# Data Fusion of ATR-FTIR and Raman Spectroscopies for Component Quantification: Applications at Hanford – 22358

Steven H. Crouse*, Stefani Kocevska*, Rupanjali Prasad*, Martha A. Grover*, Ronald W. Rousseau*
*School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, United States

## ABSTRACT

Multiple in-line technologies are often used in complex processes and can yield complementary information. For example, the Hanford site has the need for accurate quantification of concentrations in multicomponent solutions with the possibility of multiple in-line technologies. Methods exist to combine data sources through data fusion, though the application often dictates the method. In this paper, we combine experimental Raman and infrared spectroscopies and show the efficacy of data-level, feature-level, and decision-level fusion techniques. We show that a simple feature-selection method, combined with data fusion, is able to reduce the mean percent error from 15.8% to 5.6% using the same spectroscopic data. We also show the ability of established Blind Source Separation Techniques to remove non-target species for fused Raman and infrared spectroscopy data.

## INTRODUCTION

The Hanford site in Washington State currently stores 56 million gallons of used nuclear waste. The tanks have begun leaking into the surrounding environment, motivating an on-site cleanup effort. The low activity liquid waste will be immobilized as part of the Direct Feed Low-Activity Waste (DFLAW) process. During the process, the liquid supernatant in the tanks will be treated and vitrified. Currently, the DFLAW process is planned to rely on offline sampling for regulatory checks and measuring concentrations for the addition of Glass Forming Chemicals. Manual sampling is costly, time consuming, and provides an additional radiation exposure route for workers. An alternative that addresses these problems is in-line technology, whereby the concentrations of key species are measured in the process itself.

However, the complexity of this waste prevents immediate implementation of in-line process analytical technology (PAT). The waste contains at least 25 chemical constituents and 46 radionuclides. Furthermore, some of the molecular measurement techniques have spectra with overlapping peaks [1]. Beyond the complex spectra, any analytical technology will require long-term reliability and accuracy for an expected DFLAW project length of 40 years [2].

Previous work has investigated the ability for the in-line technologies of Raman Spectroscopy, Attenuated Total Reflectance – Fourier Transform Spectroscopy (ATR-FTIR), Visible Absorbance, and Laser Induced Breakdown Spectroscopy (LIBS) to characterize this waste [3-5]. Concentration quantification from these technologies can be improved by utilizing the information from multiple sensors at once for increased quantification accuracy and robustness. Multisensor data fusion is the combining of information from multiple sources to achieve a single combined output [6]. There are three levels of data fusion typical for spectroscopic measurements. These are: data-level, feature-level, and decision-level fusion [7,8]. Other distinctions for classes of fusion are described by Federico Castenedo [9].

Data-level data fusion is the combination of raw source information. In the case of in-line technology at Hanford, this is combination of the raw spectra through concatenation, addition, or other data fusion techniques that are described by Moros *et al*. [10]. Feature level data fusion involves the combining of subsets or features of the raw data, rather than the raw data itself. Decision level data fusion is the combination of outputs from regression or classification on each individual sensor. In practice, this is the combination of predicted concentrations that were determined separately for two spectroscopic measurements.

In this work, we look at ATR-FTIR and Raman spectroscopy as in-line measurements for measuring the concentrations of simulated nuclear waste. The relationship between ATR-FTIR and

Raman Spectroscopy is both complementary and redundant using the definitions of Durrant-Whyte [11]. They are complementary because they measure different physical processes. The same molecular bonds are being measured, but different vibrational modes and different physical phenomena (molecular scattering and absorbance) may be used to measure these vibrations. Also, the different penetration depths of these techniques in practice offer different perspectives of two-phase systems, which is the case for some process streams at the proposed Hanford DFLAW process [12]. In addition to being complementary, the information is also redundant. The information from each spectroscopy, on a macro scale, is measuring the same mixture. This redundant information can be combined for more accurate and robust quantification (in the case of a faulty measurement or sensor). Complementary and redundant information is very similar to the data fusion that occurs with human senses [6]. A relatable example of successful data fusion occurs with the vision of binocular animals. Two healthy eyes, in addition to increasing clarity, adds depth perception. This is information that is not available from either eye individually.

