PNAS RESEARCH ARTICLE | ENVIRONMENTAL SCIENCES

Computational chromatography: A machine learning strategy for demixing individual chemical components in complex mixtures

Mary M. Bajomo[a,b,1], Yilong Ju[c,1], Jingyi Zhou[b,d], Simina Elefterescu[e], Corbin Farr[a,b], Yiping Zhao[f], Oara Neumann[b,g], Peter Nordlander[b,g,h], Ankit Patel[c,g,i,1], and Naomi J. Halas[a,b,g,h,1,2]

Contributed by Naomi J. Halas; received July 3, 2022; accepted November 3, 2022; reviewed by Jennifer Dionne and Christopher J. Rozell

Surface-enhanced Raman spectroscopy (SERS) holds exceptional promise as a streamlined chemical detection strategy for biological and environmental contaminants compared with current laboratory methods. Priority pollutants such as polycyclic aromatic hydrocarbons (PAHs), detectable in water and soil worldwide and known to induce multiple adverse health effects upon human exposure, are typically found in multicomponent mixtures. By combining the molecular fingerprinting capabilities of SERS with the signal separation and detection capabilities of machine learning (ML), we examine whether individual PAHs can be identified through an analysis of the SERS spectra of multicomponent PAH mixtures. We have developed an unsupervised ML method we call Characteristic Peak Extraction, a dimensionality reduction algorithm that extracts characteristic SERS peaks based on counts of detected peaks of the mixture. By analyzing the SERS spectra of two-component and four-component PAH mixtures where the concentration ratios of the various components vary, this algorithm is able to extract the spectra of each unknown component in the mixture of unknowns, which is then subsequently identified against a SERS spectral library of PAHs. Combining the molecular fingerprinting capabilities of SERS with the signal separation and detection capabilities of ML, this effort is a step toward the computational demixing of unknown chemical components occurring in complex multicomponent mixtures.

surface-enhanced Raman scattering | polycyclic aromatic hydrocarbons | machine learning | nanoparticles | nonnegative matrix factorization

Despite its discovery nearly 50 y ago, surface-enhanced Raman spectroscopy (SERS) is still maturing toward a practical analytical technique for ultrasensitive chemical detection (1–3). Raman scattering, typically a very inefficient process, is enhanced by many orders of magnitude for molecules positioned in the direct vicinity of nanostructured metallic substrates, resulting in detailed SERS spectra that make detection and identification at low concentrations possible (4, 5). While SERS remains an active topic of research, its potential for high sensitivity, portability, and straightforward sample preparation could provide major advances in chemical detection/identification for biological or environmental samples over current methods that combine chromatography and mass spectrometry. However, given the multicomponent chemical complexity of environmental and biological samples (6), additional strategies are likely needed for this spectroscopic method to fulfill its technological promise.

A family of priority pollutants of great interest has been polycyclic aromatic hydrocarbons (PAHs), a hazardous class of chemicals whose molecular structure consists primarily of multiple fused benzene rings (7). PAHs are typically generated from incomplete combustion, frequently of fossil fuels, and are detectable in virtually every river and estuary worldwide (8). In biological systems, PAH metabolites bind covalently to cellular macromolecules, including DNA, and are well-known carcinogens (9, 10). In biological and environmental samples, they are typically found as multicomponent PAH mixtures and in complex matrices (11), which greatly complicates their detection and identification. Chemical methods that attempt to favor selective PAH detection on functionalized SERS substrates have been demonstrated (12–15), along with extraction protocols (16, 17), to reduce background effects due to complex matrices.

Given these challenges, the incorporation of machine learning (ML)-based strategies for digital separation or demixing together with SERS is a highly promising approach toward streamlined PAH detection and identification. Thus far, ML strategies have been combined with SERS to address problems such as the profiling of wine flavors and numerous biomedical applications (18–20). By combining the molecular fingerprinting capabilities of SERS with the signal separation and detection capabilities of ML, one can begin

Significance

Chemical contaminants are frequently found in mixtures of similar molecules; their identification typically starts with time-consuming separation steps prior to identification of individual components. Here, we examine whether chemical separations could be replaced by a machine learning-based analysis. Machine learning strategies have been developed to identify individual sources in a complex mixture of signals, known as the “cocktail party problem”, where a number of people are talking simultaneously, but the listener is trying to follow only one of the discussions. Here, we develop and apply this type of analysis to the spectroscopic signal of complex mixtures of chemicals to examine how well machine learning can distinguish the individual chemical components with no prior knowledge of their identity.

Reviewers: J.D., Stanford University; and C.J.R., Georgia Institute of Technology.

Competing interest statement: The authors declare a competing interest. The authors have patent filings to disclose, A provisional patent has been filed regarding work in this submission. Filing Date: October 31, 2022 Serial No: 63/421,094.

Copyright © 2022 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1] M.M.B., Y.J., A.P., and N.J.H. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: halas@rice.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2211406119/-/DCSupplemental.

Published December 19, 2022.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

to investigate whether individual PAHs can be identified by analyzing the SERS spectra of multicomponent PAH mixtures: an approach we have named "computational chromatography". It is worth mentioning that in this paper, we focus on unsupervised demixing (i.e., no library of known spectra is required). A library of known PAHs and mixtures might be used only for hyperparameter tuning and evaluating the demixing algorithms. However, even for these purposes, a library may be avoided, supposing we have some prior knowledge about the spectral characteristics of the components and use the performance on some downstream tasks for evaluation.

The demixing of SERS mixtures is an example of the blind source separation problem in ML, where measurement data are often modeled as an additive combination of underlying sources. A variety of methods have been designed to demix mixtures and recover the sources, among which independent component analysis (ICA) and nonnegative matrix factorization (NMF) are the most frequently used (21–24). Past attempts to demix spectra of mixtures typically have involved applying conventional ICA to a synthetic dataset (25), or to the SERS of a mixture containing only two components (26). For mixtures with more components, auxiliary algorithms have been designed to aid ICA, but the task performed was only to separate the background from the mixture (27). There are also many variants of ICA and NMF that might be very useful for demixing, as they introduce different assumptions and constraints to the problem, such as nonnegative ICA (NICA) (28, 29), sparse ICA (SICA) (30), and near-separable NMF (NSNMF). NSNMF methods, such as Extreme Ray (31) and Successive Projection Algorithm (32), are a bit different, since they directly pick the least mixed recordings from data as the estimated sources.

A key impediment to demixing is the presence of noise. Noise in the peak amplitudes and/or locations makes it difficult if not impossible to discriminate between two similar molecules. One effective strategy for dealing with this is to use a dimensionality reduction (i.e., compression) algorithm to filter out the less discriminating noncharacteristic peaks (NCPs). Such compression is especially important for NSNMF methods, because they search for extreme spectra, namely those that are most dissimilar from all other spectra in the dataset. Compression also enables demixing methods to run faster, an additional benefit. The most important information for identifying PAHs using SERS is their spectra, which consist of several prominent Raman-active spectral features, which we refer to as characteristic peaks (CPs): the background and noisy peaks are far less useful. For SERS of PAHs, roughly ten CPs can serve as a sufficiently discriminative fingerprint for the full molecular Raman spectrum, which often has many more peaks/dimensions. Hence, for a mixture of C components, we may only need roughly ~$10C$ dimensions. Examples of existing data compression algorithms designed for NSNMF include QR decomposition (33), structured random (34), and Count-Gauss (35). We call all peaks other than the CPs NCPs. None of the existing data compression algorithms or demixing methods are designed to extract and exploit the CPs, which becomes especially difficult for CPs with relatively low intensities. Moreover, these algorithms are not robust to local spectral shifts of resonant peaks, a frequently observed property in SERS spectra due to the varying interactions of molecules with SERS substrates.

Here we report a strategy that combines SERS and ML for the identification of individual components from the SERS spectra of a complex mixture of PAHs. We introduce Characteristic Peak Extraction (CaPE), a data compression algorithm that extracts CPs from SERS spectra based on counts at locations of detected peaks of the mixture. CaPE has two unique advantages over existing ML algorithms: 1) it estimates CP locations from a set of mixture SERS spectra containing any unknown components by selecting the spectral locations where peaks occur more frequently rather than just the specific locations of high-intensity peaks; and 2) it tolerates local frequency shifts of CPs, identifying peaks with small shifts across recordings as a single peak by means of a specialized clustering algorithm (see *Methods*). This combination of i) chemical sensing where SERS spectra are collected at different relative PAH concentrations and ii) demixing algorithms that can deal with small frequency shifts typically inherent in SERS spectra, enables the spectroscopic identification of individual PAHs from samples of a complex mixture in an unsupervised manner.

