Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/6181--18-9830

Development of Improved Automated GC-MS Analysis Software: Final Report

Thomas N. Loegel
Jeffrey A. Cramer
Mark H. Hammond

Navy Technology Center for Safety & Survivability
Chemistry Division

December 14, 2018

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)*
14-12-2018

2. REPORT TYPE
Memorandum Report

3. DATES COVERED *(From - To)*
1 Oct 2017 - 30 Sept 2018

4. TITLE AND SUBTITLE

Development of Improved Automated GC-MS Analysis Software: Final Report

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Thomas N. Loegel, Jeffrey A. Cramer and Mark H. Hammond

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
61-9251-G-8-5

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Research Laboratory
4555 Overlook Avenue, SW
Washington, DC 20375-5344

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/6181--18-9830

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Defense Logistics Agency Energy
8725 John J. Avenue, SW
Fort Belvoir, VA 22060-6222

10. SPONSOR / MONITOR'S ACRONYM(S)

DLA-E

11. SPONSOR / MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A: Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

FCAST is a comprehensive software package that extracts a wide variety of information from GC-MS data using mathematical, statistical, and chemometric modeling strategies, which isinvaluable when comprehensive fuel analyses are required but limited sample volumes are available. FCAST is a self-contained tool for rapid fuel identification and characterization and has been an invaluable resource in NRL fuel-based research programs, but the previous iteration of FCAST required additional development before DLA Energy could reliably employ it to assure fuel quality throughout distribution chains and collect on-site compositional data from fuel samples for comparisons to forensics library-derived compositional data. The objective of this work was thus to develop an improved FCAST, with a focus on refining and implementing GC-MS peak deconvolution and recognition methodologies (developed during an FY17 DLA Energy-funded study at NRL), optimizing software functionality to account for newly deconvolved mass spectral information, and automating front-end parameter selection.

15. SUBJECT TERMS

Gas Chromatography-Mass, Spectrometry (GC-MS), Fuel Composition and Screening Tool (FCAST), Software Automation, Peak Deconvolution, Fuels

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Jeffrey A. Cramer
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	Unclassified Unlimited	40	**19b. TELEPHONE NUMBER** *(include area code)*
Unclassified Unlimited	Unclassified Unlimited	Unclassified Unlimited			(202) 404-3419

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

This page intentionally left blank.

CONTENTS

1.0 Introduction 1
2.0 FCAST Chromatographic Peak Deconvolution 2
2.1 EWFA-MCR Algorithm Summary 4
2.2 Software Automation Considerations of EWFA-MCR Algorithm 9
2.3 Mass Channel Analysis 13
3.0 FCAST Software Automation 15
3.1 Automation of Comparison Sub-Routines 15
3.2 Default Peak Area Threshold Selection 27
3.3 Other Automation-Friendly Optimizations 29
3.4 Software Optimizations 31
4.0 Conclusions 31
5.0 Acknowledgements 32
6.0 Literature Cited 32

FIGURES

1. Visual representation of the EWFA-MCR peak deconvolution algorithm ……4

2. Summaries of the total numbers of compounds found that would have been expected to change between pairings of three surrogate fuels, utilizing multiple standard deviation multiplication constants and three different F-ratio transformations, including the control, which skipped logarithmic transformation ……19

3a. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 30 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 0.5 ……20

3b. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 0.5 ……21

3c. Class-sorted TIC results, obtained for the Surrogate Fuel 30 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 0.5 ……22

4a. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 30 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 6.0 ……23

4b. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 6.0 ……24

4c. Class-sorted TIC results, obtained for the Surrogate Fuel 30 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 6.0 ……25

5. Sum absolute TIC differences between classes, for compounds known to be different between classes, with losses being reported relative to the results obtained when a constant of 0.5 is employed……27

TABLES

1. Deconvolution outputs obtained from both the AMDIS software and NRL's automated algorithms 8

2. UVE-PLS prediction results obtained for thirty-five fuel properties with various MF thresholds. Modeling metrics in each row indicative of the most accurate results in green 11-12

3. Summary of Table 2 results, obtained by tabulating the number of green-highlighted, and thus most accurate, modeling metrics found in Table 2 13

4. Number of compounds reported for standard Agilent ChemStation analysis, vs FCAST simple, mass channel and EWFA implementations 14

5. Percent compositional contents (v/v) of the three surrogate fuels evaluated to develop the automated F-ratio threshold determination methodology 17

6. UVE-PLS prediction results obtained for five fuel properties while utilizing various peak area thresholds. Most favorable results of three replicates reported for each value. The modeling metric(s) in each row indicative of the most accurate modeling results are highlighted in green 28

