ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Pursuit of the Ultimate Regression Model for Samarium(III), Europium(III), and LiCl Using Laser-Induced Fluorescence, Design of Experiments, and a Genetic Algorithm for Feature Selection

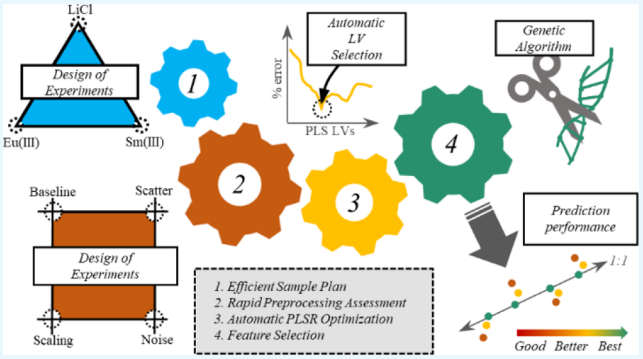
Hunter B. Andrews, Luke R. Sadergaski,* and Samantha K. Cary

Cite This: *ACS Omega* 2023, 8, 2281−2290

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Laser-induced fluorescence spectroscopy, Raman scattering, and partial least squares regression models were optimized for the quantification of samarium (0−150 μg mL^{-1}), europium (0−75 μg mL^{-1}), and lithium chloride (0.1−12 M) with a transformational preprocessing strategy. Selecting combinations of preprocessing methods to optimize the prediction performance of regression models is frequently a major bottleneck for chemometric analysis. Here, we propose an optimization tool using an innovative combination of optimal experimental designs for selecting preprocessing transformation and a genetic algorithm (GA) for feature selection. A D-optimal design containing 26 samples (i.e., combinations of preprocessing strategies) and a user-defined design (576 samples) did not statistically lower the root mean square error of the prediction (RMSEP). The greatest improvement in prediction performance was achieved when a GA was used for feature selection. This feature selection greatly lowered RMSEP statistics by an average of 53%, resulting in the top models with percent RMSEP values of 0.91, 3.5, and 2.1% for Sm(III), Eu(III), and LiCl, respectively. These results indicate that preprocessing corrections (e.g., scatter, scaling, noise, and baseline) alone cannot realize the optimal regression model; feature selection is a more crucial aspect to consider. This unique approach provides a powerful tool for approaching the true optimum prediction performance and can be applied to numerous fields of spectroscopy and chemometrics to rapidly construct models.



1. INTRODUCTION

Optical spectroscopy and chemometric analysis are key components for online monitoring applications in all fields of chemical processing. Food processing and pharmaceutical industries have applied this technology to improve process efficiency, quality, safety, and compliance.[1,2] Implementing such technology to support chemical operations in restrictive environments (e.g., radiochemical hot cells) has the potential to modernize procedures often held up by the challenges associated with traditional grab sample collection.[3,4] In addition to quantifying analytes, optical approaches can simultaneously provide chemical insight into the system and elucidate speciation, oxidation states, and complex chemical interactions. Optical techniques are also flexible and can be integrated into a variety of process batch or stream types. This approach could benefit numerous fields in nuclear science and technology (e.g., radioisotope production), but several challenges must be addressed prior to adoption at the industrial scale.[5,6]

The use of time-resolved laser-induced fluorescence spectroscopy (TRLFS) is well-established for the detection of numerous lanthanides (e.g., Ce^{3+}, Pr^{3+}, Eu^{3+}, Tb^{3+}, Gd^{3+}, Dy^{3+}, Sm^{3+}, and Tm^{3+}) and actinides (e.g., UO_2^{2+}, Am^{3+}, Cm^{3+}, Cf^{3+}, Bk^{3+}, and Es^{3+}), species highly relevant to the nuclear field.[7−18] Most research on lanthanide luminescence in aqueous environments has been concerned with Eu(III) and Tb(III). The other two visibly luminescent ions, Sm(III) and Dy(III) have received less attention because they have inferior luminescence quantum yields (i.e., more efficient nonradiative relaxation). Although the time dimension provides a unique fingerprint for various species, the analysis time is often ∼10−15 min, which is too slow for processes requiring real-time feedback.[16] Laser-induced fluorescence (LIF) spectra can be measured directly, using a charge-coupled device, to quantify complex systems with overlapping bands, matrix effects, chemical interaction(s), and baseline offsets using multivariate chemometrics.[17]

One of the most robust supervised regression techniques, partial least squares regression (PLSR),[18] models the

Received: October 13, 2022
Accepted: December 14, 2022
Published: January 3, 2023

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

covariance between two matrices corresponding to the spectra (X) and concentrations (Y) using combinations of latent variables (LVs). PLSR models require a representative training set that can be efficiently selected using design of experiments.[5,19−21] The regression analysis can be significantly improved by applying preprocessing transformations (e.g., baseline, scatter, noise, and scaling) to the spectral data.[22−24] However, this process is often subject to user bias and is time-consuming; typically, the researcher manually tests thousands of potential combinations by trial and error or settles for something far less than the true optimum. Several studies have attempted to address this bottleneck using design of experiments or machine learning algorithms to select preprocessing strategies.[20,25,26] However, the work discussed here advances both preprocessing and feature selection so that it can readily adapt to disparate data sets and increase the breadth and ease of model development while improving predictive capabilities.

Here, we automate the optimization of PLSR models, built from LIF measurements, by combining experimental design for selecting preprocessing transformations and a genetic algorithm (GA) for feature selection.[20,27,28] This new approach improves the timeliness of the model development process while preserving reasonable computing power, resulting in prediction performance that approaches the true optimum. Calibration and validation fluorescence spectral data sets were selected by determinant-optimal (D-optimal) designs to minimize the samples required in the training set, which spanned Sm(III) (0−150 μg mL^{-1}), Eu(III) (0−75 μg mL^{-1}), and LiCl (0.1−12 M) concentrations, conditions highly applicable to monitoring LiCl anion exchange column effluent streams for the ^{252}Cf Program at Oak Ridge National Laboratory.[29] The approach can also be extended to monitoring lanthanide fission product species throughout the nuclear fuel cycle. Three points of scientific advancement are covered in this work: (1) multivariate analysis enables quantitative Sm(III) and Eu(III) predictions without recording luminescence lifetimes, (2) D-optimal design (DOD) combined with a GA efficiently optimizes preprocessing and feature selection, and (3) this adaptable selection strategy is automated and provides a robust workflow for rapid model optimization. This state-of-the-art approach can assist both chemometricians and nonspecialists in their quest to find the ultimate regression model in many applications within and beyond the nuclear field.

2. EXPERIMENTAL SECTION

All chemicals were commercially obtained (ACS grade) and used as received unless otherwise stated. Lithium chloride (99% purity) was purchased from VWR Scientific, and 37% hydrochloric acid was purchased from Sigma Aldrich. Certified samarium (10,000 ± 30 μg mL^{-1}) and europium (10,000 ± 54 μg mL^{-1}) inductively coupled plasma optical emission spectroscopy standard solutions in 4% hydrochloric acid were purchased from High-Purity Standards. Samples were prepared using deionized water with Milli-Q purity (18.2 MΩ cm at 25 °C).