Results

Detection and Identification of PAHs Using SERS and ML. A schematic of the SERS substrate preparation and PAH detection is shown in Fig. 1. Au nanoshells (NSs) with a hydrodynamic diameter of 165 ± 5 nm were fabricated using a previously described method (36–38) (see *Methods*). Freshly prepared NSs were deposited onto poly-L-lysine-coated quartz substrates (Fig. 1*A*), followed by drop-dry deposition of PAH solutions in acetone onto the prepared substrates. SERS spectra of the PAHs were acquired using a Renishaw inVia Raman microscope with a 785-nm laser wavelength and a laser intensity of 55 μW. The NS were characterized by ultraviolet-visible-near infrared extinction spectroscopy while in aqueous solution (Fig. 1*B*) and scanning electron microscopy (SEM, Fig. 1*D*). The (black) experimental and (blue) theoretical extinction spectra of the aqueous NS solution show a strong dipole plasmon mode at 745 nm which corresponds to the 785-nm Raman pump laser (λ_{Ex} gray line). The strongest field enhancements for the NS aggregates are obtained at the junction between adjacent NSs: theoretical NS extinction spectra of various dimer configurations were simulated for dimers with a gap ranging from ± 4 nm (where negative gap distances refer to overlapping or fused NSs). The experimental extinction spectrum of the NSs appears to indicate the presence of both monomer and dimer plasmons in solution based on its spectral location between the calculated monomer and dimer plasmon spectral peaks. All monomer/dimer NS spectra span the Raman pump laser wavelength (785 nm) and the Stokes wavelength range. Spatial distributions of the calculated electromagnetic field enhancement for the monomer NS and for NS dimers with a ± 4-nm gap is shown in Fig. 1*C*. Although the maximum electromagnetic field enhancement occurs near the junction of dimers, there is still significant enhancement at the surface of the NS monomers. The SEM image in Fig. 1*D* shows both the size distribution and the morphology of the NSs. Three random areas in the SEM image are highlighted to represent different SERS collection areas. SERS spectra of PAH mixtures are shown in Fig. 1*E* in corresponding colors to illustrate the potential variation in SERS spectra from various collection areas on different substrates.

A schematic representation of how to extract information about the qualitative and quantitative contents of a multicomponent sample from its SERS spectra is shown in Fig. 1*E*. Given the spectra of a PAH mixture, ML methods can computationally demix the mixture and produce estimates of the underlying sources, as well as the mixing weight for each source. For the example illustrated here, the 1st mixture spectrum is a mixture of 0.8 of unit component 1 and 0.2 unit of component 2. Similarly, the other spectra can be demixed into various concentrations of component 1 and component 2.

ML-Based Demixing Algorithm. Formally, given an observation of a D -dimensional mixed spectrum x_i, we attempt to demix it as $x_i = \sum_{j=1}^{\hat{C}} w_{ij} s_j$, where $w_{ij} \in \mathbb{R}$ is the mixing weight of each

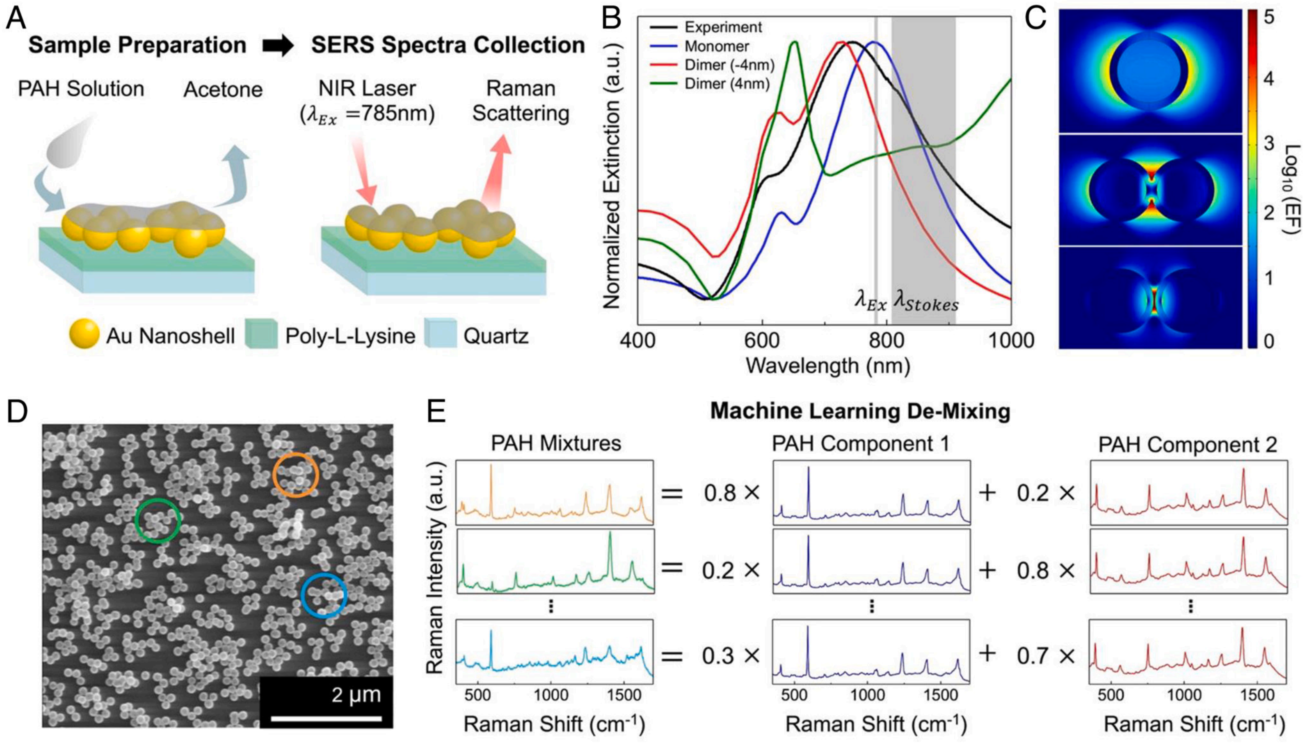

Fig. 1. Detection and identification of PAHs using SERS and ML. (*A*) Schematic depicting the preparation of the substrate and the collection of SERS spectra. (*B*) Experimental extinction spectrum of aqueous nanoshell suspension (black). Calculated extinction spectrum of nanoshell monomer (blue), a fused dimer with a -4-nm gap (red), and dimer with a 4-nm gap (green). (*C*) Spatial distribution of the calculated electric field enhancement in the x-y plane around nanoshell monomer (*Top*), dimer with a −4-nm gap (*Middle*), and dimer with a 4-nm gap (*Bottom*) at an incident wavelength of 785 nm. (*D*) SEM of SERS substrate. (*E*) Schematic of ML-based reconstruction of component PAH SERS spectra from the PAH mixture SERS spectra acquired from different areas of the SERS substrates represented by the colored areas highlighted in the SEM.

estimated source $s_j \in \mathbb{R}^D$ and $\widehat{C}$ is the number of sources. For a set of N observations $X = [x_1 | \cdots | x_N]$, this can be written as $X = SW$, where $X \in \mathbb{R}^{D \times N}$, $S \in \mathbb{R}^{D \times \widehat{C}}$, and $W \in \mathbb{R}^{\widehat{C} \times N}$. Due to the reasons we discussed above, for a better demixing result, we would like to first compress the input data to M dimensions, where $M \ll D$. Let $\tilde{X}$ denote the compressed spectra. The complete procedure of demixing includes: Part (1), obtain $\tilde{X}$ from X, and Part (2), solve $\tilde{X} = \tilde{S}\tilde{W}$, where $\tilde{X} \in \mathbb{R}^{M \times N}$, $\tilde{S} \in \mathbb{R}^{M \times \widehat{C}}$, and $\tilde{W} \in \mathbb{R}^{\widehat{C} \times N}$. We call any procedure designed to solve Part (1) a data compression algorithm and Part (2) a demixing method. We expect $\tilde{X}$ to contain only information about the CPs, which becomes trivial if we have access to clean, noiseless spectra X_{CP}. For example, $\tilde{X}$ could be as simple as all peak heights in X_{CP}. However, in practice, we typically have $X = X_{\mathrm{CP}} + z$,where z includes NCPs and background noise.

Four PAHs, Anthracene (ANTH), Pyrene (PYR), Benzo[a]pyrene (B[a]P), and Benz[a]anthracene (B[a]A), were selected from the U. S. Environmental Protection Agency's priority contaminants list (39) to produce different mixtures to test the capability of our ML-based demixing algorithm (40, 41). These PAHs were selected based on their environmental prevalence as well as their structural and spectral similarity. High-intensity peaks in the SERS spectra of each PAH were selected as ground truth peaks (GTP), on which the quality of the spectra produced by the demixing algorithm was evaluated. The demixing algorithms were first tested on the simplest multicomponent spectra: SERS spectra of a mixture of two PAHs. As shown in Fig. 2, SERS spectra of 1:1 mixtures of ANTH: PYR; ANTH: B[a]P; ANTH: B[a]A; PYR: B[a]A; PYR: B[a]P; and B[a]P: B[a]A were obtained. For each PAH mixture, 50 to 100 SERS spectra were collected from different areas of the substrate and with PAH mixtures specially prepared by varying their relative concentrations (*SI Appendix*, Figs. S1 and S2). This was done to provide the necessary variation between PAH SERS features needed for spectral separation and to meet the requirements of the demixing algorithms tested. Variation in the PAH SERS signals was created artificially in this manner, to show the capability of the SERS-ML demixing methodology. The demixing algorithms used all spectra available for each PAH mixture to produce spectra of the components of the mixture. We call these the demixed components (DCs). For each mixture, all of the strategies were able to accurately determine that the number of components $\widehat{N}_c = 2$ and produced spectra for each component. $\widehat{N}_c$ was selected from $\{2, 3, \ldots, 8\}$ and the selected number corresponded to the optimal objective value optimized by the demixing methods. An additional algorithm was employed to match the DCs to different PAHs based on spectral similarity (see *Methods*). The DCs produced from the best performing demixing algorithms, CaPE + NMF and CaPE + NICA, and the SERS spectra of the actual components of each PAH mixture are shown in Fig. 2. Most of the major GTP present in the PAH SERS spectra are also present in the corresponding DCs for each mixture, while most of the unimportant peaks or noisy peaks are ignored by CaPE, as we can see they are all set to 0 in the DCs. The exception is for mixtures containing B[a]P (Fig. 2 *D*–*F*). The demixing algorithm was only able to produce noise-free spectra for one component: B[a]P. The other component of each mixture, while containing all of the CPs of the respective PAH, also contained a significant amount of noise, which prevented a visual matching of the DC to the correct PAH component. However, it did not prevent accurate matching by the algorithm.