ABBREVIATIONS

ALS	Alternating Least Squares
AMDIS	Automated Mass Spectral Deconvolution and Identification System
ANOVA	Analysis of Variance
CUMPRESS	Cumulative Predicted Residual Error Sum of Squares
DLA	Defense Logistics Agency
DoD	Department of Defense
EFA	Evolving Factor Analysis
EPA	Environmental Protection Agency
EWFA	Evolving Window Factor Analysis
FCAST	Fuel Composition and Screening Tool
FSII	Fuel System Icing Inhibitor
GC	Gas Chromatography
GC-MS	Gas Chromatography – Mass Spectrometry
LOO-CV	Leave-One-Out Cross -Validation
LV	Latent Variable
MCR	Multivariate Curve Resolution
MF	Match Factor
MS	Mass Spectrometry
NIH	National Institute of Health
NIST	National Institute of Standards and Technology
NRL	Naval Research Laboratory
PLS	Partial Least Squares
RMSEP	Root Mean Square Error of Prediction
SVD	Singular Value Decomposition
TIC	Total Ion Chromatograph
UVE-PLS	Uninformative Variable Elimination Partial Least Squares

1.0 Introduction

NRL has, for some time, been engaged in the development of analytical methods to perform fuel quality surveillance within Navy fuel supply chains. This approach offers significant advantages over the current state-of-the-art, not only by reducing the time and manpower required to routinely measure critical specification properties, but also by providing a means by which to initially identify and characterize discrete fuel handling and performance challenges in an efficient manner. It is already known that a great deal of information regarding fuel composition and performance can be obtained from GC-MS,[1,2,3] rendering it a useful analytical technique upon which to base an in-depth analysis strategy.

Both fuel-based performance property modeling and fuel failure investigations at least potentially require the proper discrimination of hundreds, if not thousands, of discrete compounds, and chromatography combined with fully mass analyzed MS data provides this level of discrimination, as GC-MS's pre-existing status as a primary tool for the analysis and compositional characterization of mobility fuels already attests. Earlier work[4,5] performed at NRL laid the basic groundwork for the use of GC-MS in fuel modeling before the development of an effective and robust modeling strategy,[6] based on applying the PLS variant UVE-PLS[7] to data abstractions referred to herein as metaspectra. A metaspectrum is effectively a streamlined list of compounds, along with associated peak areas, derived from the original GC-MS data by comparing mass fragment data from individual chromatographic time slices with reference data from the NIST/EPA/NIH Mass Spectral Library.[8] This is useful in the context of predictive fuel property modeling because translating two-dimensional data (with a chromatographic axis and a mass spectral axis) to one-dimensional data (with a single compound identity axis) via library matching allows robust chemometric modeling techniques such as UVE-PLS to be utilized, because, as implied by their name, sets of metaspectra can be mathematically modeled in much the same way that sets of more standard spectra can be.

With in-depth fuel analysis as an explicitly targeted goal, NRL internally developed[9,10] FCAST, a comprehensive software package, to extract a wide variety of information from GC-MS data via mathematical, statistical, and chemometric modeling strategies, including detailed compositional assessments and calculated distillation curves for individual fuel samples, as well as composition-based comparisons of fuel sample pairs. The aforementioned UVE-PLS metaspectral modeling strategy is specifically utilized within FCAST to estimate critical fuel performance properties. Each of these software features can individually be invaluable when comprehensive fuel analyses are required but only limited sample volumes are available, and collectively provide a self-contained methodology for rapid fuel identification and characterization. The software has been an invaluable resource in the context of NRL fuel-based research programs, and its effectiveness has been proven during applications running the gamut from routine analyses to critical investigations into discrete fuel failures.

Manuscript approved December 14, 2018.

DLA Energy has made efforts to expand upon in-house fuel analytics in order to more ably address fuel analysis challenges in an efficient manner. GC-MS instrumentation, capable of providing detailed compositional information, is already a staple of most practicing fuel laboratories. Automated and high-throughput GC-MS analysis strategies and associated data analysis software packages, such as FCAST, can enhance the capabilities of all practicing DLA Energy laboratories to rapidly address a wide array of fuel analytic challenges, simultaneously, with minimal operator training. To this end, FCAST underwent additional development during the present work to allow the software to robustly produce fuel characterization information in an automated fashion, i.e. in the absence of *a priori* operator inputs, allowing results obtained from the software to be reliably compared across multiple laboratories while also minimizing the likelihood of operator error.

Moreover, recent in-house advances in GC-MS peak deconvolution have allowed for significant increases in the quality of the compositional information that can be obtained from GC-MS data sets. Fuel chromatography is inherently limited by the high complexity of petroleum fuel compositions, and, in practice, almost no individual fuel constituents are fully resolved from other components. Due to an insufficient peak capacity for the large number of individual components that need to be assessed within time and chromatographic efficiency constraints, as well as an insufficient resolving power of the stationary phase in the gas chromatography column relative to the many structurally similar isomers or homologs present in typical fuels, co-eluting component peaks will tend to overlap to a non-trivial degree along a data set's retention time axis. Depending upon the degree of overlap, the mass spectral database search algorithms employed by FCAST could respond to this co-elution by either completely discarding insufficiently resolved chemical information, or incorrectly assigning specious component identities to mangled mass spectral data. Obviously, then, there were significant improvements to be had by fully implementing in-house deconvolution algorithms into FCAST, and additionally accounting for their effects on associated software operations, such as fuel property predictions.

This document serves as a complement to previous NRL Memorandum reports[11,12] as it details some of the additional features incorporated into the latest version of FCAST. However, in order to maintain a streamlined software-based reporting series that can double as an up-to-date user manual, a proper FCAST update will also appear as its own NRL Memorandum report[13] concurrent with this report.