2.1. Sample Preparation. Calibration and validation samples contained samarium (0−150 μg mL^{-1}), europium (0−100 μg mL^{-1}), and LiCl (0.1−12 M) and covered the anticipated solution conditions. Sample concentrations were chosen using experimental designs built with Design-Expert (v.11.0.5.0) by Stat-Ease Inc., within the Unscrambler software package by Camo Analytics. D-optimal samples were chosen using both point and coordinate exchange and a quadratic process order and evaluated by assessing the fraction of design space.[30,31] Samples were prepared gravimetrically, using a Mettler Toledo model XS204 balance, with an accuracy of ±0.0001 and volumetric glassware. Sample concentration uncertainties were determined by standard error propagation methods described in the Supporting Information. The average relative standard deviation of each sample concentration was 1.0, 1.7, and 0.64% for Sm(III), Eu(III), and LiCl concentrations, respectively. Each sample was prepared in individual 2 mL plastic microcentrifuge tubes (VWR Scientific, 525−1160). A micro-volume (100 μL) UV fused quartz fluorescence cuvette made by ThorLabs (CV10Q1FE) was used for each measurement. The cuvette was stored on lint-free Kimwipes and periodically rinsed with dilute HCl. For each measurement, the cuvette (Z-height of 8.5 mm) was placed in a Quantum Northwest qpod 2e temperature-controlled sample compartment holder purchased from Avantes (CUV-UV/Vis-TC). The compartment had two collimating lenses (CUV-TC-QCL-UV) placed at 90°. Fluorescence measurements were performed at a constant temperature (22 °C) with an accuracy of ±0.05 °C.

2.2. Fluorescence Spectroscopy. Laser fluorescence and Stokes Raman spectra were collected with a fully automated imaging iHR 320 spectrometer (Horiba Scientific). A continuous-wave LBX 405 nm laser (Oxxius) operating at 100 mW was used as the excitation source. Thorlabs multimode fibers—a 105 μm core diameter (M105L02S-A) and a 600 μm core diameter fiber (M134L01)—were used on the excitation and emission sides, respectively. Each spectrum comprised 5585 data points. Static measurements were recorded in triplicate from 410 to 790 nm using a 600 groove mm^{-1} grating and a 100 μm slit size.

Lifetimes were measured using a Fluorolog-QM spectrometer (Horiba), a single-channel R928P PMT, and a DeltaTime kit. Lifetimes were recorded using multichannel scanning mode and a SpectralLED-390 (394 nm, fwhm 14 nm) with the DeltaHub. The operational frequency range for the pulsed diode SpectralLED light source is 0.1−2.9 kHz. The excitation wavelength overlapped the Eu(III) $^7F_0 \rightarrow {}^5L_6$ transition. Lifetimes were calculated using a fitting algorithm $D(t)$ with the one-to-four exponential PowerFit-10 application in Horiba software by eq 1

$$D(t) = \sum a_i \exp\left(\frac{-t}{\tau_i}\right) \quad (1)$$

where a_i is the preexponential factor, t is time, and τ_i is the fluorescence lifetime.

2.3. Preprocessing and Feature Selection. Several well-chosen preprocessing strategies were chosen for this work based on a previous study.[20] Transformations included standard normal variate (SNV) analysis, mean centering (MC), and Savitzky−Golay (SG) filters to account for scattering, scaling, and noise/baseline issues, respectively. SG filters contained 3 derivative levels, 4 polynomial order levels, and 29 smoothing point levels (Table S1). Preprocessing combinations were applied as follows: (1) scatter correction (SNV analysis), (2) noise/baseline (SG), and (3) scaling (MC). Additional details for each technique are provided in the Supporting Information.

A GA, a metaheuristic optimization approach developed based on natural selection concepts, was used for spectral

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

feature selection. The specific GA employed was developed in Python 3 and is described in detail elsewhere.[27] Briefly, a population of 30 binary arrays, or filters, of equal length to the spectra wavelength was randomly generated. Iteratively, the spectra matrix was multiplied by each binary array, where 1 corresponds to the wavelength being modeled, and 0 removes that portion of the spectra. The filters were built with a 10 nm resolution (i.e., spectra were filtered on/off every 10 nm) to ensure proper feature selection. The modified spectra were then used to build a PLSR model and evaluated with a test set to determine how well the filtered data could be modeled. The filters were then sorted based on how well they performed. Next, a new generation of filters was developed by combining various parts of the previous filters together. The cross-over strategies employed are detailed elsewhere.[27] The top-performing filters are passed onto the next generation without modification. Random mutation (i.e., flipping a 1 to 0) of the filters was done with a 5% probability. This served to help break away from local optimal points. If no improvement was seen in 30 generations, the GA underwent a soft reboot, wherein the top 10% of the filters were retained, and the remaining 90% of the filters were replaced with new random arrays. This is another strategy to overcome local optimal points. With the new population of filters, the next generation begins, and the process repeats until the defined number of generations is reached. The number of generations is defined by the user and typically selected by performing a feature selection with a large number of generations and evaluating when an optimal filter was reached.

2.4. Partial Least Squares Regression. PLSR, one of the most popular multivariate modeling methods, performs well for regressing spectra where the number of independent variables (X spectral matrix) is significantly larger than the number of samples. PLSR transforms spectra and a matrix of analyte concentrations into a latent space. Then vectors, referred to as LVs, are iteratively solved to explain the most covariance between the spectra and the analyte concentrations. More variance in the response matrix is explained as more LVs are added to the model until additional LVs begin to overfit the data and reduce the accuracy by modeling noise.

LV selection is typically performed using a set of test samples to evaluate model performance. This set of test samples can be either from cross-validation or an independent set and was never included in model construction. The latter option was implemented in this study because all calibration samples were needed from the optimal design. To investigate model performance versus LVs, the models were reconstructed with a different number of LVs and then used to estimate the concentrations in the validation set. Here, a prediction error metric can be compared to the number of LVs to identify an optimal number of LVs. Unfortunately, there is no definitive rule of thumb for selecting the number of LVs, and many studies are largely subject to user decision. This study employs an automated LV selection script. All regression models and data preprocessing were completed in Python 3 using modules from the Scikit-Learn library.[32]

2.5. Statistical Comparison. The root mean squared error (RMSE) was used as the primary metric for prediction error, defined in eq 2 as

$$\mathrm{RMSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

where y_i is the known concentration, $\hat{y}_i$ is the model predicted concentration, and n is the total number of samples.[33] The RMSE of prediction (RMSEP) measures the dispersion of samples not included in the training set (i.e., validation set) about the regression line. It is typical to discuss RMSE values in terms of percentages to ease comparisons. For this, the RMSE value is divided by the median of the model concentration range (RMSEP %). Lower RMSE values indicate improved model performance.