Two-Component Mixtures. The best demixing was obtained for the ANTH and PYR (Fig. 2*A*) and the ANTH and B[a]A

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

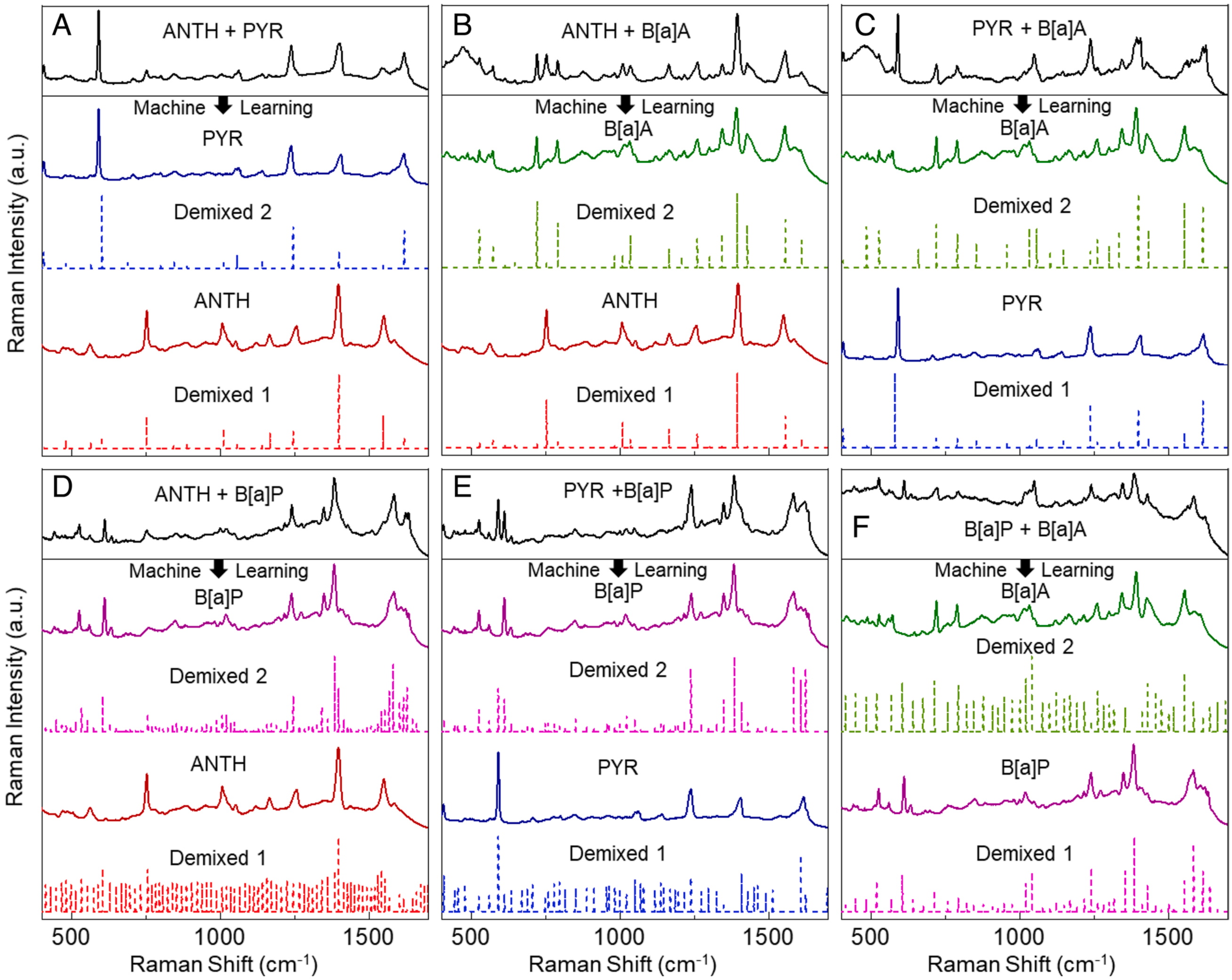

Fig. 2. Demixing of SERS spectra of mixtures with two PAHs. (*A–F*) SERS spectrum of the mixture of PAHs (solid black), SERS spectra of the components of the mixture (solid colored), and corresponding ML-based DC (dashed colored) for a mixture of (*A*) ANTH and PYR, (*B*) ANTH and B[a]A, (*C*) PYR and B[a]A, (*D*) ANTH and B[a]P, (*E*) PYR and B[a]P, and (*F*) B[a]P and B[a]A.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

(Fig. 2*B*) mixtures. All peaks present in each of the DCs matched the peaks of the corresponding PAH SERS spectra well, in both location and relative intensity. The demixing algorithm performed well in correctly attributing close peaks corresponding to the different PAHs. For the ANTH and B[a]A mixture (Fig. 2*B*), several minor features in the B[a]A SERS spectra are not present in the corresponding DC (Demixed-2). However, the majority of the most intense SERS peaks could be directly attributed to the B[a]A modes appearing in the corresponding DC. The demixing of PYR and B[a]A (Fig. 2*C*) also produces DCs that match the corresponding PAH SERS spectra well but with minor errors. The DC for PYR (Demixed-1) contains a few features with relatively low intensities corresponding to B[a]A modes at ~1,260, 1,433, and 1,554 cm^{-1}. Additionally, the DC for B[a]A (Demixed-2) contains features at ~1,616, 1,237, 1,102, 956, 853, and 659 cm^{-1} that are either too intense or are incorrectly attributed to B[a]A. None of these errors prevent the DCs from being easily matched visually or computationally to the correct PAH. In contrast, the DCs from the mixture of B[a]P and ANTH (Fig. 2*D*), B[a]P and PYR (Fig. 2*E*), and B[a]P and B[a]A (Fig. 2*F*) are not as easy to visually match as the others. The peaks in DCs produced for these mixtures are much less sparse than for the other mixtures previously discussed. The DCs corresponding to B[a]P in Fig. 2 *D–F* match the B[a]P SERS spectrum. However, instead of the DCs containing only one peak that corresponds to each B[a]P SERS feature like the other DCs previously discussed, they contain several peaks with different intensities clustered together that match the overall B[a]P SERS peak shapes. There is also the presence of some incorrectly attributed peaks in each DC. For the mixture of B[a]P and ANTH (Fig. 2*D*), the DC matched to B[a]P contains a feature at ~1,398 cm^{-1} that corresponds only to ANTH. Likewise, for the mixture of B[a]P and PYR (Fig. 2*E*), there are features in the DC matched to B[a]P at ~1,408 and 590 cm^{-1} that are, respectively, too intense and correspond only to PYR. For the mixture of B[a]P and B[a]A (Fig. 2*F*), the DC matched to B[a]P contains features at ~1,554, 1,430, 1,041, and 731 cm^{-1} that are either too intense or correspond only to B[a]A. The DCs corresponding to the other PAHs for the mixtures in Fig. 2 *D–F* contain a significant amount of noise. The noise is present at the same intensity as for the relevant peaks, making it difficult to visually distinguish these peaks from noise. The only exception is the DC corresponding to PYR in Fig. 2*E*. The characteristic PYR SERS peaks at ~1,608, 1,408, 1,238, 590, and 407 are present in the corresponding DC at a slightly higher intensity than the noise. Overall, the presence of noise does not prevent these DCs from being matched to the correct PAH algorithmically.

Understanding ML-Based Demixing Algorithms. The ML algorithm used to identify the PAH mixture components is illustrated in Fig. 3. Instead of visualizing the full spectrum of PYR and B[a]P, for simplicity here we only focus on the intensities of two frequencies, 589 cm^{-1} and 1,382 cm^{-1}, which are the spectral locations of the highest amplitude peaks of PYR and B[a]P, respectively. Thus, each spectrum is reduced from a 1,738-dimensional vector to a 2-dimensional vector. A Gaussian pulse was used to broaden each peak for visualization purposes. The calculated spectra of mixtures of two PAHs with different concentration ratios (CRs) are presented in Fig. 3*A*, and we show mixtures with different absolute concentrations in Fig. 3*B*. The pure components (shown as solid arrows) serve as the extreme vectors of a cone that contains all possible mixtures. Mixtures with higher absolute concentrations are further from the origin. Also, mixtures with the same CR lie on a ray starting from the origin. The examples from Fig. 3*A* are labeled as stars. A comparison between the DCs estimated by NMF and the pure components is shown in Fig. 3*C*. We observe some errors in the DCs (also shown as dashed arrows in Fig. 3*B*): DC 1 has a greater than expected x coordinate and DC 2 has a greater than expected y coordinate. When projected back to the full spectra, these errors become spurious peaks or peaks with incorrect relative intensities. This illustrates that the problem will become more difficult when the extreme vectors span a much smaller space as shown in Fig. 3 *D–F*. The same algorithm can only separate one of the components while missing the other. Also, in practice, there are more than two peaks in the spectra, making identifying extreme vectors much more difficult for the ML demixing.