## METHODS

The experimental data used is the dataset collected by Kocevska *et al.* [4]. Simulant mixtures were created around a 5.6 M $Na^+$ target with seven sodium salts and water as the solvent. Target species include $NO_3^-$ , $NO_2^-$ , $SO_4^{2-}$ , $CO_3^{2-}$ , and $H_2O$. These were chosen as targets based on their abundance in Hanford waste and the overlapping spectral properties of $NO_3^-$ and $CO_3^{2-}$. Non-target species used are $PO_4^{3-}$ , $C_2O_4^{2-}$ , and $CH_3COO^-$. In this work, water was not included in model evaluation because of its negligible, near-perfect accuracy in all models. Measurement conditions and preprocessing are identical to the conditions in Kocevska *et al.* [4]. An exception is Savitzky-Golay Filtering, where we used a window size of 7 for all measurements. In Kocevska's paper, a window size of 7 and 5 are used for Raman and ATR-FTIR, respectively. Our use of 7 for all trials was to mitigate the effect of pre-processing on model performance.

Blind Source Separation (BSS) can be used to decompose the spectra into its constituent sources with little prior information. In this work, we investigated how established BSS methods work on nontraditional, concatenated spectra. Non-target spectral contributions are removed to increase accuracy of quantification steps when supplied with training data of only the target species ($NO_3^-$ , $NO_2^-$ , $SO_4^{2-}$ , $CO_3^{2-}$ , and $H_2O$). This can save time and resources on calibration experiments for non-target species, as BSS allows for a training set including just target species, even in the case for overlapping spectra [4]. We used the 2-step BSS algorithm used by Maggioni *et al.* The first step, Independent Component Analysis (ICA), provides an initial guess for sources. The sources found by ICA are then fed to the second step: Multivariate Curve Resolution – Alternating Least Squares (MCR-ALS). This algorithm sharpens the guesses from ICA with more natural chemical constraints. To guide the algorithms in identifying the correct sources, single reference spectra are input in addition to the training spectra. For a more detailed description of the BSS methods used, see Maggioni *et al.* and Kocevska *et al.* [4,14]. In this work, reference spectra of the non-target species are fed to the BSS Algorithms, and the sources matching these non-targets are subtracted. This leads to three total sources being subtracted. After BSS is applied, a PLSR model trained with just target species was used to quantify species. To quantify accuracy of our prediction methods, we use Root Mean Square Error (RMSE) and percent error (PE).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(C_{actual}-C_{predicted})^2}{N}} \tag{1}$$

$$PE = \left|\frac{C_{actual}-C_{predicted}}{C_{actual}}\right| \times 100\% \tag{2}$$

**Data-level Data Fusion**

For data-level fusion, Raman and ATR-FTIR spectra were concatenated along the wavenumber axis. The general result for this is shown in Figure 1 for arbitrary reference spectra. Since the wavenumber axis loses physical meaning when concatenated, we transformed the wavenumber axis into a wavenumber index, indexing from 0 to 1471. The first 1227 wavenumbers are from Raman, while the last 245 wavenumbers are from ATR-FTIR.

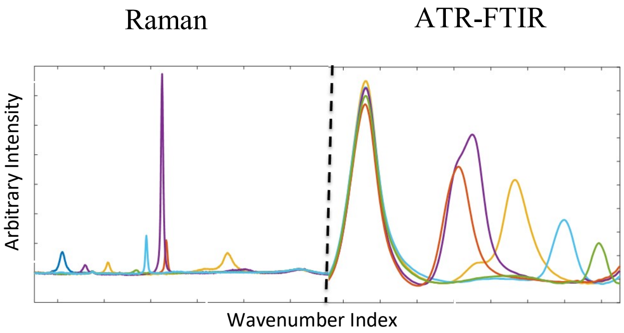

Figure 1: Concatenation of Raman (left) and ATR-FTIR (right) spectra.

**Feature-level Data Fusion**

Two feature selection methods were used in this paper. The first is the Successive Projections Algorithm (SPA). SPA is a forward selection method that is used to select wavenumbers that contain the most information, based on the wavenumber with the highest $L_2$ norm in the training set. The $L_2$ norm is a simple distance metric and determines the wavenumbers with the highest intensities. However, there is an additional constraint beyond finding the highest intensities. SPA finds these wavenumbers while minimizing mutual information between the selected wavenumbers. Mutual information in this instance is information shared as part of a single peak or component. In application, neighboring wavenumbers tend to share much information since they often share the same peaks that vary linearly with each other. Thus, SPA will not select these neighboring wavenumbers. Rather, it finds a spread of important wavenumbers for quantifying a solution rather than just the most prominent peaks. There are several descriptive references for understanding and applying the Successive Projections Algorithm [15,16].