In this study, the number of LVs in PLSR models was selected using the following procedure: (1) RMSEP values were iteratively determined for PLSR models with LVs in the user-defined range (1−10) using the validation sample set, (2) the percent reduction in RMSEP compared with the previous minimum RMSEP was calculated for each subsequent LV included in the model, and (3) the number of LVs was selected to be the last LV corresponding to a percent reduction ≥10%. This procedure was based on previous studies with LV selection and allowed for automation when building and evaluating PLSR models.[17,27]

Model prediction performance was compared using Tukey−Kramer significance tests.[34,35] Full details are provided elsewhere.[19,21] The Tukey−Kramer method performs a pairwise comparison of model RMSEPs for each analyte, assuming the null hypothesis that the mean predictions for each model are equal. The bias and standard error of prediction (SEP) ratio confidence intervals were determined for each model using a 95% confidence interval. The prediction performance of the two models was considered statistically similar when both the bias confidence interval and the SEP ratio contained 0 and 1, respectively. Additional details on the Tukey−Kramer analysis are provided in the Supporting Information.

3. RESULTS AND DISCUSSION

3.1. Fluorescence Measurements. Laporte forbidden f−f transitions lead to f-elements (e.g., lanthanides) with unique spectral fingerprints corresponding to specific elements and oxidation states.[7,8] Though f-elements typically have small Stokes shifts and band broadening, minor changes in their coordination environment can lead to small perturbations in their emission spectra.[9−12] Even though f-element absorption and emission spectra are described as "line-like," the fine structure of these spectra has proven to be useful for determining the symmetry and coordination environment of lanthanide species.[15,17]

The emission spectra of Eu(III) and Sm(III) in differing concentrations of LiCl (3−12 M) exhibit numerous lines between 550 and 725 nm (Figure 1). The emission spectra are likely affected by long-range interaction between the cations and solvent molecules and outer-sphere complexation with chloride anions.[36,38] Increasing the LiCl concentration had the most notable effect on the fine structure of the Sm(III) emission line near 645 nm given from $^4G_{5/2} \rightarrow {}^6H_{9/2}$. The $^6H_{9/2}$ band changed the most in shape and relative intensity to the other peaks in the spectrum. Band intensities increased with a more coordinating environment as LiCl concentration increased, which suggests that the transitions have an electric dipole (ED) character.[17]

Several Eu(III) lines were ascribed to transitions from the first excited $^5D_0$ state to the $^7F_{0-4}$ Stark levels. Increasing the LiCl concentration led to more changes in the fine structure and the emergence of lines not seen at lower concentrations.

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

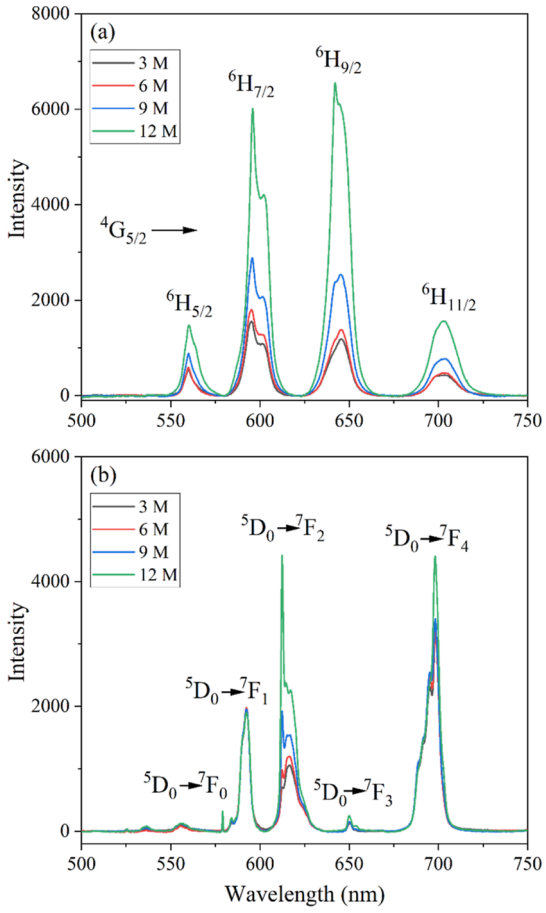


Figure 1. Emission spectrum for (a) Sm(III) and (b) Eu(III) at 100 ppm in 3−12 M LiCl. Spectra were processed using a linear baseline correction.

The $^5D_0 \rightarrow {}^7F_1$ transition has a magnetic dipole character and changes minimally with changing LiCl concentration. The $^5D_0 \rightarrow {}^7F_{0,3}$ transitions are due to ligand field effects while the transitions $^5D_0 \rightarrow {}^7F_{2,4}$ have ED character and are allowed due to the lack of symmetry at the Eu(III) site. These peaks grow significantly with increasing LiCl concentration. Changes in the fine structure for these transitions ($^5D_0 \rightarrow {}^7F_{2,4}$) are thought to be associated with changes in the coordination environment as the concentration of Cl^- increases, with the most notable change being the sharp peak near 612.4 nm.[15] A small peak appeared at 653.9 nm next to the $^5D_0 \rightarrow {}^7F_3$ transition near 650.1 nm at the highest LiCl concentration. The small, sharp peak near 579 nm appeared with increasing LiCl. The presence of one $^5D_0 \rightarrow {}^7F_0$ transition indicates the presence of at least one Eu(III) local environment.[11] This peak is associated with $^5D_0 \rightarrow {}^7F_0$, a nondegenerative forbidden transition that can gain intensity through J-mixing in different symmetries such as Cs, Cn, and Cnv.[12] Typically, this transition is seen only when Eu(III) cations are in an extremely asymmetric environment.

Comparing emission peak areas is a useful indicator for both inner-sphere and outer-sphere ligand environments for Eu(III). The slight change in the Eu(III) peak area ratio (A_2/A_1) of the $^5D_0 \rightarrow {}^7F_2$ (A_2) and $^5D_0 \rightarrow {}^7F_1$ (A_1) transitions indicates a decrease in the coordination symmetry and the formation of outer-sphere chloro complex(s) < 9 M LiCl. The A_2/A_1 ratio at higher LiCl concentrations ≥9 M indicates inner-sphere complexation with chloride (Cl^-) anions.[15,37] Comparing the relative Sm(III) peak areas from the $^4G_{5/2} \rightarrow {}^6H_{9/2}$ and $^4G_{5/2} \rightarrow {}^6H_{11/2}$ transitions, after normalization (e.g., SNV), could provide a similar indicator for the formation of inner-sphere chloride complexation at higher LiCl concentrations.

