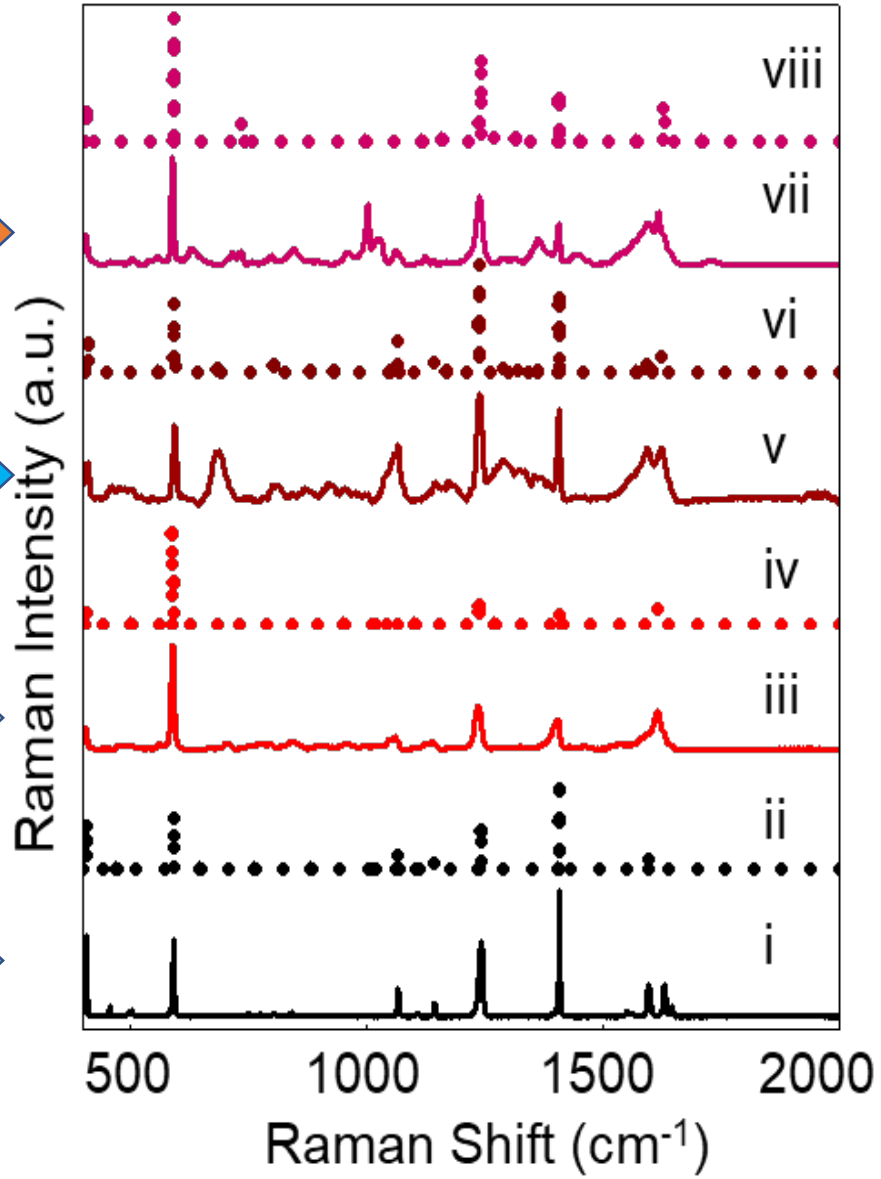
Au NP

Ag NR

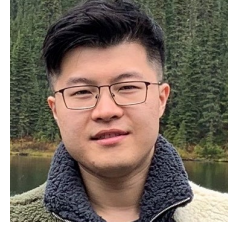
Au NS

Raman

Characteristic Peak Extractions = CaPE



ML-Based Spectral Recognition: CaPE and CaPSim



Yilong Ju

Step 1

For each molecule in the Raman library, identify potential locations (shown as grey shades) of characteristic peaks using CaPE