The second feature selection method used is a general forward selection algorithm that selects the wavenumbers that contain the most information. This method operates identically to SPA except it lacks the orthogonalization/subtraction step that gives SPA its “spreading” effect (Step 7 for SPA in Table I). In practice, the general forward selection method will choose the wavenumbers with the largest intensities. It will select these wavenumbers even if they share mutual information (part of the same peak). Figure 2 shows the distinction between these two forward selection methods in the context of reference spectra. SPA will distribute amongst all peaks while the general forward selection method will preferentially select the highest points, regardless of which peak they belong to. It is important to note that these algorithms are applied to the 36 training spectra comprised of only target species. Figure 1 demonstrates the algorithm on the set of 8 reference spectra.

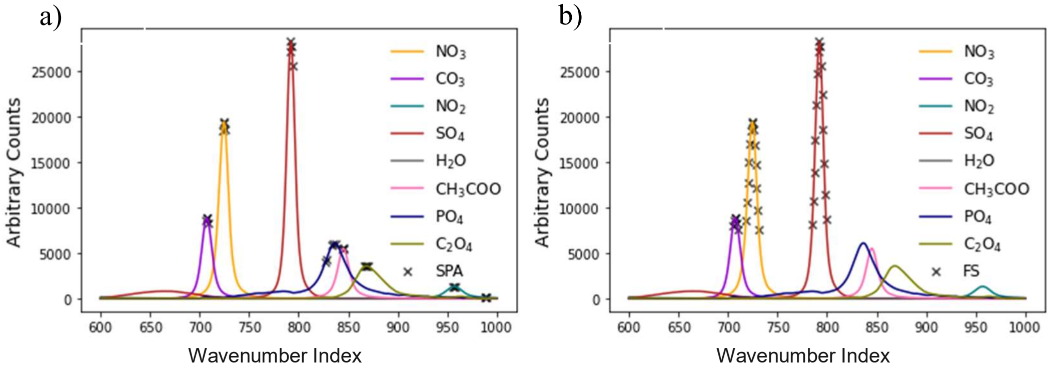

Figure 2: Visualization of forward selection algorithms: 35 features selected from a) SPA and b) a general forward selection algorithm on experimental Raman data.

Table I: SPA and General Forward Selection Algorithm

| **Successive Projection Algorithm** | **General Forward Selection Algorithm** | **Variables** |
|---|---|---|
| 1. $for\ i\ in\ W$:<br>2. $for\ j\ in\ N$:<br>3. $p_j = \|X_j\|_2$<br>4. $end\ j\ loop$<br>5. $k = argmax(p_{0:N})$<br>6. $w = w \cup k$<br>7. $X = \left(I - \frac{X_k X_k^T}{\|X_k\|_2^2}\right) X$<br>8. $end\ i\ loop$ | 1. $for\ i\ in\ W$:<br>2. $for\ j\ in\ N$:<br>3. $p_j = \|X_j\|_2$<br>4. $end\ j\ loop$<br>5. $k = argmax(p_{0:N\backslash w})$<br>6. $w = w \cup k$<br>7. $end\ i\ loop$ | $W$: number of wavenumbers to select<br>$N$: total number of wavenumbers<br>M: number of experiments/samples<br>$X$: data matrix: $MxN$<br>$p_j$: the projection of the j$^{th}$ wavenumber<br>$p_{0:N}$: $p_j$ for all wavenumbers<br>$p_{0:N\backslash w}$: $p_j$ for all wavenumbers not in $w$<br>$\|X_j\|_2$: norm of the j$^{th}$ column<br>$k$: location of maximum projection<br>$w$: output of SPA, selected wavenumbers |

### Decision-level Data Fusion

The decision-level method we use combines the outputs of the BSS-PLSR models for both Raman and ATR-FTIR. Decision-level fusion was implemented by simply averaging output concentrations from both the Raman and ATR-FTIR models. This method is an intuitive combination of model outputs.