The luminescence lifetime for Eu(III) in 12 M LiCl solution (110 ± 1 μs) was consistent with published values.[36,37] We also report lifetimes for Sm(III) in 3, 6, and 12 M LiCl solutions of 2.70, 2.87, and 3.06 μs (±0.05 μs), respectively. Minimal change in Sm(III) luminescence lifetime as a function of electrolyte concentration is consistent with Eu(III) and suggests that the number of inner-sphere water molecules surrounding these cations slightly changes. Although lifetimes provide valuable information, they do not provide explicit information regarding inner-sphere complexation with chloride ions.[37] Decay curves are provided in the Supporting Information (Figures S2 and S3). Pinpointing the exact nature of the coordination environment for Sm(III) and Eu(III) in this case with luminescence data alone is difficult. Techniques such as X-ray absorbance spectroscopy coupled with known solid-state structures would help determine exact speciation.[14]

Stokes Raman scattering peaks corresponding to the O−H stretching and bending regions were also identified from fluorescence measurements. The most intense band corresponded to the O−H stretching region, which consists of multiple overlapping bands.[17,18] The intensity of this band increased with increasing LiCl concentration, and an isosbestic point was located near 464.4 (Figure S1). The Raman water band is sensitive to any perturbation that impacts water structure (e.g., ionic strength and temperature), but it was insensitive to Sm(III) and Eu(III) at the low concentrations evaluated in this study.

3.2. Selecting Sample Concentrations. Multivariate regression models were built using the calibration and validation sets shown in Table 1. The Sm(III), Eu(III), and LiCl concentration ranges covered those expected in the expected application. Each sample was selected by D-optimal experimental design, a useful approach for minimizing the number of samples required in spectral training sets. One-factor-at-a-time methods are more commonly used for selecting samples.[38] For example, a three-factor set varied at five levels would require 125 samples (5^3). Ten required model points were augmented with 15 lack-of-fit (LOF) points. LOF samples fall within the factor space (i.e., no vertex points) and were included as either calibration or validation samples. LOF fit points maximize both the distance to other runs and the determinant of the information matrix $X'X$ while satisfying the optimality criterion. Here, the calibration set contained 15 samples, while the validation set contained 10 samples covering the factor space for each variable.

3.3. Selection of Optimal Preprocessing Combinations. Multivariate regression models benefit from preprocessing strategies that remove artifacts (i.e., unwanted variation) from spectral data.[20,22−26] Removing artifacts highlights the relevant structure within a spectrum, making the data more amenable to regression analysis. Numerous techniques are available, but preprocessing strategies are often selected through trial and error or experience.[23] There are no clear guidelines for when to use specific techniques or combinations of techniques. In our previous study, the DOD of experiments was used to determine an optimal preprocessing combination with fewer trials than evaluating all possible combinations (26 vs 576).[20] The D-optimal approach provides a simple approach

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

Table 1. D-Optimal Selected Analyte Concentrations With Space and Build Types[a]

run	Sm(III) (μg mL^{-1})	Eu(III) (μg mL^{-1})	LiCl (M)	space type	build type
1*	66.0	75.0	10.6	plane	LOF
2	**0.0**	**45.0**	**4.9**	**plane**	**model**
3	24.8	34.9	0.1	plane	LOF
4	**89.6**	**0.0**	**4.9**	**plane**	**model**
5	16.5	0.0	5.4	plane	LOF
6	**150**	**0.0**	**12.0**	**vertex**	**model**
7	**150**	**26.6**	**0.1**	**edge**	**model**
8*	150	41.4	9.5	plane	LOF
9*	150	50.6	4.0	plane	LOF
10*	65.3	7.5	0.1	plane	LOF
11	**53.6**	**75.0**	**0.1**	**edge**	**model**
12	**89.3**	**44.6**	**12.0**	**plane**	**model**
13	**0.0**	**75.0**	**12.0**	**vertex**	**model**
14*	24.5	75.0	6.3	plane	LOF
15*	42.8	21.8	9.0	interior	LOF
16*	91.5	70.1	4.8	nterior	LOF
17	**0.0**	**0.0**	**12.0**	**vertex**	**model**
18*	90.0	34.9	5.5	interior	LOF
19	91.5	45.8	0.1	plane	LOF
20*	133	75.0	0.1	edge	LOF
21	**0.0**	**0.0**	**0.1**	**vertex**	**model**
22*	0.0	40.5	12.0	edge	LOF
23	**150**	**75.0**	**7.8**	**edge**	**model**
24	150	13.5	6.3	plane	LOF
25	80.3	0.0	12.0	edge	LOF

[a]*LOF points included in the validation set. Required model points are bolded. Abbreviations include LOF.

to testing a reasonable number of combinations without requiring specialized coding experience. This study replicates this preprocessing strategy using a different analyte and spectroscopy system and then expands this approach by applying a GA for feature selection.

PLSR models were built using multiple preprocessing strategies. The simplest strategy, namely no preprocessing (NP), was used to benchmark the improvements afforded by the advanced strategies. The DOD strategy included 26 unique SNV, SG, and MC preprocessing combinations, whereas the user-defined design (UDD) included all possible SNV, SG, and MC preprocessing combinations (576 samples). Details for each design are provided in Table S1 and elsewhere.[20]

A Python script was developed to first build an NP PLSR model and determine the optimal number of LVs based on the CV procedure detailed in Section 2.4. A similar script was developed to iteratively build PLSR models using the DOD or UDD preprocessing combinations and automatically select the optimal number of LVs in each case. The number of LVs was limited to ≤10 LVs. Although the Y matrix comprised three concentrations (i.e., LiCl, Sm, and Eu), the Sm and Eu spectral intensity and shape changed as a function of LiCl concentration due to the formation of chloride complexes. Therefore, it was reasonable to consider more than three LVs. All PLSR models were evaluated by comparing the predictions of the validation samples to their known concentrations, as represented by the RMSEP % value. Each analyte was modeled individually (e.g., PLS-1) to better identify preprocessing strategies particular to each correlated behavior and make model comparisons more concise.

The NP PLSR model used 10 LVs to model Sm(III) concentration and resulted in an RMSEP % of 7.13%. The best-performing DOD and UDD models both used SG smoothing and derivatives for preprocessing. Both the DOD and UDD smoothing used a window of 61 points and a third-order polynomial. The DOD model used a second-order derivative, while UDD selected the first-derivative transformation. The Sm(III) DOD and UDD models both used 9 LVs and resulted in similar RMSEP % values of 6.41 and 6.21%, respectively. For Eu(III), the NP PLSR model used 9 LVs and had an RMSEP % value of 5.26%. DOD smoothing used a window of 27 points and a fifth-order polynomial, whereas UDD smoothing used a window of 13 points and a third-order polynomial. The DOD model performed similarly to the NP model, resulting in a model with 9 LVs and an RMSEP of 5.22%. The UDD selection built a model with a first derivative, resulting in a model with fewer LVs (6) and a lower RMSEP % of 4.41%.