2.0 FCAST Chromatographic Peak Deconvolution

For many reasons, not least of which being that predictive fuel property models are generally only as good as the GC-MS data upon which they are based, the performance boundaries of fuel chromatography become a significant limiting factor for the downstream performance of fuel characterization tools such as FCAST. While multidimensional approaches, longer columns and

slower heating rates can offer some benefits, they will not necessarily fully resolve co-eluting fuel compounds, let alone in a manner suitable for practical, real-world fuel analysis applications.

When multiple compounds co-elute, their chromatographic peaks overlap, and the mass spectral data used for library comparisons will thus contain contributions from all co-eluting compounds. Depending on the degree of overlap, library searches either yield incorrect peak assignments due to comingled mass spectral data, or result in information loss via the elimination of poorly resolved mass spectral data within overlapped peaks. The end user will not see the discarded compounds in the library search report, but will see occurrences of misidentified compounds. In the case of FCAST, at least, these convolved mass spectra can theoretically be discriminated against by means of a NIST/EPA/NIH Mass Spectral Library search-based goodness-of-fit metric known as the match factor, or MF. Briefly, the library searches used to produce compound identifications are conducted by comparing the distances of the experimental mass spectrum from any given library mass spectrum, after data preprocessing, in a multidimensional m/z space. These identifications utilize MF values, which scale from 0 to 1000 in the present work and are inversely proportional to the distances indicated previously, to determine how closely any given mass spectrum matches with any given library entry, and the match resulting in the highest MF value, indicating the closest shape-based correspondence, is typically reported to the practitioner as the most likely identification result. However, by virtue of the sheer number of discrete mass spectral data collections within a typical GC-MS data set, fuel assessments at NRL have previously uncovered significant numbers of instances in which high (i.e. favorable) MF values were obtained from nonsensical library matches. Co-elution is also responsible for the reporting of multiple instances of the same compound, at significantly different retention times, throughout a single analysis, as the mingling of mass spectra from co-eluting compounds can exacerbate instances of poor discrimination between similar compounds or isomers in the NIST/EPA/NIH Mass Spectral Library.

Previous work at NRL[14] indicated that the deconvolution methodology utilized by the freely available AMDIS software package[15] was sub-optimal for thorough fuel analysis efforts, which suggested the necessity of a more customized peak deconvolution solution. While the challenge of mathematically deconvolving co-eluting mass spectral data is well-known and has been addressed previously in the literature using many different approaches,[16,17] the complexity and unpredictability of fuel compositions leaves both more basic techniques and techniques that rely upon a priori knowledge[18,19,20] with a somewhat limited applicability in the context of realistic fuel modeling applications. In a similar vein, applying an individually customized deconvolution strategy to every single sample's data within an overall data set prior to modeling is impractical. What was deemed necessary, instead, was a robust technique that is as generally applicable as possible, across both the entirety of a single fuel's GC-MS data and entire GC-MS data sets, such that it can be relied upon to extract meaningful chemical information across the diverse fuel data sets in an automated fashion.

2.1 EWFA-MCR Algorithm Summary

The algorithmic development of Evolving Window Factor Analysis (EWFA)[21] and Multivariate Curve Resolution (MCR)[22] into a novel, integrated technique suitable for the automated, unsupervised deconvolution of GC-MS fuel data has been covered in greater detail elsewhere.[23,24,25] At a conceptual level, EWFA is used to track the appearances, disappearances, and overlap of underlying data factors, known as loadings, across the retention time axes of multi-peak chromatographic data; MCR is used to refine the shapes of these loadings to more accurately reflect underlying chemical phenomena; and NIST/EPA/NIH library searches are used to produce MF values that serve as quality control metrics by which to filter out superfluous data artifacts, leaving only the loadings that can serve as meaningful representations of the individual mass spectra convolved within the parent GC-MS data. The basic outline of the combined EWFA-MCR algorithm developed to address the fuel peak deconvolution task at hand can be found in Figure 1.