## RESULTS AND DISCUSSION

### BSS for Fused Data

In this work, we extend the ideas of Blind Source Separation to dealing with multiple input sources of data. Kocevska *et al*. has previously shown the ability of blind source separation to process ATR-FTIR and Raman spectra for waste measurements. In the case of concatenated spectra, these spectra can be processed as a single spectrum. We have shown the efficacy of Kocevska's BSS algorithm at recognizing the sources from concatenated spectra in Figure 3. It can be seen that the algorithms recognize a single source (a combination of Raman and ATR-FTIR) for all of the species present in solution. After being identified, these sources can be subtracted from the spectra. Figure 4 shows the resulting quantification after the successful removal of non-target species ($PO_4^{3-}$ , $C_2O_4^{2-}$ , and $CH_3COO^-$) from the fused spectra.

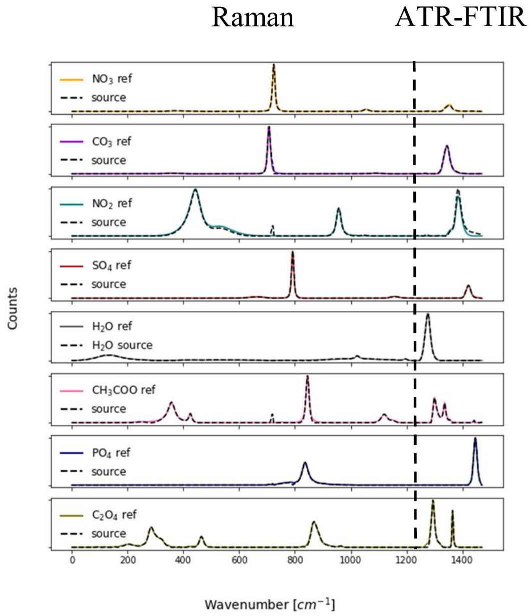

Figure 3: Recognition of sources by the BSS algorithm for concatenated spectra.

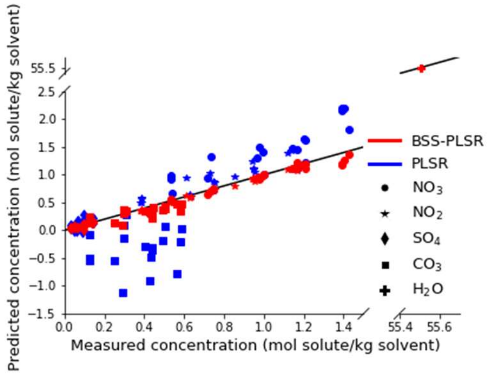

Figure 4: Results of Data Fusion with BSS vs. without BSS. Spectra were fused using a max peak normalization scheme.

## Scaling

Concatenating spectra introduces a problem of scaling. For the y-axis, Raman "counts" can be greater than 30,000, while ATR-FTIR absorbance is typically lower than 1.2. On the x-axis, Raman is sampled every $cm^{-1}$ leading to 1227 wavenumbers. ATR-FTIR is sampled every 3.71 $cm^{-1}$ leading to 245 wavenumbers. In addition, each method gives measurements with different peak widths, peak shapes, peak spacing, and sensitivity to species in solution. There is no known "physically motivated" scaling that accounts for all of these spectral differences. However, a simple scaling factor can adjust the spectra to the same magnitude so that they can be processed. We investigate two separate scalings for data-level fusion: normalization and standard scaling. Normalization is adjusting the scale of Raman and ATR-FTIR so that all points lie between 0 and 1. Standard scaling adjusts the mean of the data to 0 and adjusts the standard deviation to 1 for Raman and ATR-FTIR separately. These scaling schemes are shown below in

Figure 5. It is possible to use standard scaling for every wavenumber rather than the entire sets of Raman and ATR-FTIR data, but this inadvertently amplifies noise so it was not investigated in this work.