The LiCl NP model used 8 LVs and had an RMSEP % value of 7.05%. The best-performing DOD model utilized SG smoothing with a window of 61 points, a third-order polynomial, and a second-order derivative. The LiCl DOD model used 7 LVs and had an RMSEP % of 6.50%. The top UDD model used SG smoothing with a window of 45 points and a fifth-order polynomial and, again, a second-order

Table 2. Model Information and Comparison of Predictive Performance before and after GA[a]

model information				unfiltered		GA-filtered	
Sm(III)	Run No.	SNV, SG, MC	LVs	RMSEP	RMSEP %	RMSEP	RMSEP %
NP			10	5.346	7.13	0.852	1.14
DOD	15	0,(2,3,61),0	9	4.805	6.41	0.680	0.91
UDD	506	0,(1,3,61),0	9	4.661	6.21	0.844	1.13
Eu(III)	Run No.	SNV, SG, MC	LVs	RMSEP	RMSEP %	RMSEP	RMSEP %
NP			9	1.971	5.26	1.314	3.50
DOD	16	0,(0,5,27),0	9	1.959	5.22	1.334	3.56
UDD	303	0,(1,3,13),0	6	1.654	4.41	1.333	3.55
LiCl	Run No.	SNV, SG, MC	LVs	RMSEP	RMSEP %	RMSEP	RMSEP %
NP			8	0.423	7.05	0.221	3.68
DOD	15	0,(2,3,61),0	7	0.390	6.50	0.128	2.14
UDD	373	0,(2,5,45),0	6	0.240	4.00	0.180	3.00

[a]Model information notation refers to SNV (0 = off and 1 = on), SG (derivative order, polynomial order, and window length), and MC (0 = off and 1 = on).

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

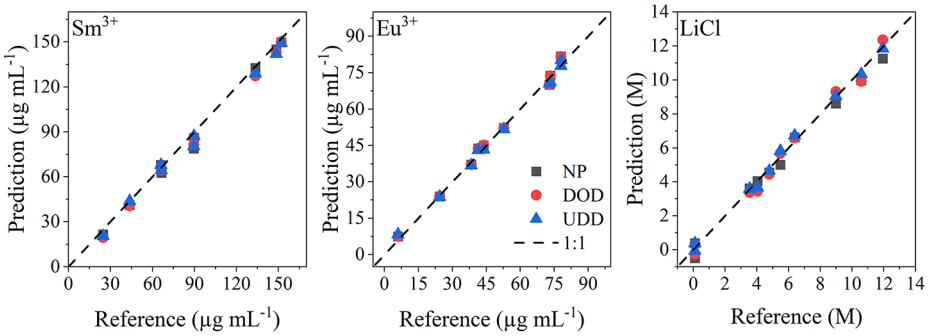


Figure 2. Comparison of NP, DOD, and UDD model predictions to reference values for Sm(III), Eu(III), and LiCl. The 1:1 dashed line represents a perfect match.

derivative. The LiCl UDD model used 6 LVs and had a lower RMSEP % of 4.00%. Details on the top-performing models are provided in Table 2. The various model predictions are compared to the known concentration values of the validation samples for each analyte in Figure 2.

For Sm(III), the DOD and UDD models resulted in lower RMSEP % values than the NP model, as well as a reduction in LVs. Additionally, the RMSEP % values for each top-performing DOD and UDD model appeared to be similar, likely a result of similar preprocessing strategies. However, RMSEP % values can be deceiving. A Tukey–Kramer test revealed no statistical difference between the three Sm(III) models at a 95% confidence level. Although this reveals that preprocessing may not be needed when modeling Sm(III) in this study, it does reveal that the DOD approach to preprocessing selection does result in a preprocessing combination which is not different from that of the UDD approach, offering significant time savings and favoring automated model construction.

For Eu, the UDD model resulted in a lower RMSEP % with fewer LVs, but the Tukey–Kramer test revealed there was no statistical difference between the three models. For LiCl, the RMSEP % values were decreased by the DOD- and UDD-selected preprocessing approaches, which were similar to derivative/smoothing SG filters. For LiCl, the Tukey–Kramer tests revealed that the preprocessed models were not statistically different from the NP model; however, the DOD and UDD models were. Additionally, the number of LVs used in the models decreased when preprocessing was applied. From this, we can infer that preprocessing may help to make models more resistant to issues with overfitting and robust to potential artifacts encountered in application versus the lab setting (e.g., baseline drift).

3.4. GA Feature Selection. Feature selection provides another strategy to improve model performance. In its simplest form, this can involve trimming the spectra prior to modeling to remove wavelength regions where there is little spectral response. At a much higher level, feature selection can involve filtering spectra to permit the modeling of only distinct spectral regions. Although PLSR models are well known for modeling data sets with high dimensionality, the removal of regions with weak correlations to the response matrix (i.e., concentrations) allows the model to better weigh the highly correlated regions and be less impacted by secondary effects.[27] Five 150-generation GAs were applied to the NP, top DOD, and top UDD models for each analyte. The top filter, out of the five GA runs, was selected for final testing. 150 generations were sufficient to achieve an optimal GA filter, and the GA-derived filters in all five runs generally agreed (see Figures S4 and S5). The number of LVs was held equal to those discussed previously as the GA-derived filters were developed.

The final GA filters were developed for the Sm(III), Eu(III), and LiCl NP models in Figure 3a–c, respectively. The colored regions correspond to the spectral features, which were selected by the GA to be regressed by PLSR models. These

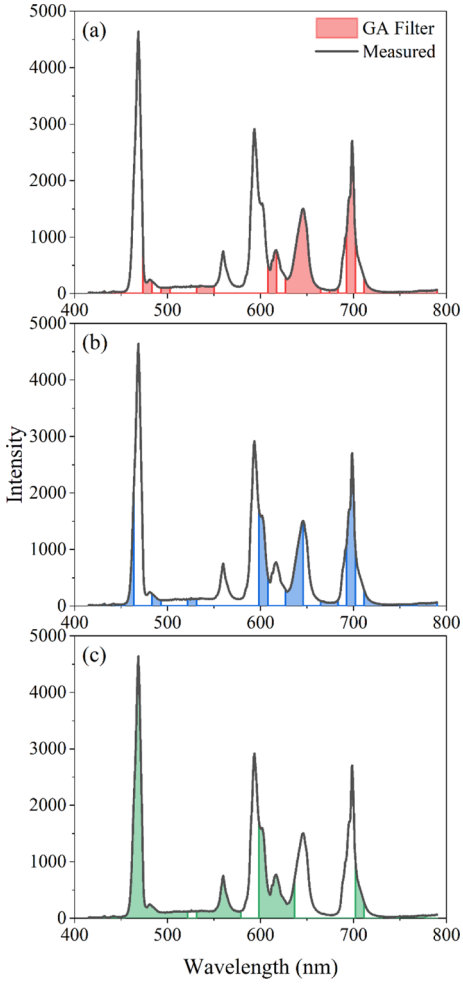


Figure 3. GA-selected features for (a) Sm(III), (b) Eu(III), and (c) LiCl to be regressed by PLSR models. The shown filters are for NP models to allow for better visualization without varying the preprocessing of the impacting signal shape. The spectrum shown is that of Sample 20.