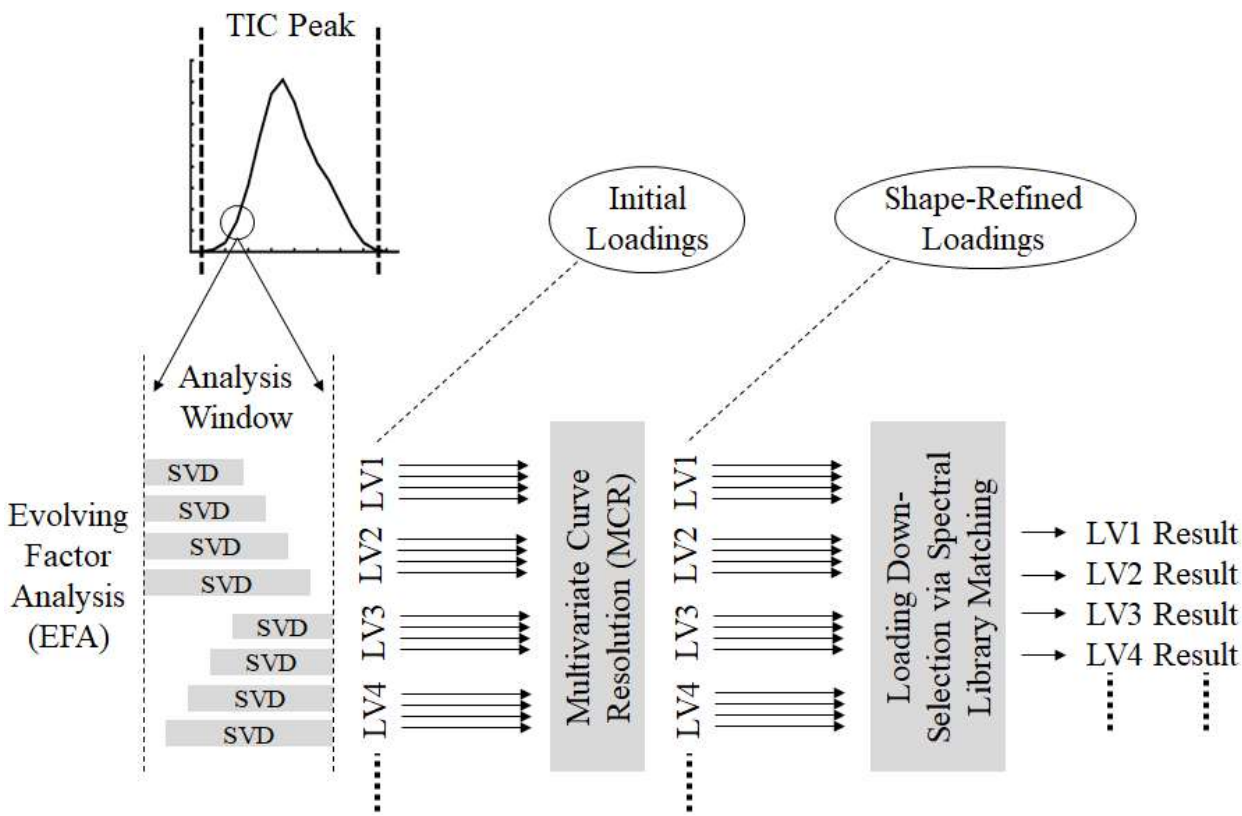

Figure 1. Visual representation of the EWFA-MCR peak deconvolution algorithm.

EWFA-MCR algorithm description. The algorithm proceeds by first locating peaks in the TIC produced from the original GC-MS data set and defining corresponding analysis window sizes as the number of variables corresponding to the width of the individual peak to be analyzed plus two variables, with even-numbered widths being rounded up by an additional variable to ensure the existence of central variables in subsequent analysis windows. A window of the defined size is then moved across the peak, with the mandated central variable of the window beginning at one end of the peak along the retention time axis and ending at the opposite end, allowing for additional retention time variables outside of the TIC peak to inform deconvolution procedures to further account for cross-peak chemical information.

At each possible window location, a full evolving factor analysis (EFA) operation is performed within the window, in the form of repeated SVD operations performed on increasingly large portions of a data matrix, proceeding in both the forward and reverse directions. In the forward direction, SVD is performed on a data subset, initially defined starting from the first row/column, which increases in size by one row/column per SVD operation until the last row/column is included in the SVD. In the reverse direction, the same stepwise increase in data subset size proceeds in the opposite direction from the last row/column instead. These overlapping EFA operations inherently include the thorough interrogation of contiguous data subsets within individual peaks, which ably allows for the identification of compounds that do not themselves contribute to any given peak's entire TIC area. It should also be noted here that this methodology allows the combined EWFA-MCR algorithm to function without a pre-defined preset value for either parent peak width or component peak width, thus allowing for a level of automation-friendly flexibility unavailable to deconvolution methodologies and/or software applications requiring such preset values.

SVD mathematically breaks a given data matrix down into its underlying LVs, which can be represented as scores, loadings, and singular values. The core decomposition of SVD can be represented by the following equation:

$$\mathbf{R} = \mathbf{USV}^{\mathrm{T}} \tag{1}$$

In this equation, **R** is the original data set (in this case, the portion of the GC-MS data being analyzed), $\mathbf{V}^{\mathrm{T}}$ is the transposed matrix of loadings that the developed algorithm uses to estimate the shapes of deconvolved mass spectral data, and **US** is the product of the scores and singular values that, combined, indicate the significance of the corresponding loadings to the variance within the original data set.

The EWFA portion of the combined algorithm relies upon these multiple executions of SVD for two primary uses. First, the absolute values of the loadings can be interpreted along the original mass spectral data axis to assess underlying sources of mathematical variance which should, in turn, correlate to underlying sources of chemical variance. Second, the **US** product can be

subjected to a seemingly trivial threshold value of $1x10^{-15}$ to ensure that deconvolved mass spectral loadings have at least a minimal significance to the original data set before being collected during the course of the EWFA-MCR algorithm, thus reducing the number of superfluous compound identifications.