More than Two Components in a Mixture. The demixing strategies were tested on more complex multicomponent spectra (Fig. 4): SERS spectra of a mixture of the four PAHs. SERS spectra of mixtures of ANTH, PYR, B[a]P, and B[a]A in various ratios were collected (Fig. 4*A*). The relative ratios of PAHs used in demixing the spectra of four PAHs were similar to the ratios used for demixing two PAHs. They both included spectra of the PAHs mixed equally and spectra with each PAH at a higher

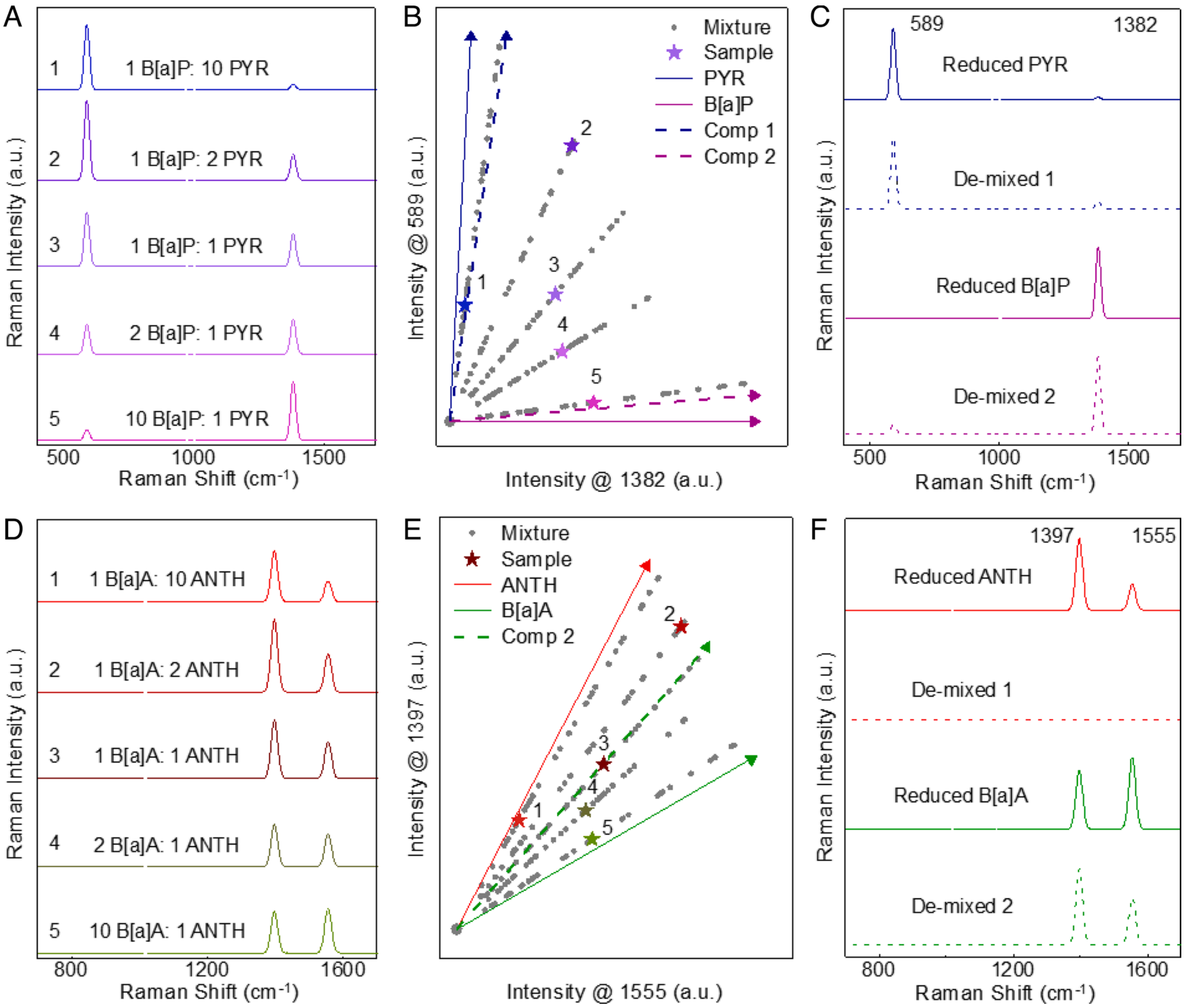

Fig. 3. Visualization of PAH spectra in the feature space. (*A*) Synthetic mixture spectra of B[a]P: PYR with ratios of: 1) 1:10, 2) 1:2, 3) 1:1, 4) 2:1, and 5) 10:1. We only focus the most important peaks of PYR and B[a]P, which are at 589 and 1,382 cm^{-1}, respectively, while setting elsewhere to 0. This reduces each spectrum to a 2D vector. (*B*) Synthetic mixture of PYR and B[a]P in the feature space with intensities focused at 1397 and 1555 cm^{-1}. Arrows show the direction of pure components. Each gray dot corresponds to a mixture spectrum. Each star corresponds to an example shown in *A*. (*C*) Spectra of mixture components reduced PYR, DC 1, reduced B[a]P, and DC 2. (*D*) Synthetic mixture spectra of B[a]A: ANTH with five ratios. (*E*) Synthetic mixture of ANTH and B[a]A in the feature space with intensities focused at 589 and 1,382 cm^{-1}. (*F*) Spectra of mixture components reduced ANTH, DC 1, reduced B[a]A, and DC 2.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

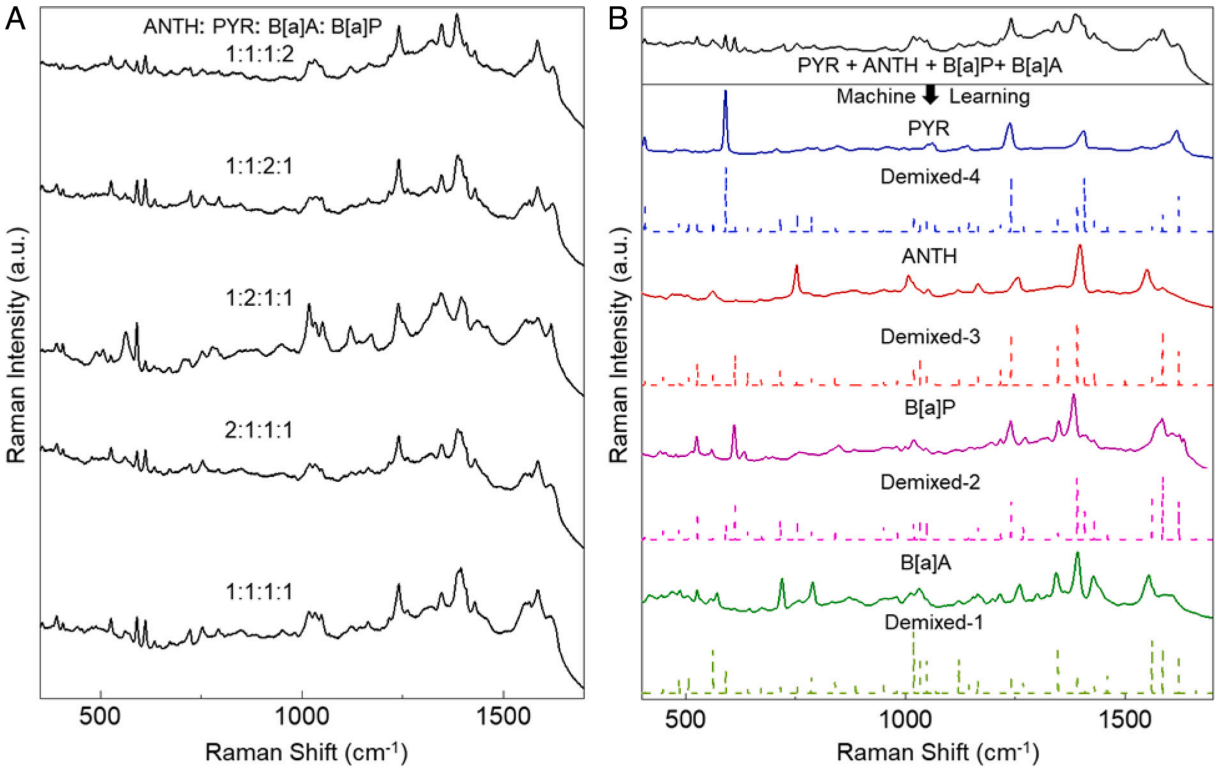

Fig. 4. NMF demixing of Mixture SERS spectra with four PAHs. (*A*) SERS spectra of ANTH, PYR, B[a]A, and B[a]P in different mixture ratios. (*B*) SERS spectrum of a mixture of PAHs (solid black), SERS spectra of the components of the mixture (solid colored), and corresponding derived components (DCs) (dashed colored) for a mixture of ANTH, PYR, B[a]A, and B[a]P.

concentration than the other(s). All mixture SERS spectra contain features from the individual PAHs. However, there is significant overlap in the major peaks from each of the different PAHs in the 1,300 cm^{-1} to 1,500 cm^{-1} range, making the separation and identification more challenging.

The DCs produced from the demixing algorithm and the SERS spectra of the components of the PAH mixture are shown in Fig. 4*B*. Unlike the demixing results from two PAHs, the demixing of four PAHs resulted in DCs more unlike the PAH SERS spectra. This is also reflected in the quantitative assessment of the demixing. The best result is the DC corresponding to PYR (Demixed-4). The five most intense peaks match the most intense peaks in the PYR SERS spectrum extremely well. However, there are a few lower intensity peaks present in the DC that do not match PYR. The Demixed-2 spectrum that corresponds to B[a]P also has a similar result with most of the major peaks present with some low-intensity noise. There is also the absence of a distinguishing B[a]P feature at ~1,350 cm^{-1} in the Demixed-2 spectrum. The Demixed-3 and Demixed-1 spectra, corresponding to ANTH and B[a]A respectively, do not match their respective SERS spectra compared with the other DCs. There are also some misattributed or noisy peaks with high intensities and the missing CPs. Despite these errors, the simple matching algorithm is still able to match the DCs to the correct PAHs. Also, CaPE successfully picks up the CP locations while ignoring most of the unimportant and background peaks.