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281–2290

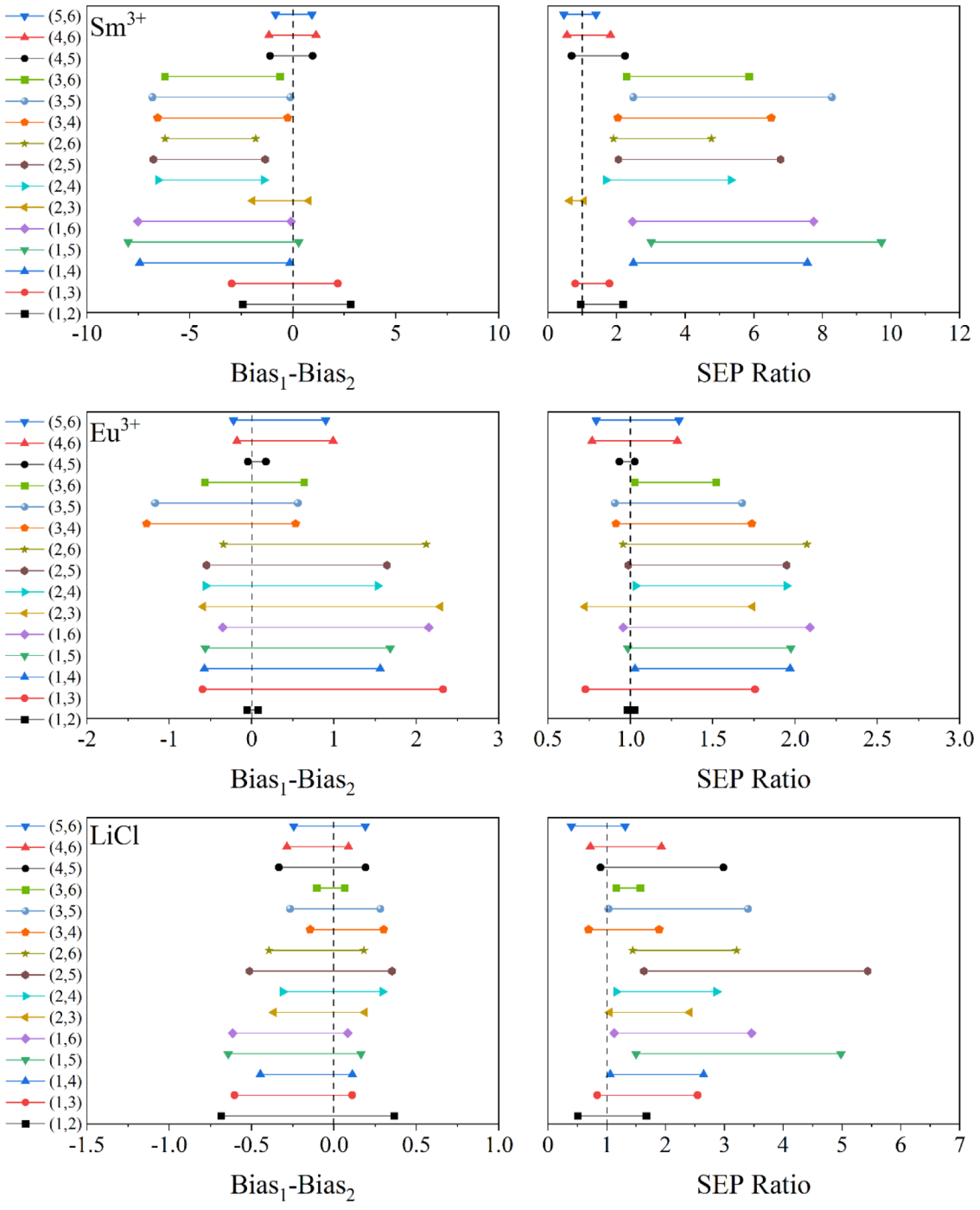


Figure 4. Confidence intervals for (a) bias and (b) SEP for all comparisons. Prediction performance is statistically similar between designs when the confidence interval crosses the dashed vertical line for bias and SEP. The model key is as follows: (1) NP, (2) DOD, (3) UDD, (4) GA-NP, (5) GA-DOD, and (6) GA-UDD, where (1,5) is a comparison of the NP and GA-DOD models.

regions correspond to analyte peaks, peaks of matrixed analytes that need to be incorporated into the models, or baseline regions that help the model adjust accordingly. Each species has a unique filter, suggesting that it would be unrealistic for users to manually select the optimal regions through trial and error. The filters were overlaid with the spectra from sample 20 to visualize shared analyte features but comparing the GA filters to the pure spectra shown in Figure 1 offers additional insight. The PLSR *x*-weights for the three GA-filtered NP models are shown in Figure S6.

In Figure 3a, the entire 650 nm peak was selected by the filter as it experienced little interference, and the regression coefficients indicated this as the major regression feature selected for the Sm(III) regression model. Several regions between 475 and 550 nm were also selected, likely to help the model normalize the spectra being regressed to a shifting baseline. The Eu(III) filter selected many similar regions to the Sm(III) filter. This is indicative of the similar emission signals and the models' ability to deconvolute these emissions. Here, the filters appeared to serve the purpose of removing signals

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

irrelevant to the model, thereby increasing the models' regression efficacy.

The upper portion of the 600 nm Sm(III) peak was selected as the lower half was convoluted with a Eu(III) peak. Likewise, the lower half of the 700 nm peak was selected as this peak was convoluted with a Eu(III) emission peak. At the center of the 700 nm peak, the regression coefficients logically reveal that the Eu(III) model weighing this emission is far greater than that of Sm(III). The Eu(III) filter contained the lower half of the Raman water band which also weighed moderately in its regression coefficients.

The LiCl filter selected greater portions of the spectrum. It utilized the entire Raman water band as expected (Figure S1), along with many other portions of Sm(III) and Eu(III) peaks. The *x*-weights (Figure S6) show that these emission peak tails and baselines were included in the LiCl model because they are highly sensitive to LiCl concentration (Figure 1). These results demonstrate the power of GA feature selection; not only does the method improve prediction performance, but it can also inform the user about the modeling interactions.

The GA-filtered models resulted in significant RMSEP % reductions for each species. The RMSEP values are provided in Table 2. For Sm(III), the GA filters reduced RMSEP % values by −84.1, −85.9, and −81.9% for the NP, DOD, and UDD models, respectively. The true RMSEP value for the GA-filtered Sm model is likely best represented by the analyte uncertainty, as described in Section 2.1. For Eu, the GA filters reduced RMSEP % values by −33.3, −31.9, and −19.4% for the NP, DOD, and UDD models, respectively. For LiCl, the GA filters reduced RMSEP % values by −47.8, −67.1, and −25.1% for the NP, DOD, and UDD models, respectively.

The Tukey−Kramer test statistically compared each of these models. The bias and SEP ratio confidence intervals are plotted in Figure 4. When comparing models, if both the bias and SEP ratio confidence intervals crossed the vertical reference lines, they were considered statistically equivalent. For Sm(III), the GA-filtered NP, DOD, and UDD models were statistically different from their unfiltered counterparts. These three models were not statistically different from one another, indicating that applying GA feature selection to NP, DOD, or UDD preprocessing strategies produced equivalent prediction performance. Applying feature selection may provide a more direct way to optimize a model compared to preprocessing. Even though the prediction performance was statistically equivalent when comparing NP to DOD and UDD, preprocessing adds the benefit of protecting the model against effects such as scattering or baseline drift. Therefore, the DOD approach is recommended to efficiently identify optimal preprocessing strategies.