Prior to interpreting the loadings as actually being informative of sources of chemical variance, however, individual SVD loading results are refined by means of MCR, applied to the data as it exists within the analysis window, with initial loading results serving as the initial estimates to be refined. The MCR portion of the overall algorithm performs ALS-based refinements repeatedly over the course of 1,000 iterations, though, as a stringent test for convergence, the algorithm is also designed to terminate early if an iteration produces a set of results whose maximum difference from the previous set of results is only 10^{-10} of the maximum result value, or if the average root mean square difference between the data as reconstructed from the refined results and the original data subset itself is only 10^{-10}.

The number of loadings that can be obtained from any given EWFA window is only limited by the size of the window itself. This means that large windows can produce large numbers of loadings, and all of these loadings could theoretically be subjected to further evaluations. However, not only would this be time consuming, but it is not likely that loadings associated with lower significances, as described above, will yield chemically meaningful information. An output constraint was thus implemented in the EWFA-MCR algorithm in which only chemical compound identifications obtained from the three most significant loadings obtained from any given window would be further considered for the purposes of overall chemical profiling. This substantially reduces the total number of SVD operations (and, incidentally, overall algorithm calculation times) while still allowing for substantial data deconvolution capabilities.

The shape refined loadings thus far collected are then interpreted as if they actually were the mass spectra required to identify individual components by means of NIST/EPA/NIH Mass Spectral Library searches. This library matching provides a built-in quality control methodology, an internal, MF-based metric that can be used to interrogate loadings in a manner more directly relevant towards characterizing complex compositions. As might be apparent from considering the large number of SVD operations performed for any given TIC peak, such an extensive procedure produces far more information than is useful in the present context. In the MF-based algorithm, only the results corresponding to the highest MF value are maintained for any given retention time/LV combination. The maintenance of retention time-specific results also allows for the subsequent derivation of retention time information for deconvolved component sub-peaks, information that can inform downstream FCAST operations such as fuel property modeling. However, these deconvolved results can also be further summarized to apply to the respective TIC peaks from which they were derived, providing proportional information regarding peak

composition. This proportional peak composition information, in turn, can be used to assign the proportional contributions of identified compounds to the original TIC peak areas.

FCAST itself will report to the end-user any compound identifications that represent at least some user-selected threshold of the data's TIC area, though such values falling below 0.1% will be reported as <0.1%. The compound identifications obtained via the EWFA-MCR deconvolution algorithm in the present work were initially assigned a threshold value of 0.001%, which served as an additional safeguard against superfluous compound identifications. It will be shown later in this report that such a threshold value also makes an effective default FCAST threshold value in a more general sense.

Comparison of EWFA-MCR peak deconvolution results with AMDIS software outputs. The automated peak deconvolution algorithm thus developed at NRL requires hours to fully deconvolve a typical GC-MS data set when it is run through FCAST, not primarily as a consequence of its repeated performance of SVD across various subsets within the entirety of such a data set, but primarily as a consequence of the large number of times it must consult a NIST mass spectral database to evaluate SVD results. This is as opposed to the minutes required to deconvolve these same data sets using the freely available AMDIS software, which is commonly utilized in GC-MS data deconvolution operations. Because a generally comprehensive understanding of fuel composition can be found to be very important in the context of fuel investigations, the longer analysis times associated with the NRL-developed peak deconvolution algorithm are defensible, provided that the additional analysis times required by the EWFA-MCR algorithm are, in fact, providing more comprehensive results than can be achieved using AMDIS. As indicated earlier in this report, previous work at NRL has indicated as much, but only indirectly, i.e. because AMDIS was not providing sufficiently comprehensive results during preliminary work, NRL developed its own data deconvolution strategies, and thus a side-by-side comparison was done during the present work. As the AMDIS software is sometimes relied upon for peak deconvolution in non-AMDIS software applications as well, any comparison between NRL-developed peak deconvolution algorithms and AMDIS outputs might also productively inform comparisons between FCAST and said software applications.

To perform a meaningful and realistic deconvolution comparison, AMDIS was used to analyze twelve different petrochemical fuel samples, covering a wide range of jet and diesel fuel grades and fuel compositions, selected from our internal GC-MS data archives. When the AMDIS software allowed for as much, analysis settings were adjusted to closely mimic the parameters either found to be optimal at NRL (see sections 2.2 and 3.2 of the present report) or otherwise utilized during the EWFA-MCR algorithm's development work: the mass range to examine was kept at 35-356 m/z, the Match Factor (MF) threshold was set to 750, the peak threshold was set at 0.001%, the type of analysis was defined as "Simple" (with no internal standards or other calibration inputs), and the NIST 2011 Mass Spectral Library was utilized. Otherwise, the AMDIS

software's analysis setting were adjusted in an attempt to maximize the number of components expected to be found: no adjacent peak subtractions were employed, "High" resolution was utilized (the highest setting), and "Very High" sensitivity was utilized (again, the highest setting). Because preliminary work did not indicate that shape requirements had a consistent impact on component predictions, these requirements were simply set to "Medium." Finally, because individual chemical components might possess one of a large number of widths in a convolved GC-MS data set, component width values of 8, 16, and 32 (the maximum allowed value) were each evaluated and are reported upon separately herein. While it might be possible to aggregate the results obtained when utilizing multiple component width values in an attempt to identify components that might not appear when using any single given value, such an operation would not typically be performed by an end-user of the AMDIS software.