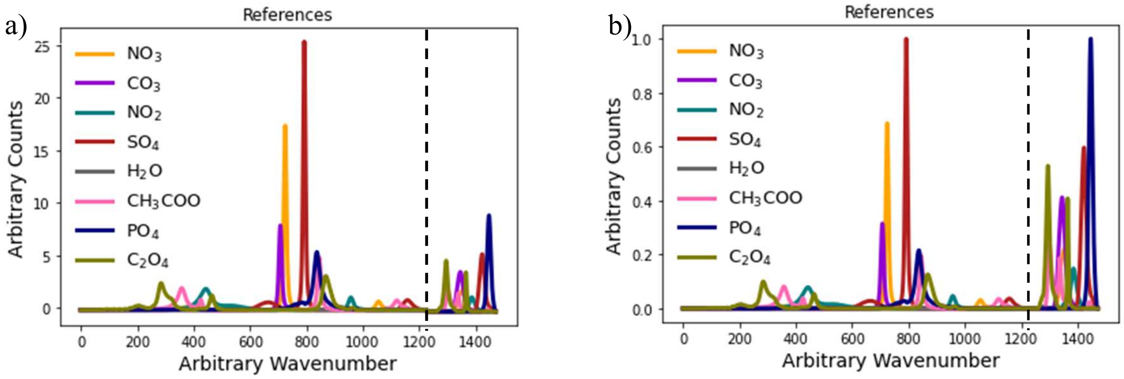

Figure 5: a) Mean centered with unit variance b) Normalized to highest measured peak.

## Feature Selection

Feature selection is fundamentally motivated by the physical features of a spectra. Certain wavenumbers correspond to concentration of species while others do not. Feature selection, therefore, allows for the most important information to be used in model-building. This selected data is combined with the intent that the selected information leads to more accurate quantification. There are competing effects when optimizing the number of features to select. Too few selected features will neglect substantial information about the target species. Too many features included will cause uninformative wavenumbers to be included that can disrupt quantification. This optimization process is shown in Figure 6. In our approach, wavenumbers are selected for each individual method (ATR-FTIR and Raman) and are then combined using a normalization approach to scaling. Global minima for combined RMSE were found for the SPA and general forward selection algorithms.

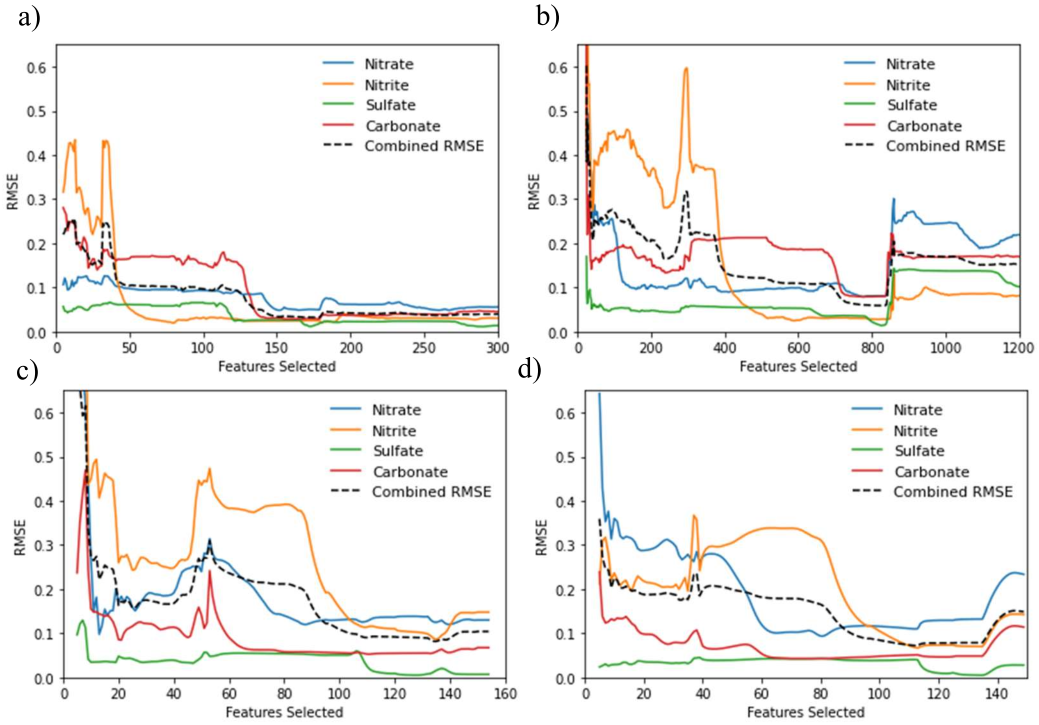

Figure 6: Selection algorithms with optimum number of selected features: a) SPA Raman (166 features), b) forward selection Raman (823 features), c) SPA ATR-FTIR (140 features), and d) forward selection ATR-FTIR (111 features).