For Eu(III), the GA-filtered models were statistically different from the unfiltered versions, but they were equivalent amongst themselves and to other unfiltered models which used different preprocessing strategies. This was due to the bias confidence intervals (Figure 4). These results argue for the direct application of the GA feature selection to the NP model. The results for the LiCl models were equivalent to those of the Sm(III) models.

These results validate previous findings that DOD and UDD arrive at equivalent models, although feature selection provides a far more impactful change in model performance.[20] Again, it should be stated that while preprocessing may be unnecessary in this case because it did not significantly lower RMSEP, many literature reports indicate that processing improves the robustness of a model to changing conditions, making it worthwhile to investigate.[3,4,18,20,22−26] These findings are crucial considering future automation in even more complex systems. UDD preprocessing selection requires ∼4 min of computation time vs. the ∼13 s required for the DOD preprocessing selection routine. As these routines become looped in a greater automation scheme, the difference in time becomes significant. To rapidly construct quantification models, the authors recommend investigating both an NP model with a GA-derived filter and a DOD preprocessing selection combined with GA feature selection.

The limits of detection (LODs) for the top-performing models were estimated to benchmark their performance for future comparisons. Here, a pseudounivariate approach (LOD_{pu}) was used to estimate the LODs for the PLSR models. This approach, as detailed by Ortiz et al., extends the univariate LOD recommendations of the International Union of Pure and Applied Chemistry (IUPAC) to multivariate models.[39] The LOD_{pu} definition and equation are provided in the Supporting Information, along with results from a univariate approach (Table S2). The LOD_{pu} values calculated for the top models were 2.95 μg mL^{-1} for Sm (DOD/GA), 1.66 μg mL^{-1} for Eu (NP/GA), and 0.287 M for LiCl (DOD/GA). It is important to note that these LOD_{pu} values are estimates and have been shown to be either consistent with or conservative when compared to more involved LOD confidence bands.[40]

4. CONCLUSIONS

In this work, multivariate regression models were developed to quantify Sm(III), Eu(III), and LiCl concentrations based solely on spectral variations (i.e., without recording lifetimes). A rapid workflow in Python combined DOD preprocessing selections and GA-selected spectral features to optimize prediction performance. The reduction in RMSEP shows the benefit of GA feature selection. The novel preprocessing and feature selection strategy will increase the breadth and ease of spectroscopy monitoring applications and improve their predictive capabilities. It is critical to streamline each aspect of the model development process, performance optimization, and implementation to fully adopt this approach in the challenging environments encountered in nuclear industry applications. These needs were addressed by reducing the quantity of materials in the training set and establishing an automated system to calibrate and optimize regression performance. In essence, the trial-and-error methodologies of the past evolved, with the help of intelligent design, into a resourceful method that will extend into future generations. The next steps will include applying this strategy to disparate data sets.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c06610.

It includes an extended discussion on preprocessing techniques, error propagation, statistical comparisons, DOD matrix, example spectra, luminescence decay curves, and details describing the performance of the GA. (PDF)

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290

AUTHOR INFORMATION

Corresponding Author

Luke R. Sadergaski − *Radioisotope Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States;* orcid.org/0000-0003-0248-2114; Email: sadergaskilr@ornl.gov

Authors

Hunter B. Andrews − *Radioisotope Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States;* orcid.org/0000-0002-2091-9415

Samantha K. Cary − *Radioisotope Science and Technology Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37830, United States;* orcid.org/0000-0003-0398-7106

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c06610

Author Contributions

The manuscript was written using the contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

This research was supported by the US Department of Energy Isotope Program, managed by the Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the U. S. Department of Energy under contract DE-AC05-00OR22725. This work used resources at the Radiochemical Engineering Development Center operated by the Oak Ridge National Laboratory.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors wish to thank Laetitia Delmau for helpful discussions regarding lanthanide speciation and Kyle Morgan for assistance. This work was supported by the ^{252}Cf Program at Oak Ridge National Laboratory.

REFERENCES

(1) Buckley, K.; Ryder, A. G. Applications of Raman spectroscopy in biopharmaceutical manufacturing: A short review. *Appl. Spectrosc.* **2017**, *71*, 1085−1116.

(2) Jin, H.; Lu, Q.; Chen, X.; Ding, H.; Gao, H.; Jin, S. The use of Raman spectroscopy in food processes: A review. *Appl. Spectrosc. Rev.* **2016**, *51*, 12−22.

(3) Kirsanov, D.; Rudnitskaya, A.; Legin, A.; Babain, V. UV-VIS spectroscopy with chemometric data treatment: an option for on-line control in nuclear industry. *J. Radioanal. Nucl. Chem.* **2017**, *312*, 461−470.

(4) Colle, J.-Y. C.; Manara, D.; Geisler, T.; Konings, R. J. M. Advances in the application of Raman spectroscopy in the nuclear field. *Spectrosc. Eur.* **2020**, *32*, 9−13.

(5) Sadergaski, L. R.; DePaoli, D. W.; Myhre, K. G. Monitoring the caustic dissolution of aluminum alloy in a radiochemical hot cell using Raman spectroscopy. *Appl. Spectrosc.* **2020**, *74*, 1252−1262.

(6) Sadergaski, L. R.; Myhre, K. G.; Delmau, L. H. Multivariate chemometric methods and Vis-NIR spectrophotometry for monitoring plutonium-238 anion exchange column effluent in a radiochemical hot cell. *Talanta Open* **2022**, *5*, 100120.

(7) *Lanthanide Luminescence: Photophysical, Analytical and Biological Aspects*; Hanninen, P., Harma, H., Eds.; Springer Ser. Fluoresc., Springer-Verlag: Berlin Heidelberg, 2010.

(8) Hasegawa, M.; Ohmagari, H.; Tanaka, H.; Machida, K. Luminescence of lanthanide complexes: From fundamental to prospective approaches related to water- and molecular-stimuli. *J. Photochem. Photobiol., C* **2022**, *50*, 100484.

(9) Melby, L. R.; Rose, N. J.; Abramson, E.; Caris, J. C. Synthesis and fluorescence of some trivalent lanthanide complexes. *J. Am. Chem. Soc.* **1964**, *86*, 5117−5125.

(10) Cho, U.; Chen, J. K. Lanthanide-based optical probes of biological systems. *Cell Chem. Biol.* **2020**, *27*, 921−936.

(11) Moore, E. G.; Samuel, P. S.; Raymond, K. N. From antenna to assay: Lessons learned in lanthanide luminescence. *Acc. Chem. Res.* **2009**, *42*, 542−552.

(12) Choppin, G. R. In *Lanthanide Probes in Life, Chemical and Earth Sciences: Theory and Practice*; Bunzli, J.-C. G., Choppin, G. R., Eds.; Elsevier Science Publishers B.V.: Amsterdam, 1989; pp 1−41.