In Table 1, the AMDIS results thus obtained are compared to the deconvolved results found when utilizing the developed EWFA-MCR peak deconvolution algorithm. It can be seen in this table that the number of unique compounds that can be identified via the AMDIS software, regardless of component width utilized, are, on average, lower than the number of unique compounds identified after utilizing the EWFA-MCR algorithm, thus indicating the enhanced utility of the EWFA-MCR algorithm thus developed. AMDIS only outperforms EWFA-MCR by greater than 12 compounds identifications in the case of a single fuel.

Grade	unique compounds found before deconvolution via FCAST	unique compounds reported by AMDIS (component width = 8)	unique compounds reported by AMDIS (component width = 16)	unique compounds reported by AMDIS (component width = 32)	unique compounds identified via EWFA-MCR
JP-5	108	250	247	260	342
JP-5	89	228	231	226	333
JP-8	122	286	278	273	376
JP-8	129	347	338	353	394
Jet A	119	294	295	299	366
Jet A	80	232	237	243	324
F-76	101	284	289	282	367
F-76	72	199	207	205	293
MGO	111	386	407	410	398
MGO	117	348	368	359	387
Alt. Diesel (CHCD)	112	337	367	387	383
Alt. Diesel (HEFA)	110	250	249	255	228

Table 1. Deconvolution outputs obtained from both the AMDIS software and NRL's automated algorithms.

2.2 Software Automation Considerations of EWFA-MCR Algorithm

While the majority of steps taken during the present work to enhance FCAST's automated analysis capabilities will be reported upon in a separate section of the present report, specific work regarding the implementation of the EWFA-MCR algorithm into the current iteration of FCAST will be reported upon here. It should be noted, however, that this work also informed the selection of FCAST's default MF threshold value, which has implications outside of fuel property modeling.

Improved Fuel property modeling via UVE-PLS. Before discussing fuel property modeling, some details must be conveyed regarding how said modeling was performed. UVE-PLS was employed, as opposed to basic PLS, because the technique was found during previous GC-MS data modeling work to more ably focus subsequent models on the chemical compounds that would be expected to influence the fuel properties being predicted. In UVE-PLS, the amount of relevant information possessed by individual compounds in a metaspectral data set is determined by using regression coefficients derived from the overall stability found during LOO-CV. To determine this stability, a number of randomized variables equal to a third of the total possible size of any given metaspectrum are added to the original metaspectral data. After this, the actual variables determined to be as inconsistently informative as their randomized counterparts during cross-validation are eliminated from the final model. In the present work, being inconsistently informative is defined as having a regression coefficient average/regression coefficient standard deviation ratio lower than that obtained for 85% of the random variables. The value of 85% was chosen to maintain consistency with the statistical F-test which, when applied to interim CUMPRESS results with an 85% confidence interval and utilizing a maximum of 10 LVs, was used to select a number of LVs to be employed for the final UVE-PLS models that would minimize model overfitting (i.e. models too well-calibrated to the initial training data that they are no longer optimal for use with non-training data).

Investigation into the default match factor to employ for fuel property modeling. Obviously, the use of data deconvolution significantly impacts the compositional information contained within the metaspectra produced downstream of the compound identifications performed via library matching. The UVE-PLS models to be employed for the FCAST fuel property modeling thus were required to be reconstructed after deconvolving the metaspectra used as the training data. One aspect of this model reconstruction was a reconsideration of the MF threshold to employ. In other words, an updated decision was required as to how high a quality any given loading-based library match would be required to be in order to contribute to a corresponding metaspectrum. This is because the deconvolved loading information would be expected to be of a different quality than the original convolved mass spectral data. Although it would be possible to assign an individually optimized MF threshold for each model, it is believed that tailoring modeling solutions to this extent would unacceptably risk model overfitting. In addition, the MF threshold settled upon via this portion of the present work would be expected to be an acceptable default value for FCAST in a general sense.

The UVE-PLS modeling results obtained for thirty-five fuel properties can be found in Table 2. The results shown in this table are actually the most favorable results found from among three modeling operation replicates, as the variances inherent in the use of random variables for UVE-PLS can impact individual modeling results. These results include RMSEP error values (lower values indicate lower errors and, hence, more accurate models), correlation coefficient (R^2) values calculated from predicted and calibration property values, and, for certain entries, the numbers of compounds, out of the first ten most significant (see Equation 1 and its description) in each model, that would be expected to correlate to the fuel property being modeled (such as aromatic compounds in the case of the aromatics models). In addition, the modeling results reported for the detection of FSII as per ASTM D5006 indicate how significant FSII itself is in the corresponding model, with lower FSII compound number / total compound number ratios indicating a higher significance. Table 3 further summarizes Table 2's results by tabulating the number of green-highlighted, and thus most accurate, modeling metrics found in Table 2.

Interestingly, the RMSEP and R^2 values would seem to indicate that the use of MF threshold values of about 600 would provide for more accurate quantitative modeling results. However, modeling results that would more accurately reflect underlying chemical phenomena might be better pursued by employing a higher MF threshold value of about 700-750 (with 850 not being similarly considered due to its inability to reliably provide low RMSEP values). This latter observation would be consistent with higher MF thresholds excluding the maximum possible numbers of unwanted interferences, thus allowing models to more directly focus upon compounds of interest.