Comparing Demixing Algorithms. The performance of different demixing methods with or without using the CaPE algorithm is shown in Fig. 5. Fig. 5*A* shows the area under the precision-recall curve (AUPRC) for known mixtures, which demonstrates the best possible performance for each algorithm, while Fig. 5*B* shows if the DCs match the PAHs for unknown mixtures, demonstrating the generalization performance. The AUPRC measures how well the DCs reconstruct the matched PAHs in terms of the recovery of CPs. We use a similarity metric close to the cosine similarity for the matching process (see *Methods*). A perfect recovery of the underlying PAH will lead to an AUPRC close to one. We also consider existing data compression algorithms for NSNMF, including QR decomposition, structured random compression, and Count-Gauss. Using CaPE, the AUPRCs for all mixtures are improved by a large margin, especially for the more difficult ones, like B[a]P + ANTH and the four mixtures, where the AUPRCs are relatively low. These results indicate that CaPE is able to extract CPs effectively. Each spectrum processed by CaPE has only 18 to 106 dimensions, which is much lower than the original 1,738 dimensions of the original acquired spectra. These lower dimensional representations of mixture spectra also make it much easier to identify which PAH each DC is, as shown in Fig. 5*B*. We plot the performance of matching DCs to PAHs averaged over multiple tests. In each test, we leave one mixture unseen (i.e., unknown) and use the rest to tune the hyperparameters. The task in the right panel is more difficult, since the test mixtures only contain unseen PAH components. CaPE-rank and CaPE-threshold are two variants of CaPE (see *Methods*). We calculate the proportion of correctly matched PAHs by matching the DCs to a small library of eight PAHs using a similarity metric. Four of the PAHs are not present in any mixtures. Thus, if a demixing method is not performing well, it may miss all the PAHs. Without CaPE, existing demixing methods can match half of the correct PAHs at most. For NSNMF, using existing data compression algorithms reduces the performance. However, CaPE enables the demixing methods to recover many more components correctly. NICA + CaPE is almost able to match all the PAH components stably in both of the test settings in (B).

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

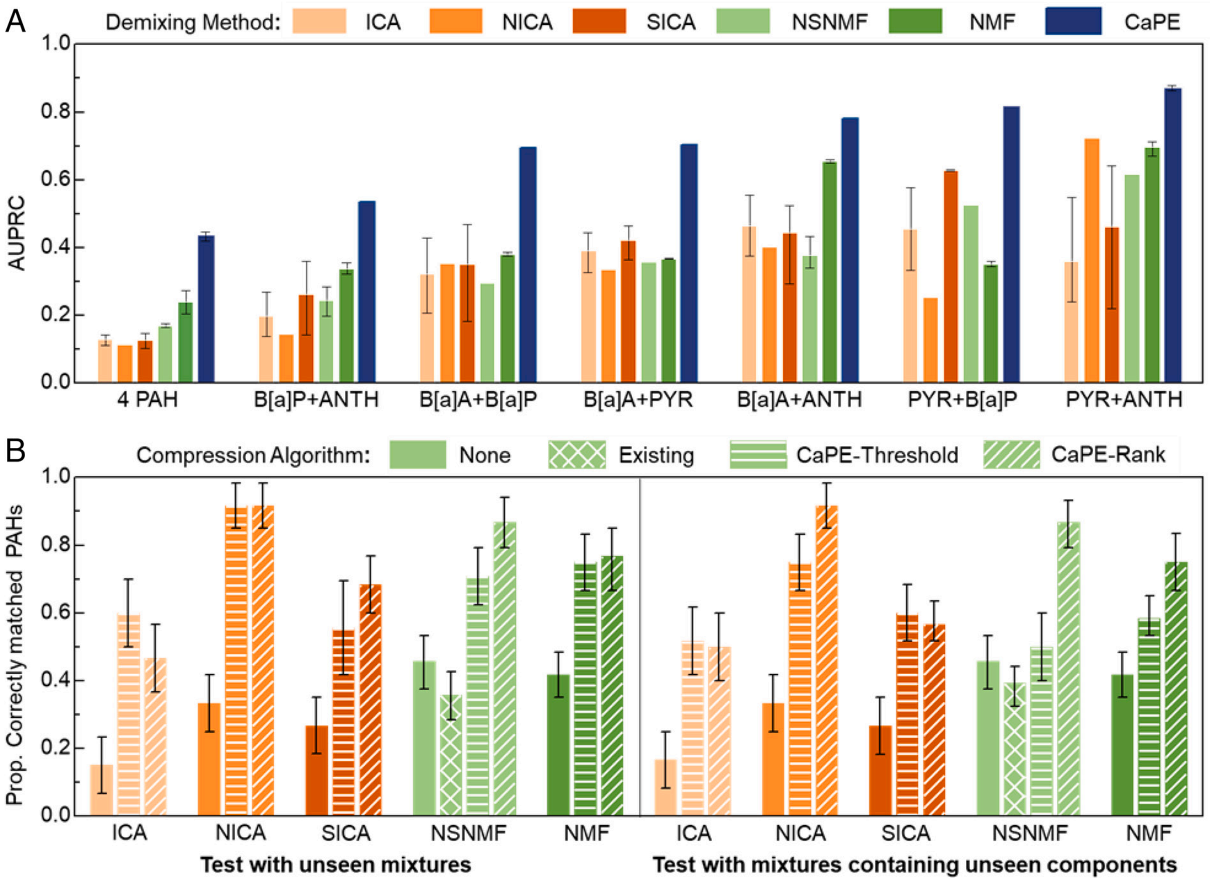

Fig. 5. Performance comparison between demixing methods with or without CaPE. (*A*) AUPRC compared between existing demixing methods with or without using CaPE on different mixtures. The bars of CaPE show the results of the best combination of an existing demixing method and CaPE. The bars of NSNMF show the better results between using or not using existing data compression algorithms. (*B*) Proportion of correctly matched PAHs after demixing, averaged over multiple tests. In each test, we leave one mixture unseen and use the rest to tune the hyperparameters. *Left*: the left-out mixture may contain a PAH already included in other mixtures. *Right*: the left-out mixture does not such contain a PAH component. All error bars show the 95% CI across five runs.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

Discussion and Conclusions

Past work attempting to demix PAH mixtures required curated libraries like Raman spectra, but such curated libraries for SERS do not yet exist, and if they are generated they may be incomplete, highly substrate-dependent, or possess procedure-dependent artifacts. In addition, the variability of SERS measurements in different spatial regions of the substrate, conventionally considered a nuisance, is in fact, from an ML or information theoretic point of view, a desired feature because varying CRs of components provides more information useful for the demixing algorithm (Fig. 3*B*). This feature is particularly essential for unsupervised demixing without the use of libraries. Also, there is a lack of attention to frequency shifts of SERS peaks due to variations in molecular orientation and binding affinity to SERS substrates, a characteristic property of SERS. This might worsen the performance of widely used ML methods, like ICA, NMF, and their variants. We address these past limitations by developing a computational-sensing-based technique for demixing mixtures that does not require any knowledge of the underlying mixture components. It employs a codesign of chemical sensing that measures SERS samples at various points on a substrate and a demixing strategy that can deal with frequency shifts and low-intensity CPs in SERS spectra. In addition, the key of the strategy is CaPE.

CaPE uses a count-based criterion because 1) some components may have relatively low concentrations in the mixture, and hence the intensities of all their peaks are low, and 2) some NCPs or noise may have the same level of intensity as some low-intensity CPs. By counting the number of peak occurrences at a particular location (wavenumber) across all recordings, we can find hotspots where CPs are likely to be located. Ideally, the count for every CP should be close to the total number of recordings, whereas the counts for NCPs should tend to be much lower since their locations may be shifted over the entire Stokes spectral region. CaPE also has a spatial maximum pooling (42) operation, commonly used in the architecture of convolutional neural networks to enable invariance to small local shifts of objects in the input image, a critical part of successful computer vision algorithms (40, 43).

Based on our quantitative evaluations, CaPE offers a great value for the problem of SERS demixing: more CPs will be assigned to the correct DC and more DCs will be matched to the correct PAHs. In addition, CaPE also compresses the data and relieves the constraints on time or space complexity when choosing demixing methods. CaPE is necessary for achieving the best possible demixing performance, as shown in Fig. 5*A*, where we tune the hyperparameters in the demixing method and data compression algorithm jointly according to the average performance (details in *Methods*). Also, CaPE is not only effective in a single mixture, but it also works for other unknown mixtures, no matter which demixing method it is combined with. This was shown in Fig. 5*B*, which tests whether the algorithm can generalize by leaving one mixture unseen and only using the rest to tune the hyperparameters. Despite these gains, CaPE only has three hyperparameters, for each of which are searched over three or four values. Hence, the tuning effort required is small. Also, although we use a library of known mixtures during tuning, in practice, we may not need a library if we have some prior knowledge about the spectra of the potential components. In addition, for evaluation, since we currently match the DCs to a library of known chemicals, new chemicals can be discovered and added to the library if they do not match any existing ones with high confidence. In future work, we could also avoid using a library by evaluating the algorithms on some downstream tasks. Another promising direction is to explore how CaPE-enhanced unsupervised demixing compares with nonblind or semiblind demixing methods that use libraries or dictionaries (44, 45). It is possible to obtain good demixing results, but a large dictionary may increase the running time of these already time-consuming algorithms. Besides, if the mixture contains a component that is not in the dictionary, we still need to learn and add it to the dictionary online, which is similar to what an unsupervised approach does.