**Method Performance**

Table II tabulates the RMSE of the 7 methods compared in this paper. Identical BSS preprocessing is applied to all 7 methods in the table. Raman and ATR-FTIR are standard spectroscopic methods. Normalized and standard-scaled are data-level data fusion techniques. Forward selection and SPA, both used with a normalization scheme, are feature-level data fusion. The last technique, appropriately named "Decision Fusion", is a decision-level data fusion technique.

From Table II, it can be seen that the simple forward selection algorithm outperforms SPA in respect to root mean squared error. This difference can be attributed to the orthogonalization step of SPA. As shown in Figure 2, SPA tends to avoid points that are collinear. This leads to fewer wavenumbers being chosen corresponding to target species and, inadvertently, more being chosen from non-target species locations. Figure 7 shows the general forward selection method in practice and how these optimized spectra appear in relation to the references. In Figure 7, the bold lines represent selected features of the spectra, while dotted lines represent features that are not selected. It can be seen that the spectral features associated with $NO_3^-$ , $NO_2^-$ , $SO_4^{2-}$ , $CO_3^{2-}$ , and $H_2O$ (target species) are largely preserved. The exception to this is part of the sulfate spectrum in the ATR-FTIR range.

Table II: Results from different data processing strategies applied to experimental data from Kocevska *et al*.

| RMSE: | Nitrate | Nitrite | Sulfate | Carbonate | Mean RMSE |
|---|---|---|---|---|---|
| **Raman** | 0.137 | 0.047 | 0.094 | 0.191 | 0.117 |
| **ATR-FTIR** | 0.135 | 0.145 | 0.008 | 0.058 | 0.087 |
| **Normalized Data Fusion** | 0.079 | 0.038 | 0.026 | 0.118 | 0.065 |
| **Standard-Scaled Data Fusion** | 0.057 | 0.065 | 0.059 | 0.198 | 0.095 |
| **Forward Selection with Normalization** | **0.056** | **0.019** | **0.005** | **0.037** | **0.029** |
| **SPA with Normalization** | 0.105 | 0.050 | 0.006 | 0.046 | 0.052 |
| **Decision Fusion with Averaging** | 0.128 | 0.084 | 0.043 | 0.097 | 0.088 |

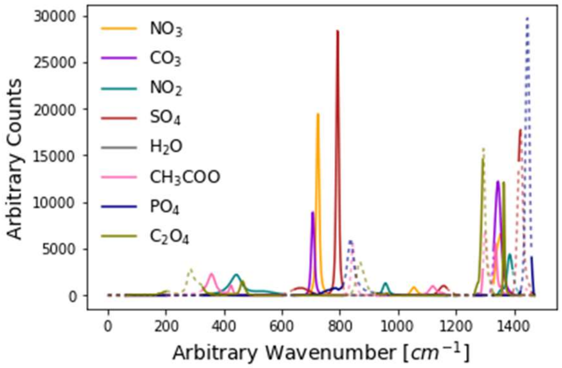

Figure 7: Selected features from Forward Selection applied to reference spectra, dotted line indicates removed portion of the spectra.

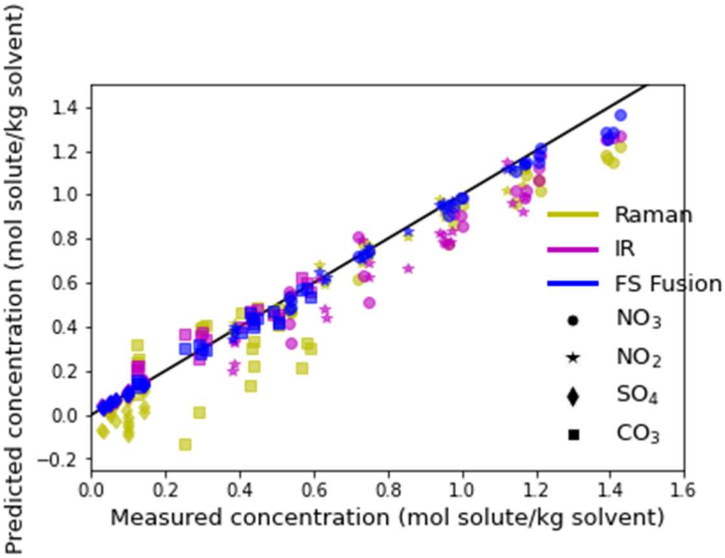

Figure 8: Parity plot showing results of Forward Selection Data Fusion compared with results from Raman and ATR-FTIR. Identical BSS preprocessing and PLSR models were applied to all 3 trials.