Because FCAST is intended for use with as wide a variety of fuel samples as possible, it is considered somewhat more important for FCAST's practical applicability that it accurately assess future fuel samples at the compound level. When this is considered alongside the fact that an MF threshold value of 750 is the second-most effective threshold to use for quantitative property modeling after 600 (i.e. in the context of producing low RMSEP results), a default MF threshold value of 750 is deemed appropriate for use in FCAST with respect to both fuel modeling and non-fuel modeling applications.

It is also apparent from Table 2 that lubricity and sulfur content are not accurately modeled from GC-MS data. This is not unexpected, since lubricity is a function of trace levels of carboxylic acids and other surface active agents, and GC-MS is not necessarily diagnostic for such species. It may be possible to more accurately model lubricity with a GCxGC model, but that would be the topic of a future research effort. Similarly, organosulfur (and organonitrogen) compounds are not easily detected in standard unit resolution GC-MS and would probably require specific detectors, which, again, would be the topic of additional research efforts.

		600	650	700	750	800	850
Density	*RMSEP*	0.0031	0.0036	0.0034	0.0036	0.0038	0.0038
(ASTM D4052)	R^2	0.98	0.98	0.98	0.98	0.97	0.97
Flash Point	*RMSEP*	5.0306	4.9861	4.9893	5.0845	5.1655	5.3533
(ASTM D56 & D93)	R^2	0.83	0.83	0.83	0.82	0.82	0.80
Viscosity, -20C	*RMSEP*	0.7123	0.4671	0.4255	0.3851	0.3634	0.4948
(ASTM D445)	R^2	0.59	0.83	0.85	0.88	0.89	0.80
Viscosity, 40C	*RMSEP*	1.0301	1.0882	1.0401	0.8390	1.0748	1.1288
(ASTM D445)	R^2	0.65	0.60	0.64	0.77	0.61	0.57
Pour Point	*RMSEP*	5.5614	5.3798	5.6692	5.1629	5.3286	5.7850
(ASTM D97 & D5949)	R^2	0.91	0.91	0.90	0.92	0.91	0.90
Cloud Point	*RMSEP*	3.0381	3.1305	3.2973	3.4205	3.4868	3.6472
(ASTM D2500 & D5773)	R^2	0.80	0.78	0.76	0.74	0.73	0.71
Freeze Point	*RMSEP*	2.2023	2.2273	2.2354	2.2953	2.1642	2.1664
(ASTM D2386 & D5972)	R^2	0.84	0.84	0.83	0.82	0.84	0.84
Cetane Index	*RMSEP*	1.5040	1.5058	1.6179	1.5130	1.5580	1.5426
(ASTM D976)	R^2	0.87	0.87	0.85	0.87	0.86	0.86
Aromatics, FIA	*RMSEP*	0.5611	0.4059	0.7685	0.7881	0.5533	1.1136
(ASTM D1319)	R^2	0.94	0.97	0.90	0.89	0.95	0.78
	Aromatics (of 1st 10)	7	7	8	8	8	6
Aromatics,	*RMSEP*	0.4437	0.4793	0.3890	0.3803	0.8162	0.9428
HPLC	R^2	0.99	0.99	0.99	1.00	0.98	0.97
(ASTM D6379)	*Aromatics (of 1st 10)*	6	4	4	4	6	6
Olefins, FIA	*RMSEP*	0.7024	0.6980	0.5928	0.6430	0.6993	0.3406
(ASTM D1319)	R^2	0.87	0.88	0.91	0.89	0.87	0.97
	Olefins (of 1st 10)	1	2	3	4	4	6
Saturates, FIA	*RMSEP*	0.9464	0.7191	0.8842	0.7592	0.5875	1.6901
(ASTM D1319)	R^2	0.96	0.98	0.97	0.97	0.99	0.88
	Saturates (of 1st 10)	4	5	5	7	5	6
Naphthalene	*RMSEP*	0.2373	0.2143	0.2585	0.1813	0.2845	0.3103
Content	R^2	0.90	0.92	0.88	0.94	0.86	0.83
(ASTM D1840)	*Naphthalenes (of 1st 10)*	4	4	4	2	2	3
Hydrogen Content	*RMSEP*	0.3732	0.4515	0.4354	0.4518	0.4632	0.4954
(ASTM D3343 & D3701)	R^2	0.98	0.96	0.97	0.96	0.96	0.96
Fuel System Icing	*RMSEP*	0.0201	0.0195	0.0205	0.0209	0.0237	0.0237
Inhibitor, FSII	R^2	0.58	0.61	0.57	0.55	0.42	0.42
(ASTM D5006)	*FSII # / Total #*	136/563	75/401	41/274	51/178	103/163	68/99
Ash Content	*RMSEP*	0.0015	0.0018	0.0016	0.0019	0.0017	0.0018
(ASTM D482)	R^2	0.68	0.54	0.61	0.50	0.58	0.52
Smoke Point	*RMSEP*	0.6657	0.8883	1.6160	0.2764	0.7312	1.0334
(ASTM D1322)	R^2	0.93	0.88	0.60	0.99	0.92	0.83

Table 2 (part 1). UVE-PLS prediction results obtained for thirty-five fuel properties with various MF thresholds. Modeling metrics in each row indicative of the most accurate results in green.