Despite the strengths, there are still opportunities for improvement. For example, from Fig. 2 *D–F*, we see that CaPE still picked some NCP locations, which indicates that our assumption about the more spread-out distribution of NCPs and noisy peaks might be violated in some cases. This peak range selection may also confuse the demixing method, since if NCPs and noisy peaks are included, the similarity between different components will increase. Hence it becomes more difficult in such cases for demixing methods to separate the mixture. In future work, it is worth exploring how we can refine the criterion of peak selection in CaPE to improve it. Furthermore, we can explore ways to estimate some of the hyperparameters directly from data, making the tuning step less needed. For example, we may be able to estimate the distance threshold for clustering peak counts by just gauging from its distribution. In this way, we may also use different thresholds for different CPs, making the algorithm more flexible.

We must acknowledge that any demixing algorithm has fundamental limitations. For example, if two components have the same spectral peaks then demixing is impossible, as this violates the requirement (the source matrix being full-rank) for the identifiability of NMF (46). In these cases (e.g., in Fig. 2 when the mixture contains B[a]P which shares many CPs with other PAHs), the bottleneck of the demixing performance is not CaPE but the demixing algorithm. Nevertheless, CaPE is still able to improve the performance to some degree given the uncertainty/ambiguity caused by overlapping peaks and noise. To summarize, the SERS-ML tandem methodology will open the door for rapid diagnostic, fieldable identification, and detection of at-risk chemicals based on their molecular structure.

Methods

Materials. (3-aminopropyl) triethoxysilane (ES, 99%), tetrachloroauric acid ($HAuCl_4 \cdot 3H_2O$), tetrakis hydroxymethyl phosphonium chloride, poly-L-lysine hydrobromide (MW 150,000−300,000) poly-L-Lysine (PLL) ANTH, PYR, B[a]P, and B[a]A were purchased from Sigma-Aldrich. Formaldehyde (37%), sulfuric acid (H_2SO_4, 100%), hydrogen peroxide (H_2O_2, 30%), potassium dihydrogen phosphate (KH_2PO_4), and 200-proof ethanol were obtained from Fisher Scientific. All chemicals were used as received without further purification. Quartz slides were purchased from Fisher Scientific.

Sensing. SERS measurements were acquired with a Renishaw inVia Raman microscope (Renishaw) with 785-nm excitation wavelength and 55-μW laser power at the samples. Backscattered light was collected using a 63× water immersion objective lens (Leica) with a 20-s exposure time. The extinction measurements were performed on a Cary 5000 UV/Vis/NIR Varian spectrophotometer. SEM measurements were performed using a FEI Quanta 400 field emission SEM at an acceleration voltage of 20-kV scanning electron microscope. The SEM samples were prepared by evaporating a droplet of aqueous NS solution onto a silicon wafer.

SERS Substrate Preparation. Cleaned quartz slides were modified with 0.01% w/v aqueous solution of PLL (MW 150,000 to 300,000) for 20 min to facilitate the attachment of a dispersed monolayer of NSs on the quartz surface. The quartz substrates were cleaned by immersing in "piranha solution" (H_2SO_4:H_2O_2 = 3:1), followed by rinsing with deionized water (18.3 MΩ, Millipore). Au NSs were fabricated using a method previously described. NS of inner and outer radii [r_1, r_2] = [63, 86] with a strong dipole plasmon mode at 780 nm were used for the SERS studies. NS were immobilized by depositing 100 μL of the Au NS suspension on PLL-coated quartz substrate with isolated wells (9-mm diameter) for a minimum of 6 h. The quartz slides coated with Au NS film were rinsed with water and acetone followed by incubation with 10 μL of 100 μM PAH solution. Before acquiring the SERS spectra, the substrates were fully immersed in Milli-Q water.

Calculation of the Optical Properties of the SERS Substrate. This was performed using COMSOL Multiphysics software. The nanoshell was modeled as a silica core of radius 60 nm coated with 14-nm Au layer on a quartz substrate. The junction of the fused dimer was smoothed to a curve of radius 3 nm. The dielectric constant of Au was obtained from Johnson & Christy (47). The refractive index of the silica core and the substrate was 1.5. The medium refractive index was 1.33 for NSs dispersed in the aqueous solution. The electric field enhancement was calculated as $|E_{ex}/E_0|^2 \times |E_{stokes}/E_0|^2$, where the stokes shift was 350 cm^{-1}. The dimers simulated for field enhancement were filled with PAH inside the junction and under longitudinal polarized light. The PAH refractive index was taken to be 1.49 (48).

Computational Methods. This section includes the details of the preprocessing method we use, as well as the notation and the full procedure of the CaPE algorithm.

Baseline Removal. The existence of a common background in the spectra for different PAHs will increase the correlation between PAHs and hence make demixing more difficult. Hence, people typically use a baseline removal algorithm (49) to remove this overall trend in the spectra. This step does not affect any spectral peaks. We also apply this step to our SERS data before all analyses.

CaPE. As visualized in Fig. 3, we can simplify spectra of PAH mixtures to better understand how the ML demixing works–we reduce the dimensionality of the spectra by only observing the most CPs in each PAH component (see *SI Appendix*, Tables S1–S4 for the CPs). This spectra compression step reduces the similarity between spectra caused by noisy NCPs. *SI Appendix*, Fig. S4 shows how the mixture spectra distribute when one of the picked peaks is an NCP. In the extreme case, all samples lie on a straight line no matter the CR, and hence, the components become unidentifiable. Therefore, we hypothesize that we can improve the identifiability by picking CPs and compressing the spectra before inputting the data into ML demixing methods.

However, when given a mixture whose components are unknown, it becomes challenging to find CPs since 1) we cannot refer to a library of spectra for the CP locations; 2) some components may have relatively low concentrations in the mixture, and hence the intensity of their CPs is low as well; 3) some NCPs may have the same level of intensity as some low-intensity CPs. We may need a nontrivial peak detector to distinguish between CPs and NCPs. Also, the compression algorithms proposed for NSNMF in the existing work do not have an interpretation related to the SERS demixing task. Thus, in this paper, we proposed a simple yet effective algorithm able to reduce mixture spectra to a lower dimension and keep most of the important information.

Notation. Let $X \in \mathbb{R}^{d\times N}$ denote the input mixture spectra, where N is the number of recordings and d is the dimension. In our setting, $d = 1,738$ and N ranges from 60 to 120. We use $x \in X$ to denote a single recording from X. Let $L_x \in \mathbb{R}^d$ be a binary vector whose i th element $L_{x,i} = 1$ if there is a peak at x_i else 0. The CaPE algorithm contains two stages: 1) estimate a range of locations for each CP. CPs from all components in a mixture are considered. 2) reduce the mixture spectra to a lower dimension by applying max pooling over every estimated range of CP locations. Max pooling is an operation that selects only the maximum intensity over a given range; all others are discarded. The resulting vector will contain the intensities of the maximal CPs.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

Estimating Ranges of CP Locations: CaPE-Rank. Step 0–Smoothing: we smooth X by applying a smoothing kernel to each x. We use a moving average kernel with kernel size K for all experiments. We also tried a Gaussian kernel, but our preliminary results show that it is not as helpful as a moving average kernel. Step 1–Peak Detection: we detect peaks for each x and obtain L_x. We use a trivial peak detector whose only criterion is a minimum prominence of 0.02. Prominence is the vertical distance between a peak and its lowest contour line. Each x is normalized to have an intensity range of [0, 1]. Thus, this small prominence threshold would suffice to detect any reasonably sized peaks. Step 2–Count Peaks: Calculate L_X, the count of all detected peaks for X, defined as $L_X = \sum_{x \in X} L_x$. Step 3–Select Peaks: Select the top $M' = M$ candidate peaks in terms of peak counts given in L_X. Suppose the indices for these M' peaks are $i_1, \ldots, i_{M'}$. Then, we obtain $L_{X,M'}$, where $L_{X,M',i} = L_{X,i}$ if $i \in \{i_1, \ldots, i_{M'}\}$, else 0. After this step, we typically see that some selected peaks are very close to each other, which corresponds to the frequency shifts of peaks in different recordings. Step 4–Cluster Selected Peaks: Cluster the selected peaks with a distance threshold D_t, i.e. the peaks in the same cluster will be at most D_t away from each other. Now we have M'_c clusters whose ranges are denoted by R_j, $j = 1, \ldots, M'_c$, where $R_j = [r_{j,l}, r_{j,u}]$ and $1 \le r_{j,l} \le r_{j,u} \le d$. Step 5–Aggregate Peak Counts: Since we consider each cluster as a result of the horizontal shift of one peak, we aggregate all peak counts within a cluster by summing them up to obtain a single peak count value C_j for each cluster, i.e. $C_j = \sum_{i \in R_j} L_{X,M',i}$. Step 6–Repeat: Repeat Step 1 to 5 with $M' \leftarrow M' + 0.2M$ until $M'_c \ge M$. Then, we pick the top M clusters in terms of the total counts C_j as our final estimated ranges of CP locations $R_{(j)}$, $j = 1, \ldots, M$, where $R_{(j)}$ denotes the j^{th} range ordered descendingly by the corresponding $C_{(j)}$.