(13) Ueno, Y.; Sasaoka, N.; Morishige, K.; Shigematsu, T.; Nishikawa, Y. Fluorescence properties of samarium (III) and europium (III) in mineral acids, and their use for fluorometry. *Bunseki Kagaku* **1988**, *37*, 263−268.

(14) Löble, M. W.; Keith, J. M.; Altman, A. B.; Stieber, S. C. E.; Batista, E. R.; Boland, K. S.; Conradson, S. D.; Clark, D. L.; Lezama Pacheco, J.; Kozimor, S. A.; Martin, R. L.; Minasian, S. G.; Olson, A. C.; Scott, B. L.; Shuh, D. K.; Tyliszczak, T.; Wilkerson, M. P.; Zehnder, R. A. Covalency in lanthanides. An x-ray absorption spectroscopy and density functional theory study of LnCl6x− (x = 3, 2). *J. Am. Chem. Soc.* **2015**, *137*, 2506.

(15) Arisaka, M.; Kimura, T.; Suganuma, H.; Yoshida, Z. Direct evidence for enhanced inner-sphere chloro complexation of Eu(III) and Cm(III) in anion exchange resin phase studied by time-resolved laser-induced fluorescence spectroscopy. *Radiochim. Acta* **2002**, *90*, 193−197.

(16) Moulin, C.; Decambox, P.; Mauchien, P.; Pouyat, D.; Couston, L. Direct uranium(VI) and nitrate determinations in nuclear reprocessing by time-resolved laser-induced fluorescence. *Anal. Chem.* **1996**, *68*, 3204−3209.

(17) Sadergaski, L. R.; Andrews, H. B. Simultaneous quantification of uranium(VI), samarium, nitric acid and temperature with combined ensemble learning, laser fluorescence, and Raman scattering for real-time monitoring. *Analyst* **2022**, *147*, 4014−4025.

(18) Casella, A. J.; Levitskaia, T. G.; Peterson, J. M.; Bryan, S. A. Water O−H stretching Raman signature for strong acid monitoring via multivariate analysis. *Anal. Chem.* **2013**, *85*, 4120−4128.

(19) Sadergaski, L. R.; Toney, G. K.; Delmau, L. D.; Myhre, K. G. Chemometrics and experimental design for the quantification of nitrate salts in nitric acid: Near-infrared spectroscopy absorption analysis. *Appl. Spectrosc.* **2021**, *75*, 1155−1167.

(20) Sadergaski, L. R.; Hager, T. J.; Andrews, H. B. Design of experiments, chemometrics, and Raman spectroscopy for the quantification of hydroxylammonium, nitrate, and nitric acid. *ACS Omega* **2022**, *7*, 7287−7296.

(21) Bondi, R. W.; Igne, B.; Drennen, J. K., III; Anderson, C. A. Effect of experimental design on the prediction performance of calibration models based on near-infrared spectroscopy for pharmaceutical applications. *Appl. Spectrosc.* **2012**, *66*, 1442−1453.

(22) Mishra, P.; Biancolillo, A.; Roger, J. M.; Marini, F.; Rutledge, D. N. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *Trends Anal. Chem.* **2020**, *132*, 116045.
(23) Engel, J.; Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Downey, G.; Blanchet, L.; Buydens, L. M. C. Breaking with trends in pre-processing? *Trends Anal. Chem.* **2013**, *50*, 96−106.
(24) Rinnan, A. Pre-processing in vibrational spectroscopy—When, why and how. *Anal. Methods* **2014**, *6*, 7124−7129.
(25) Gerretzen, J.; Szymańska, E.; Jansen, J. J.; Bart, J.; van Manen, H.-J.; van den Heuvel, E. R.; Buydens, L. M. C. Simple and effective way for data preprocessing based on design of experiments. *Anal. Chem.* **2015**, *87*, 12096−12103.
(26) Storey, E. E.; Helmy, A. S. Optimized preprocessing and machine learning for quantitative Raman spectroscopy in biology. *J. Raman Spectrosc.* **2019**, *50*, 958−968.
(27) Andrews, H. B.; Myhre, K. G. Quantification of lanthanides in a molten salt reactor surrogate off-gas stream using laser-induced breakdown spectroscopy. *Appl. Spectrosc.* **2022**, *76*, 877−886.
(28) Myakalwar, A. K.; Spegazzini, N.; Zhang, C.; Anubham, S. K.; Dasari, R. R.; Barman, I.; Gundawar, M. K. Less is more: Avoiding the LIBS dimensionality curse through judicious feature selection for explosive detection. *Sci. Rep.* **2015**, *5*, 13169.
(29) Robinson, S. M.; Benker, D. E.; Collins, E. D.; Ezold, J. G.; Garrison, J. R.; Hogle, S. L. Production of Cf-252 and other transplutonium isotopes at Oak Ridge National Laboratory. *Radiochem. Acta* **2020**, *108*, 737−746.
(30) Smucker, B.; Krzywinski, M.; Altman, N. Optimal experimental design. *Nat. Methods* **2018**, *15*, 559−560.
(31) Zahran, A.; Anderson-Cook, C. M.; Myers, R. H. Fraction of design space to assess prediction capability of response surface designs. *J. Qual. Technol.* **2003**, *35*, 377−386.
(32) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B. Scikit-Learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(33) Westad, F.; Marini, F. Validation of chemometric models—A tutorial. *Anal. Chim. Acta* **2015**, *893*, 14−24.
(34) Fearn, T. Comparing standard deviations. *NIR News* **1996**, *7*, 5−6.
(35) Cederkvist, H. R.; Aastveit, A. H.; Næs, T. A comparison of methods for testing differences in predictive ability. *J. Chemom.* **2005**, *19*, 500−509.
(36) Nehlig, A.; Elhabiri, M.; Billard, I.; Albrecht-Gary, A. M.; Lützenkirchen, K. Photoexication of europium(III) in various electrolytes: Dependence of the luminescence lifetime on the type of salts and the ionic strength. *Radiochim. Acta* **2003**, *91*, 37−44.
(37) Tanaka, F.; Yamashita, S. Luminescence lifetimes of aqueous europium chloride, nitrate, sulfate, and perchlorate solutions. Studies on the nature of the inner coordination sphere of the europium(III) ion. *Inorg. Chem.* **1984**, *23*, 2044−2046.
(38) Czitrom, V. One-factor-at-a-time versus designed experiments. *Am. Stat.* **1999**, *53*, 126−131.
(39) Ortiz, M. C.; Sarabia, L. A.; Herrero, A.; Sánchez, M. S.; Sanz, M. B.; Rueda, D.; Giménez, M. E.; Meléndez, M. E. Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares. *Chemom. Intell. Lab. Syst.* **2003**, *69*, 21−33.
(40) Allegrini, F.; Olivieri, A. C. IUPAC-Consistent Approach to the Limit of Detection in Partial Least-Squares Calibration. *Anal. Chem.* **2014**, *86*, 7858−7866.

https://doi.org/10.1021/acsomega.2c06610
ACS Omega 2023, 8, 2281−2290