From Figure 8 and Table II, it is observed that the forward selection feature-level fusion outperforms all other methods for the prediction of $NO_3^-$, $NO_2^-$, $SO_4^{2-}$, and $CO_3^{2-}$. The improvement in quantification is explained by the feature selection and fusion performed. All preprocessing and quantification steps, including BSS and the PLSR model, are identical for all trials. Thus, the only factors changing for each trial are how the data is fused and the scaling used. The forward feature selection method used tends to select features of the target species based on the training data. These features corresponding to target species are useful for building an accurate model, while unimportant features are removed before fusing the data and do not influence spectra quantification. Fusing these select wavenumbers prioritizes important information being fed to the model.

It can be seen that the Forward Selection with a Normalization Scaling scheme outperforms all other methods for quantifying all species. These improved RMSE values correspond to improved model performance. The mean percent error values corresponding to Figure 8 are: 43.2% (Raman), 15.8% (ATR-FTIR), and 5.6% (Forward Selection Data Fusion). This 5.6% error represents a 2.8 times improvement in the mean accuracy of quantification over ATR-FTIR measurements. In the context of process application, feature selection and data fusion can be used without adding significant computation time. Once important wavenumbers are determined (which can be found for every batch or for every subset of measurements), unimportant wavenumbers can be eliminated from the spectra and any preprocessing (such as smoothing and Blind Source Separation) and quantification (such as PLSR) can proceed as normal with the concatenated spectra.

Decision-level and data-level fusion still have use even though their performance did not match the forward selection method. Decision Fusion had a mean percent error of 25.5% for the target species. Normalized concatenation had a mean percent error of 16.9%, performing between Raman and ATR-FTIR overall. Table II shows that normalized data fusion outperforms Raman and ATR-FTIR for nitrate and nitrite, while not matching the performance of ATR-FTIR for sulfate and carbonate. Data-level fusion does not require optimization of features that is required for feature-level fusion, while still providing additional accuracy beyond the base methods of Raman and ATR-FTIR for some species. Data-level and decision-level fusion may also find use in producing robust measurements when sensor measurements are faulty or disturbed.

## CONCLUSIONS

In this work, we showed how blind source separation tools can be used effectively on concatenated spectra comprised of both Raman and ATR-FTIR spectroscopies. In addition, we demonstrated that feature-level data fusion can be used to increase quantification accuracy of target-species concentrations in a solution that also includes non-target species. A general forward-selection method with a normalization scaling scheme outperformed all other methods tested for measuring all target anions. The mean percent error was improved by a factor of 2.8 for the forward-selection method over ATR-FTIR alone and a factor of 7.7 over Raman; the resulting average quantification error was reduced to 5.6%. This method is practical for implementation, operates on the simple principal of selecting important variables, and offers improvements on the base methods of Raman and ATR-FTIR.

The results of this work may have an enabling positive effect on waste processing at Hanford. Quick and accurate measurements are desired to reduce off-line analysis. Combining measurement technologies in intelligent ways can increase measurement accuracy and robustness, leading to a more efficient vitrification process that operates within stricter tolerances. Future work in this area must still be done on the robustness of data fusion, particularly when process measurements or sensors fail. In addition, there are likely further improvements that may be made to the feature selection algorithm so that it selects wavenumbers based on different criteria. The efficacy of data fusion reported here is a small piece of a much larger picture for real-time in-line monitoring at Hanford.

## RERFERNCES

[1] Lines, A. M., Tse, P., Felmy, H. M., Wilson, J. M., Shafer, J., Denslow, K. M., Still, A. N., King, C., & Bryan, S. A. (2019). Online, Real-Time Analysis of Highly Complex Processing Streams: Quantification of Analytes in Hanford Tank Sample. *Industrial & Engineering Chemistry Research*, *58*, 21194–21200. https://doi.org/10.1021/acs.iecr.9b03636

[2] U.S. Department of Energy. (2019). Hanford Lifecycle Scope, Schedule and Cost Report. *Doe/Rl-2018-45*.