Estimating Ranges of CP Locations: CaPE-Threshold. Steps 0 to 2 are the same as CaPE-rank. Step 3–Select Peaks: Select the candidate peak locations with peak count $\ge pN$, where $0 < p \le 1$. In other words, suppose the set of indices for these M' peaks is $I = \{i_1, \ldots, i_{M'}\} = \{i \in \{1, \ldots, d\} \,|L_{X,i} \ge pN\}$. Then, we obtain $L_{X,M'}$, where $L_{X,M',i} = L_{X,i}$ if $i \in \{i_1, \ldots, i_{M'}\}$, else 0. Steps 4 to 5 are the same as CaPE-rank. We do not repeat the above steps in CaPE-threshold. Therefore, the resulting M clusters are the estimated ranges of CP locations $R_{(j)}$, $j = 1, \ldots, M$.

Compressing Spectra to Lower Dimensions. Given $R_{(j)}$, $j = 1, \ldots, M$, we apply max pooling over $R_{(j)}$ to the input data. For the demixing, we only need the resulting intensities. However, to evaluate the DCs afterward, including matching PAHs in the library and calculating AUPRC with the matched PAHs, we also need to set the location of each peak. We choose the cluster center for simplicity, and it turns out that this performs decently well. Let $\text{Mid}\left(R_{(j)}\right) = \lfloor (r_{(j),l} + r_{(j),u})/2 \rfloor$ denote the center of $R_{(j)}$. Then, let x' denote the sparse spectrum we create from x. Each of its entry is

$$x'_i = \begin{cases} \max_{k \in R_{(j)}} x_k, & \text{if } \exists j \in \{1, \ldots, M\}, s.t.\ i = \text{Mid}\left(R_{(j)}\right) \\ 0, & \text{otherwise} \end{cases} \quad \forall i = 1, \ldots, d. \qquad [1]$$

Note that although x' is still a d-dimensional vector, it is M-sparse. We call a spectrum that contains only the M nonzero entries of x' a compressed spectrum $\tilde{x}$. We only need to feed $\tilde{x}$ into the demixing methods. By applying Eq. **1** to every $x \in X$, we obtain a set of sparse spectra $X' \in \mathbb{R}^{d \times N}$ and a set of compressed spectra $\tilde{X} \in \mathbb{R}^{M \times N}$. According to our setup, $M \ll d$.

Comparing CaPE-Rank and CaPE-Threshold. The major difference is that CaPE-rank iterates until the output has a given dimension M, while CaPE-threshold does not iterate. Thus, CaPE-threshold has simpler steps but may produce outputs of different dimensions for different inputs. They are also similar in the sense that CaPE-rank picks the candidate peaks by the ranking of their counts, which is equivalent to a threshold of the M^{th} peak count. The intuition behind CaPE-threshold is that since, ideally, every recording in the data should contain the CPs but not NCPs, the counts for CPs should be close to N, while the counts for NCPs might be much lower. Therefore, in CaPE-threshold, we pick the candidate peaks according to a threshold proportional to N. In our experiments, we find the best value of M for CaPE-rank lies in [30, 50], and the best value of p for CaPE-threshold lies in $[0.1, 0.2]$.

Matching a DC to a PAH Using Similarity. We define the similarity between two spectra as the inner product between them after normalizing each to range [0, 1]. We do not use the cosine similarity because the l_2-norm is highly sensitive to NCPs and background noise in the spectra, even if they have low intensity. Suppose we have two spectra for the same PAH. One is clean, noiseless, only contains the CPs, while the other contains the same CPs but is noisier. Then, the noisier one will have a much larger l_2-norm since the spectra are high dimensional. In other words, two spectra will have a very different normalizing multiplier even if they have exactly the same CPs. This might cause an issue in the matching process—the similarities between different pairs of spectra may not have a consistent scale.

We first calculate the similarities between each DC and each PAH. Since each PAH has multiple recordings, we take the average of the similarities between the DC and each recording. Then, we pick the DC-PAH pair with the highest similarity and remove the DC-PAH pairs containing any DC or PAH already matched. We continue this step until every DC is matched to a PAH.

AURPC. Precision is the ratio between the number of detected CPs and the number of all detected peaks in a DC. Recall is the ratio between the number of CPs detected in a DC and the total number of CPs in the corresponding PAH. The precision-recall curve contains precision-recall pairs obtained by varying the peak detection threshold, which is the minimum height of a peak, from 0 to 1 by a 0.002 interval after normalizing the spectrum to range [0, 1]. We allow a tolerance of 12 indices when counting if the peak locations match, which corresponds to around 10 cm^{-1}. If multiple peaks match the same CP of a PAH, we only count once.

Implementation. All code written by us is in Python 3.7. We use the Python code by Ouedraogo et al. (29) for NICA, the MATLAB code of SparseICA-EBM (30) for SICA, and the FastICA function in the Python package Scikit-Learn (50) for ICA. Since NICA and SICA only accept an input of $d \times C$, where d is the dimension of spectra and C is our guess of the number of sources, while the data have a shape of $d \times N$, where $N \gg C$ is the number of observations, we use the PCA function in Scikit-Learn to extract the top C principal components before feeding the data into NICA or SICA. For NMF, we use the NMF function in Scikit-Learn. For NSNMF, we use the Python package Nimfa (51) for X-ray and SPA, as well as the existing data compression algorithms, including QR decomposition, structured random compression, and Count-Gauss. We use the AgglomerativeClustering function in Scikit-Learn for the clustering of peak counts, with n_clusters = None and distance_threshold = D_t.

Hyperparameter Tuning. We use grid search for all hyperparameter tuning and tune the demixing method and the data compression algorithm together if a compression algorithm is applied. For all demixing methods, we tune the guess of the number of sources C from $\{2, 3, 4, 5, 6, 7, 8\}$. For ICA, we tune the negentropy approximation function from {logcosh, exp, cube}. For NICA, we use 0.1 for the stop tolerance and 100,000 for the maximum number of iterations. We do not find substantial differences between the performance using different values for these two hyperparameters. For SICA, we tune the sparsity parameter λ from $\{0.0001, 0.01, 1\}$ and the smoothing parameter $\in$ from $\{0.001, 0.1, 10\}$. For NMF, we tune the regularization strength for the sources α_W from $\{0.01, 0.1, 1\}$ and the regularization strength for the coefficient α_H from $\{0.01, 0.1, 1\}$. The implementation of NSNMF methods does not contain hyperparameters to tune. However, we can still tune the data compression algorithms for NSNMF. The QR decomposition does not have any hyperparameters. For the structured random compression, we tune the number of power iterations from $\{0, 1, 5, 20\}$, the oversampling parameter from $\{1, 5, 10, 20, 50\}$ and the minimum compression level from $\{5, 10, 20, 40, 80\}$. For Count-Gauss, we tune the oversampling factor from $\{5, 10, 20, 50\}$. For both variants of CaPE, we tune K from $\{1, 5, 9\}$ and D_t from $\{12, 24, 36, 48\}$. For CaPE-rank, we tune M from $\{30, 40, 50\}$. For CaPE-threshold, we tune p from $\{0.05, 0.1, 0.2, 0.4\}$. These value ranges are all determined by the preliminary experiments.

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.

Data, Materials, and Software Availability. All study data have been archived using Zenodo (https://zenodo.org/record/7406328#.Y496iOzMLBI) at DOI: 10.5281/zenodo.7405563.

ACKNOWLEDGMENTS. This research was supported by Welch Foundation grants C-1220 (N.J.H. and M.M.B.) and C-1222 (P.N. and J.Z.) and by the National Institute of Environmental Health Sciences of the National Institutes of Health [P42ES027725].

Author affiliations: [a]Department of Chemistry, Rice University, Houston, TX 77005; [b]Laboratory for Nanophotonics, Rice University, Houston, TX 77005; [c]Department of Computer Science, Rice University, Houston, TX 77005; [d]Department of Materials Science and Nanoengineering, Rice University, Houston, TX 77005; [e]Department of Biochemistry, University of Houston, Houston, TX 77204; [f]Department of Physics and Astronomy, University of Georgia, Athens, GA 30602; [g]Department of Electrical and Computer Engineering, Rice University, Houston, TX 77005; [h]Department of Physics and Astronomy, Rice University, Houston, TX 77005; and [i]Department of Neuroscience, Baylor College of Medicine, Houston, TX 77030

Author contributions: A.P., P.N., and N.J.H. designed research; M.M.B., Y.J., J.Z., S.E., C.F., and O.N. performed research; Y.Z. contributed new reagents/analytic tools; M.M.B., Y.J., J.Z., and O.N. analyzed data; Y.Z. contributed a data set which allowed us to perform initial testing of our computational analysis; and M.M.B., Y.J., J.Z., O.N., A.P., P.N., and N.J.H. wrote the paper.