Step 2

Extract peaks from these learned locations by looking for maximum responses

Step 3

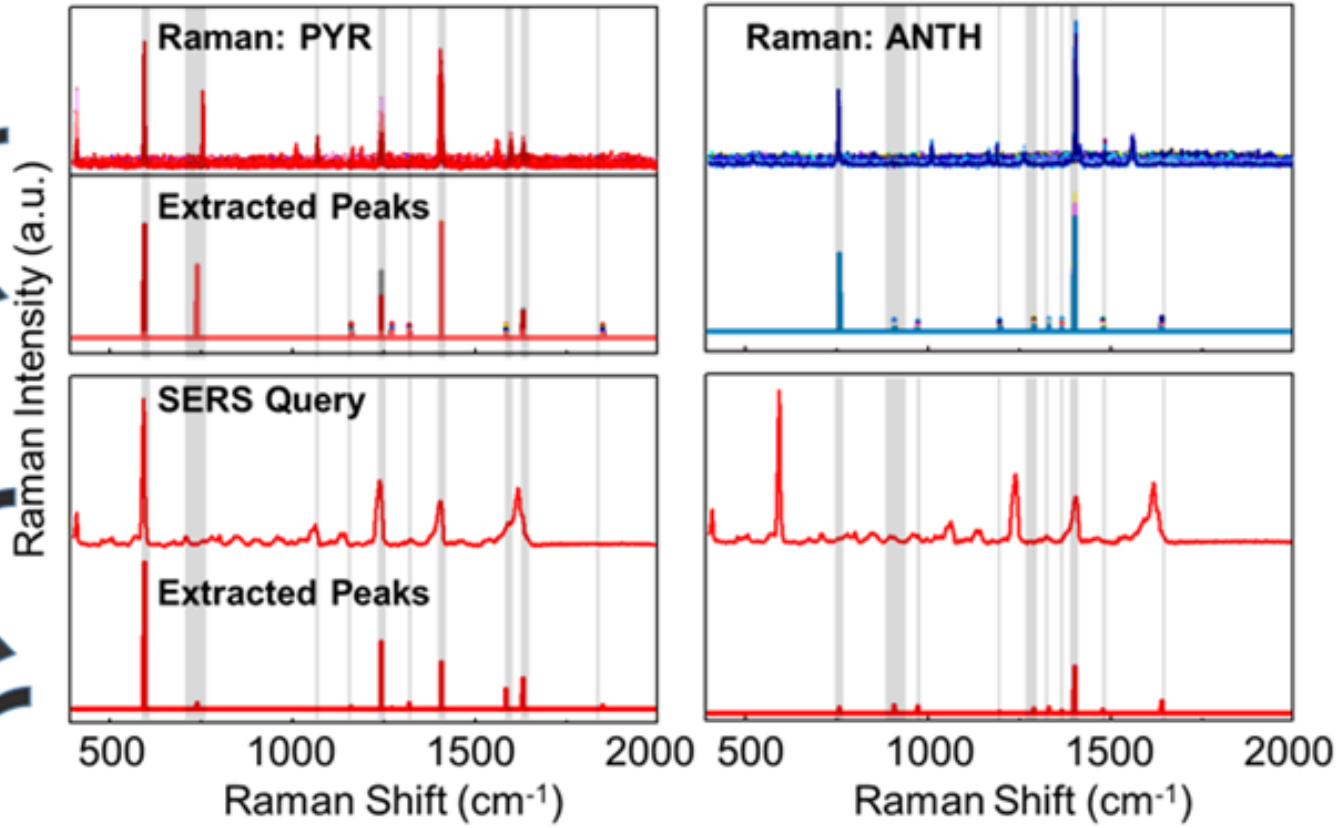
Extract peaks from the SAME locations from the query SERS spectrum

Step 4

Calculate the inner product between extracted peaks from the Raman reference and the SERS query, obtaining **CaPSim**

Step 5

Predict the query as the molecules with the highest similarity



Similarity = 0.242 **Similarity = 0.077**

Prediction: PYR

Machine Learning-enhanced SERS + SEIRA Detection of Polycyclic Aromatic Hydrocarbons in Human Placenta

Smoking during pregnancy is associated with increased risk of:

- miscarriage,
- prematurity,
- stillbirth,
- low birth weight,
- perinatal morbidity,
- sudden infant death syndrome (SIDS),
- adverse neurodevelopmental disorders

such as Attention-Deficit Hyperactivity (ADHD),

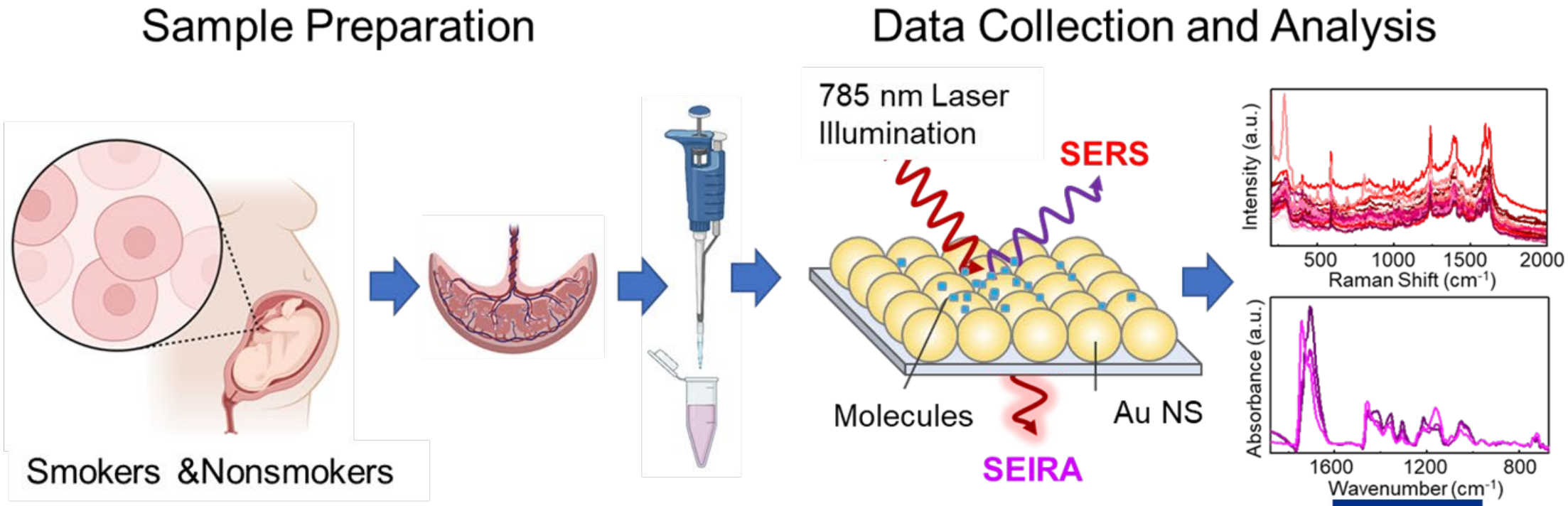
- anxiety, and depression

PAHs can cross the placenta from the mother to the fetus, exposing the fetus to these chemicals

Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

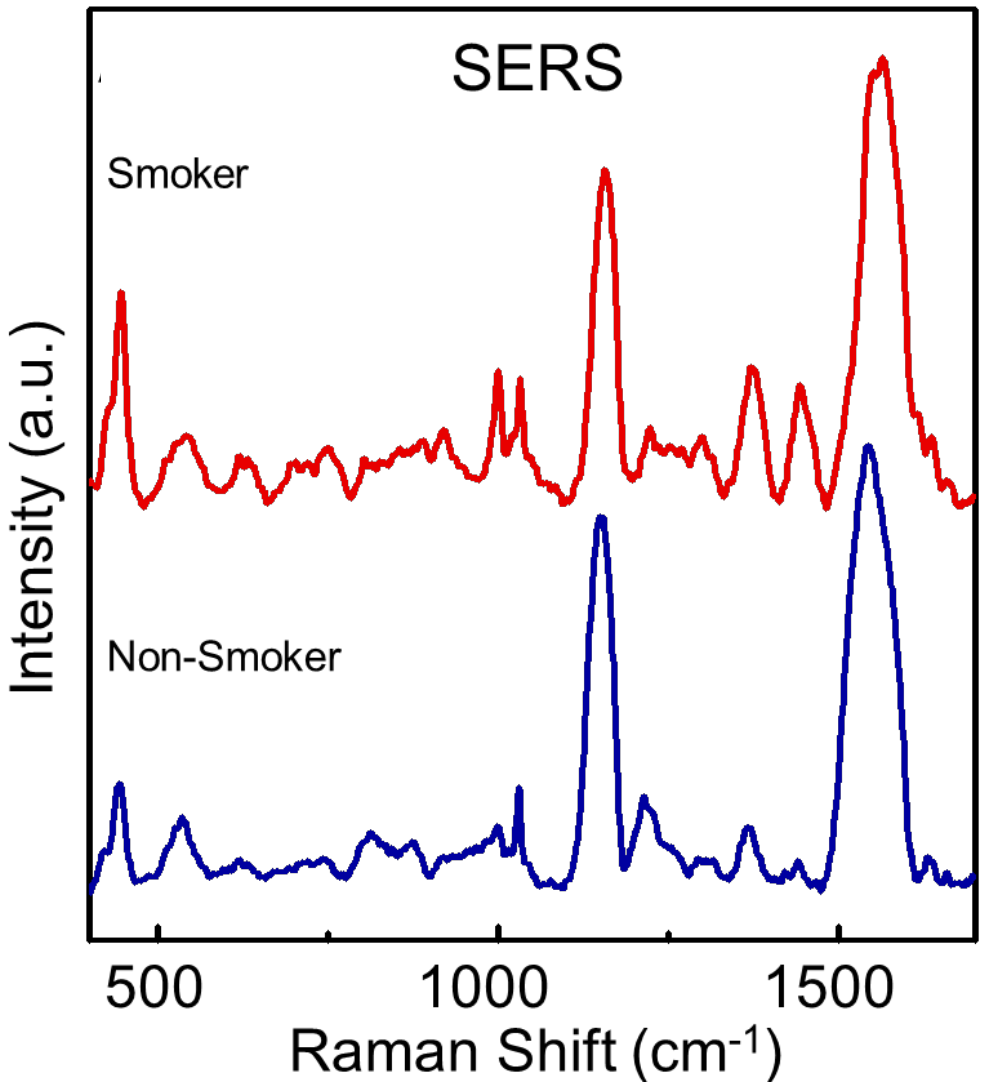
Machine Learning-enhanced SERS + SEIRA Detection of Polycyclic Aromatic Hydrocarbons in Human Placenta