[3] Lines, A. M., Hall, G. B., Asmussen, S., Allred, J., Sinkov, S., Heller, F., Gallagher, N., Lumetta, G. J., & Bryan, S. A. (2020). Sensor Fusion: Comprehensive Real-Time, On-Line Monitoring for Process Control via Visible, Near-Infrared, and Raman Spectroscopy. *ACS Sensors*, *5*(8), 2467–2475. https://doi.org/10.1021/acssensors.0c00659

[4] Kocevska, S., Maggioni, G. M., Rousseau, R. W., & Grover, M. A. (2021). Spectroscopic Quantification of Target Species in a Complex Mixture Using Blind Source Separation and Partial Least-Squares Regression: A Case Study on Hanford Waste. *Industrial and Engineering Chemistry Research*, *60*(27), 9885–9896. https://doi.org/10.1021/acs.iecr.1c01387

[5] Stone, M. E., Diprete, C. C., Farrar, M. E., Howe, A. M., Miera, F. R., & Poirier, M. R. (2017). *WTP Real-Time , In-Line Monitoring Program Task 2 : Determine the Technical Basis for Process Control and Task 5 : Process Control Challenges*. *December*.

[6] Hall, D. L., & Llinas, J. (2016). An introduction to multi-sensor data fusion. *Sensors, Nanoscience, Biomedical Engineering, and Instruments*, *85*(1).

[7] Moros, J., & Laserna, J. (2011). New Raman - Laser-Induced Breakdown Spectroscopy Identity of Explosives Using Parametric Data Fusion on an Integrated Sensing Platform. *Analytical Chemistry*, 6275–6285. https://doi.org/10.1021/ac2009433

[8] Borras, E., Ferre, J., Boque, R., Mestres, M., Acena, L., & Busto, O. (2015). Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Analytica Chimica Acta*, *891*. https://doi.org/10.1016/j.aca.2015.04.042

[9] Castanedo, F. (2013). A review of data fusion techniques. *The Scientific World Journal*, *2013*. https://doi.org/10.1155/2013/704504

[10] Moros, J., & Laserna, J. J. (2015). Unveiling the identity of distant targets through advanced Raman-laser-induced breakdown spectroscopy data fusion strategies. *Talanta*, *134*, 627–639. https://doi.org/10.1016/j.talanta.2014.12.001

[11] H. F. Durrant-Whyte, “Sensor models and multisensor integration,” *International Journal of Robotics Research*, vol.7, no. 6, pp. 97–113, 1988.

[12] Poirier, M. R., Howe, A. M., Miera, F. R., Stone, M. E., & DiPrete, C. C. (2018). *WTP Real-Time In-Line Monitoring Program : Task 1: LAW and EMF Analytes and Properties - Functional Requirements*. March.

[13] Vincent Mazet (2021). Background correction (https://www.mathworks.com/matlabcentral/fileexchange/27429-background-correction), MATLAB Central File Exchange. Retrieved November 12, 2021.

[14] Maggioni, G. M., Kocevska, S., Grover, M. A., & Rousseau, R. W. (2019). Analysis of Multicomponent Ionic Mixtures Using Blind Source Separation: A Processing Case Study. *Industrial and Engineering Chemistry Research*, *58*(50), 22640–22651. https://doi.org/10.1021/acs.iecr.9b03214

[15] Gillis, N. (2014). Successive nonnegative projection algorithm for robust nonnegative blind source separation. *SIAM Journal on Imaging Sciences*, *7*(2), 1420–1450. https://doi.org/10.1137/130946782

[16] Araújo, M. C. U., Saldanha, T. C. B., Galvão, R. K. H., Yoneyama, T., Chame, H. C., & Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, *57*(2), 65–73. https://doi.org/10.1016/S0169-7439(01)00119-8

**ACKNOWLEDGEMENTS**

Support by the U.S. Department of Energy under Cooperative Agreement DE- FC01-06EW07053, entitled “The Consortium for Risk Evaluation with Stakeholder Participation III,” is gratefully acknowledged.