1. J. Langer *et al.*, Present and future of surface-enhanced Raman scattering. *ACS Nano* **14**, 28–117 (2020).
2. J. Plou *et al.*, Prospects of surface-enhanced Raman spectroscopy for biomarker monitoring toward precision medicine. *ACS Photonics* **9**, 333–350 (2022).
3. D. W. Li, W. L. Zhai, Y. T. Li, Y. T. Long, Recent progress in surface enhanced Raman spectroscopy for the detection of environmental pollutants. *Microchim. Acta* **181**, 23–43 (2014).
4. K. Kneipp, H. Kneipp, I. Itzkan, R. R. Dasari, M. S. Feld, Ultrasensitive chemical analysis by Raman spectroscopy. *Chem. Rev.* **99**, 2957–2976 (1999).
5. J. Kneipp, H. Kneipp, K. Kneipp, SERS-A single-molecule and nanoscale tool for bioanalytics. *Chem. Soc. Rev.* **37**, 1052–1060 (2008).
6. B. I. Escher, H. M. Stapleton, E. L. Schymanski, Tracking complex mixtures of chemicals in our changing environment. *Science* **367**, 388–392 (2020).
7. H. I. Abdel-Shaf, M. S. M. Mansour, A review on polycyclic aromatic hydrocarbons: Source, environmental impact, effect on human health and remediation. *Egypt. J. Pet.* **25**, 107–123 (2016).
8. A. B. Patel, S. Shaikh, K. R. Jain, C. Desai, D. Madamwar, Polycyclic aromatic hydrocarbons: Sources, toxicity, and remediation approaches. *Front. Microbiol.* **11**, 562813 (2020).
9. B. Moorthy, C. Chu, D. J. Carlin, Polycyclic aromatic hydrocarbons: From metabolism to lung cancer. *Toxicol. Sci.* **145**, 5–15 (2015).
10. Anonymous, Toxicological profile for polycyclic aromatic hydrocarbons, US Department of health & human Services, public health service, agency for toxic substances and disease registry, Washington, DC, August, 1985. *J. Toxicol., Cutan. Ocul. Toxicol*. **18**, 141–147 (1999).
11. E. Hussar, S. Richards, Z. Q. Lin, R. P. Dixon, K. A. Johnson, Human health risk assessment of 16 priority polycyclic aromatic hydrocarbons in soils of Chattanooga, Tennessee, USA. *Water Air Soil Pollut.* **223**, 5535–5548 (2012).
12. C. L. Jones, K. C. Bantz, C. L. Haynes, Partition layer-modified substrates for reversible surface-enhanced Raman scattering detection of polycyclic aromatic hydrocarbons. *Anal. Bioanal. Chem.* **394**, 303–311 (2009).
13. L. L. Qu *et al.*, Humic acids-based one-step fabrication of SERS substrates for detection of polycyclic aromatic hydrocarbons. *Analyst* **138**, 1523–1528 (2013).
14. I. Lopez-Tocon, J. C. Otero, J. F. Arenas, J. V. Garcia-Ramos, S. Sanchez-Cortes, Multicomponent direct detection of polycyclic aromatic hydrocarbons by surface-enhanced raman spectroscopy using silver nanoparticles functionalized with the viologen host lucigenin. *Anal. Chem.* **83**, 2518–2525 (2011).
15. Y. H. Kwon, K. Sowoidnich, H. Schmidt, H. D. Kronfeldt, Application of calixarene to high active surface-enhanced Raman scattering (SERS) substrates suitable for in situ detection of polycyclic aromatic hydrocarbons (PAHs) in seawater. *J. Raman Spectrosc.* **43**, 1003–1009 (2012).
16. M. Zhang, X. L. Zhang, B. F. Qu, J. H. Zhan, Portable kit for high-throughput analysis of polycyclic aromatic hydrocarbons using surface enhanced Raman scattering after dispersive liquid-liquid microextraction. *Talanta* **175**, 495–500 (2017).
17. C. Liu *et al.*, Silver nanoparticle aggregates on metal fibers for solid phase microextraction–surface enhanced Raman spectroscopy detection of polycyclic aromatic hydrocarbons. *Analyst* **140**, 4668–4675 (2015).
18. Y. X. Leong *et al.*, Surface-enhanced Raman scattering (SERS) taster: A machine-learning-driven multireceptor platform for multiplex profiling of wine flavors. *Nano Lett.* **21**, 2642–2649 (2021).
19. N. Kim *et al.*, Surface enhanced Raman scattering artificial nose for high dimensionality fingerprinting. *Nat. Commun.* **11**, 207 (2020).
20. F. Safir *et al.*, Detecting pathogenic bacteria in blood with combined acoustic bioprinting, Raman spectroscopy, and machine learning. arXiv [Preprint] (2022). https://doi.org/10.48550/arXiv.2206.09304 (Accessed 9 June 2022).
21. A. Hyvarinen, E. Oja, Independent component analysis: Algorithms and applications. *Neural Networks* **13**, 411–430 (2000).
22. Anonymous, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. P. Comon, C. Jutten, Eds. (Academic Press, 2010), pp. 1–831.
23. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
24. Y. X. Wang, Y. J. Zhang, Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. Knowl. Data Eng.* **25**, 1336–1353 (2013).
25. G. R. Naik, D. K. Kumar, Determining number of independent sources in undercomplete mixture. *EURASIP J. Adv. Signal. Process.* 694850 (2009), 10.1155/2009/694850.
26. J. L. Abell, J. Lee, Q. Zhao, H. Szu, Y. P. Zhao, Differentiating intrinsic SERS spectra from a mixture by sampling induced composition gradient and independent component analysis. *Analyst* **137**, 73–76 (2012).
27. J. Yao, H. Su, Z. X. Yao, Blind source separation of coexisting background in Raman spectra. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **238**, 118417 (2020).
28. M. D. Plumbley, Algorithms for nonnegative independent component analysis. *IEEE Trans. Neural. Networks* **14**, 534–543 (2003).
29. W. S. B. Ouedraogo, A. Souloumiac, C. Jutten, "Non-negative independent component analysis algorithm based on 2D givens rotations and a newton optimization" in *9th International Conference on Latent Variable Analysis and Signal Separation* (Springer-Verlag, St Malo, France, 2010), pp. 522–529.
30. Z. Boukouvalas, Y. Levin-Schwartz, V. D. Calhoun, T. Adali, Sparsity and independence: Balancing two objectives in optimization for source separation with application to fMRI analysis. *J. Franklin Inst. Engr. Appl. Math.* **355**, 1873–1887 (2018).
31. A. Kumar, V. Sindhwani, P. Kambadur, Fast conical hull algorithms for near-separable non-negative matrix factorization.Proceedings of the 30th International Conference on Machine Learning. *PMLR* **28**, 231–239 (2013).
32. N. Gillis, S. A. Vavasis, Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**, 698–714 (2014).
33. A. R. Benson, J. D. Lee, B. Rajwa, D. F. Gleich, "Scalable methods for nonnegative matrix factorizations of near-separable tall-and-skinny matrices" in Advances in Neural Information Processing Systems. **27**, (2014).
34. M. Tepper, G. Sapiro, Compressed nonnegative matrix factorization is fast and accurate. *IEEE Trans. Signal Process.* **64**, 2269–2283 (2016).
35. M. Kapralov, V. K. Potluru, D. P. Woodruff, "How to fake multiply by a gaussian matrix" in Proceedings of The 33rd International Conference on Machine Learning. *PMLR* **48**, 2101–110 (2016).
36. S. J. Oldenburg, R. D. Averitt, S. L. Westcott, N. J. Halas, Nanoengineering of optical resonances. *Chem. Phys. Lett.* **288**, 243–247 (1998).
37. S. J. Oldenburg, J. B. Jackson, S. L. Westcott, N. J. Halas, Infrared extinction properties of gold nanoshells. *Appl. Phys. Lett.* **75**, 2897–2899 (1999).
38. B. E. Brinson *et al.*, Nanoshells made easy: Improving Au layer growth on nanoparticle surfaces. *Langmuir* **24**, 14166–14171 (2008).
39. L. Keith, W. Telliard, ES&T special report: Priority pollutants: Ia perspective view. *Environ. Sci. Technol.* **13**, 416–423 (1979).
40. Z. W. Li, F. Liu, W. J. Yang, S. H. Peng, J. Zhou, A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural. Netw. Learn. Syst.* **33**, 6999–7019 (2021). 10.1109/tnnls.2021.3084827.
41. J. Wright, Y. Ma, *gh-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications* (Cambridge Univ. Press, 2022).
42. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Comm. ACM* **60**, 84–90 (2017).
43. J. D. Lee, J. Yang, Z. Wang, What does CNN shift invariance look like? A visualization study. *Mach. Learn. The 16th European Conference on Computer Vision (Springer, Cham, 2020)*, pp. 196–210.
44. J.-L. Starck, Y. Moudden, J. Bobin, M. Elad, D. Donoho, "Morphological component analysis" in *Wavelets XI* (SPIE, 2005), pp. 209–223.
45. S. Rambhatla, J. Haupt, "Semi-blind source separation via sparse representations and online dictionary learning" in *2013 Asilomar Conference on Signals, Systems and Computers* (IEEE, 2013), pp. 1687–1691.
46. X. Fu, K. Huang, N. D. Sidiropoulos, On identifiability of nonnegative matrix factorization. *IEEE Signal Process. Lett.* **25**, 328–332 (2018).
47. P. B. Johnson, R. W. Christy, Optical constants of noble metals. *Phy. Rev. B* **6**, 4370–4379 (1972).
48. N. Kumawat, P. Pal, M. Varma, Diffractive optical analysis for refractive index sensing using transparent phase gratings. *Sci. Rep.* **5**, 16687 (2015).
49. Z. M. Zhang, S. Chen, Y. Z. Liang, Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst* **135**, 1138–1146 (2010).
50. F. Pedregosa *et al.*, Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
51. M. Zitnik, B. Zupan, NIMFA: A python library for nonnegative matrix factorization. *J. Mach. Learn. Res.* **13**, 849–853 (2012).

Downloaded from https://www.pnas.org by NATIONAL RESEARCH COUNCIL CANADA on May 12, 2025 from IP address 132.246.3.117.