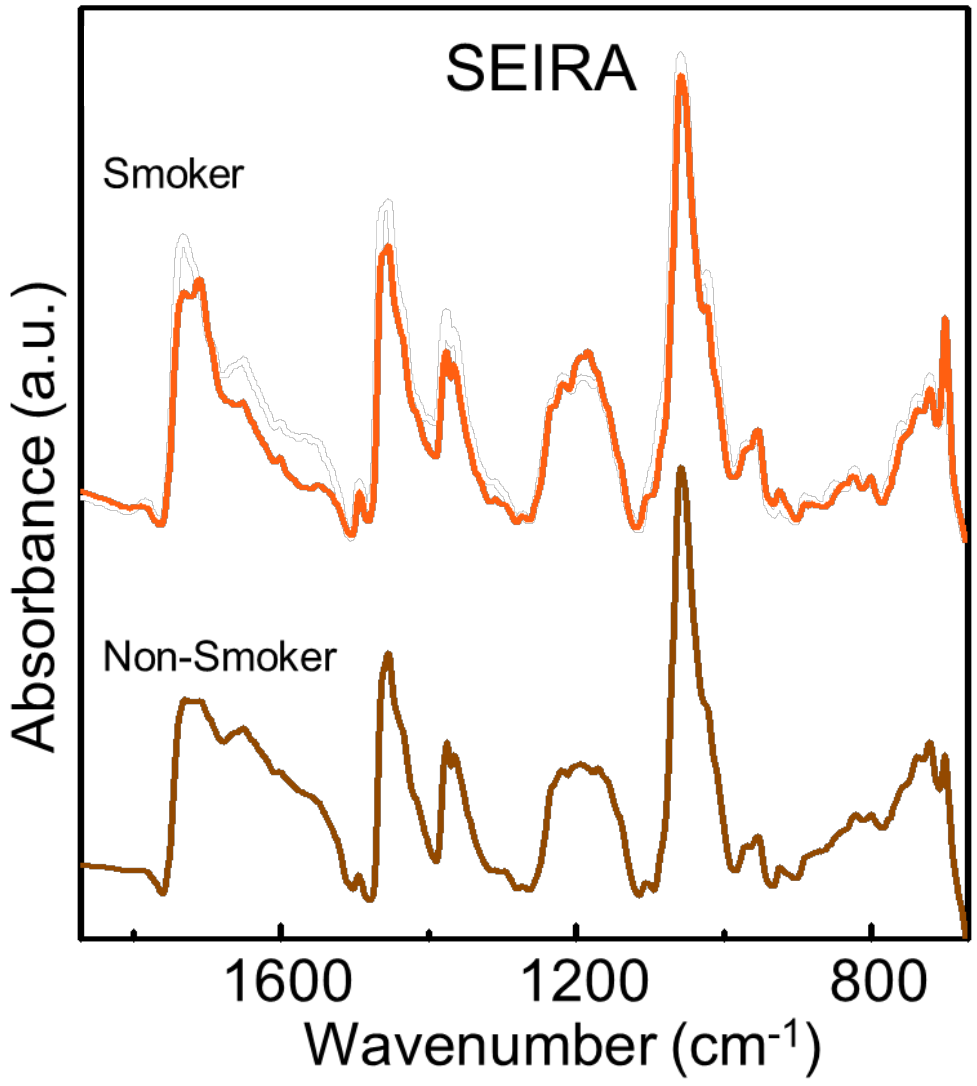


Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

SERS and SEIRA of Human Placenta

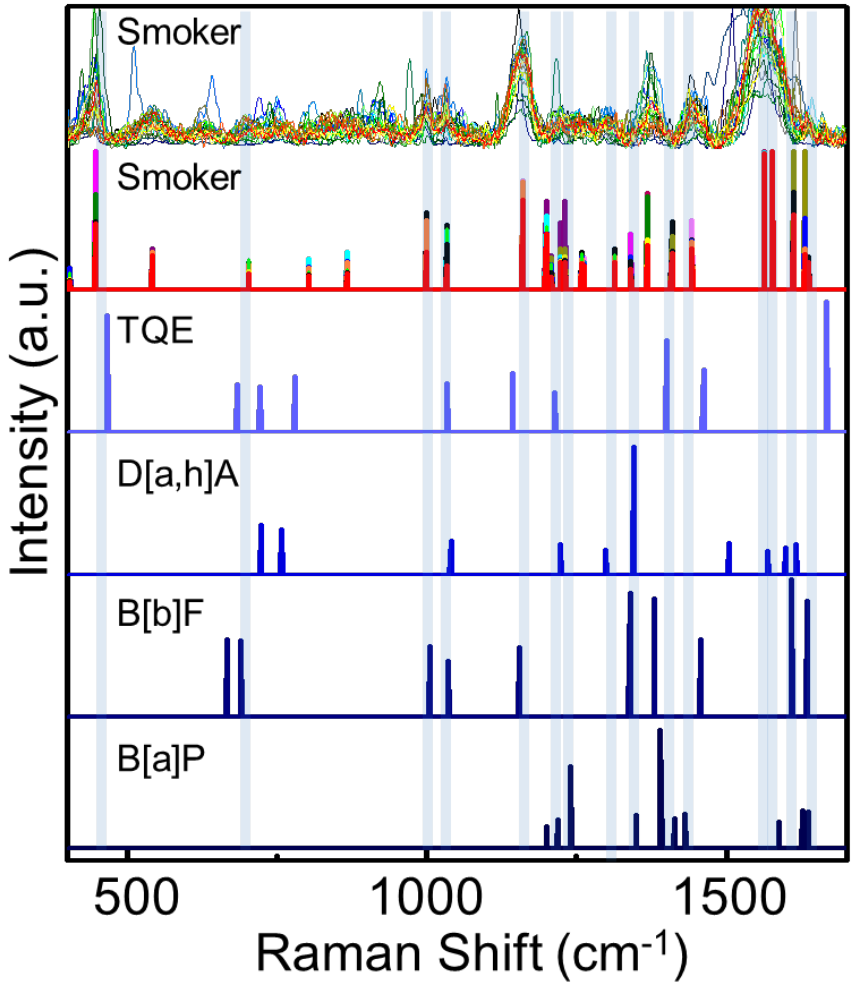


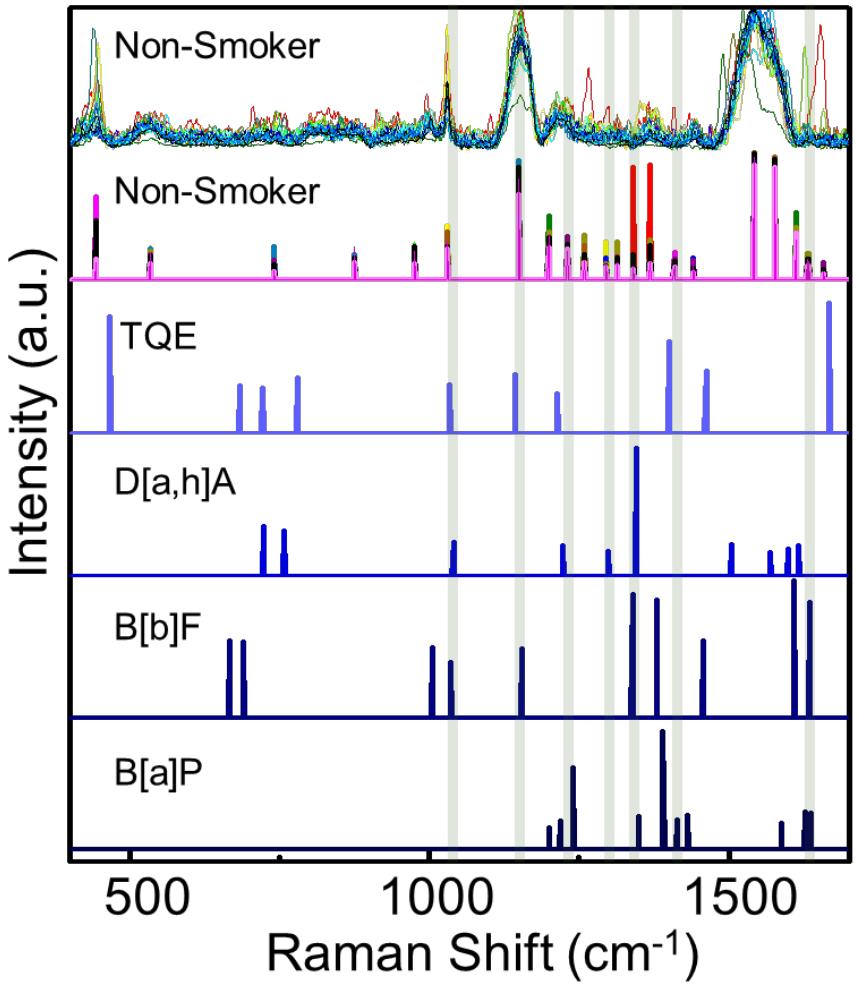


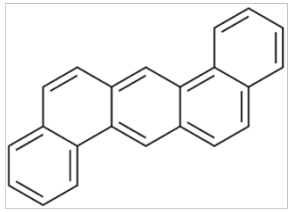
Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM

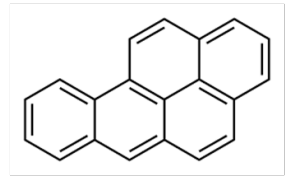
CaPE Analysis of Smoker vs. Non-Smoker SERS data:



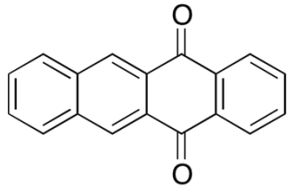




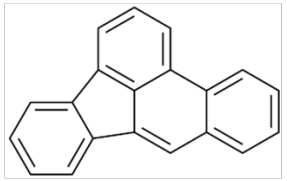
D[a,h]A



B[a]P



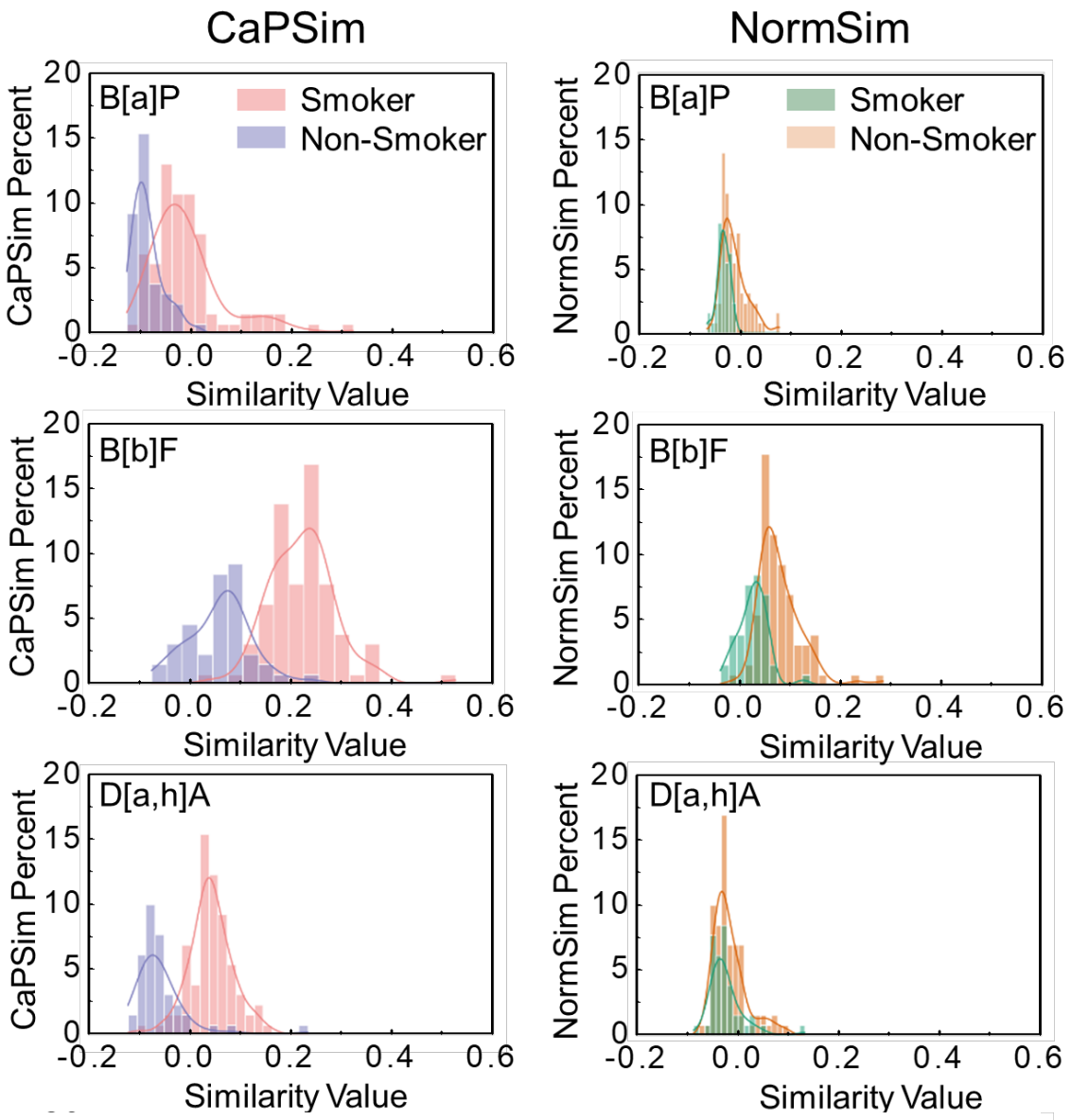
TQE



B[b]F

Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM


CaPSim
NormSim
B[a]P
Smoker
Non-Smoker
B[b]F
D[a,h]A
CaPSim Percent
NormSim Percent
Similarity Value
0
5
10
15
20
-0.2
0.0
0.2
0.4
0.6

CONCLUSIONS:

- ML can enable the detection of specific PAHs in complex mixtures without separations
- Characteristic Peak Extraction (CaPE) simplifies SERS spectra so a Raman spectral library can be used for identification with CaPSim
- CaPE and CaPSim can differentiate PAHs in the placenta of smokers versus non-smokers

Baylor College of Medicine

BCM-RICE SUPERFUND RESEARCH PROGRAM