		600	650	700	750	800	850
Distillation, IBP	*RMSEP*	7.7751	7.8725	7.9236	9.0739	8.0901	8.2605
(ASTM D86)	R^2	0.81	0.81	0.80	0.74	0.80	0.79
Distillation, 10%	*RMSEP*	6.3601	6.3304	6.2913	6.2534	6.3395	7.9354
(ASTM D86)	R^2	0.94	0.94	0.94	0.94	0.94	0.90
Distillation, 20%	*RMSEP*	4.4267	4.4373	4.4463	4.6547	5.0372	5.1180
(ASTM D86)	R^2	0.98	0.98	0.97	0.97	0.97	0.97
Distillation, 50%	*RMSEP*	4.6961	5.1436	4.8903	5.1366	5.1733	5.6745
(ASTM D86)	R^2	0.98	0.98	0.98	0.98	0.98	0.98
Distillation, 90%	*RMSEP*	7.2654	7.5152	7.9337	7.7594	8.3251	8.7608
(ASTM D86)	R^2	0.97	0.97	0.97	0.97	0.97	0.96
Distillation, FBP	*RMSEP*	9.6089	9.7122	9.7573	9.3430	10.4595	11.0333
(ASTM D86)	R^2	0.96	0.96	0.96	0.96	0.95	0.94
Existent Gum	*RMSEP*	1.5327	1.5836	1.5996	1.7661	1.5130	1.8145
(ASTM D381)	R^2	0.64	0.62	0.61	0.53	0.65	0.50
Lubricity	*RMSEP*	0.0416	0.0482	0.0398	0.0405	0.0504	0.0617
(ASTM D5001)	R^2	0.56	0.41	0.60	0.58	0.35	0.03
Acid Number	*RMSEP*	0.0523	0.0521	0.0496	0.0540	0.0552	0.0571
(ASTM D974 & D3242)	R^2	0.60	0.60	0.64	0.57	0.55	0.52
Storage Stability	*RMSEP*	0.9090	0.8443	0.9737	0.9954	1.0387	0.9878
(ASTM D5304)	R^2	0.31	0.40	0.21	0.17	0.10	0.18
Particulates (ASTM	*RMSEP*	4.4994	4.5444	4.5673	4.4679	4.6687	4.5320
D2276, D5452 & D6217)	R^2	0.35	0.33	0.33	0.36	0.30	0.34
K.F. Water Titration	*RMSEP*	11.4125	8.7232	15.6786	12.5989	7.0911	10.9182
(ASTM D6304)	R^2	0.74	0.85	0.51	0.69	0.90	0.76
Carbon Residue	*RMSEP*	0.0186	0.0254	0.0244	0.0223	0.0237	0.0287
(ASTM D4530)	R^2	0.80	0.64	0.66	0.72	0.68	0.53
Carbon Residue	*RMSEP*	0.0181	0.0188	0.0158	0.0197	0.0166	0.0212
(ASTM D524)	R^2	0.74	0.72	0.80	0.69	0.78	0.64
Demulsification	*RMSEP*	2.8584	2.8236	2.7572	2.8374	3.1170	2.9826
(ASTM D1401)	R^2	0.26	0.27	0.31	0.27	0.11	0.19
Sulfur,	*RMSEP*	153.71	153.39	204.76	246.51	342.29	282.01
Wave. Dis. XRF	R^2	0.88	0.88	0.79	0.69	0.40	0.59
(ASTM D2622)	*S-Containing (1st 10)*	0	0	0	0	0	0
Sulfur, by mass	*RMSEP*	0.0625	0.0630	0.0673	0.0736	0.0846	0.0999
(ASTM D1 and	R^2	0.91	0.90	0.89	0.87	0.83	0.76
ASTM D4294)	*S-Containing (1st 10)*	0	1	0	1	0	1
Water Index	*RMSEP*	1.6898	2.0568	1.2604	1.7378	1.8587	3.8847
(ASTM D3948)	R^2	0.94	0.91	0.97	0.94	0.93	0.70

Table 2 (part 2). UVE-PLS prediction results obtained for thirty-five fuel properties with various MF thresholds. Modeling metrics in each row indicative of the most accurate results in green.

	600	650	700	750	800	850
Number Of Times This Match Factor Yields The Lowest RMSEP Value	11	5	5	8	5	1
Number Of Times This Match Factor Allows For The Most Appropriate Compounds To Influence Modeling	2	2	3	3	2	3

Table 3. Summary of Table 2 results, obtained by tabulating the number of green-highlighted, and thus most accurate, modeling metrics found in Table 2.

Fuel property model reconstruction. After deciding upon an MF threshold value of 750 for the production of metaspectra, each fuel property model reported upon in Table 2 was reconstructed 100 times to compensate for the random variance associated with UVE-PLS. The model providing the lowest RMSEP amongst the 100 reconstructions, for each fuel property, was thus selected as the best model, and the corresponding lists of compounds will be utilized in FCAST. It should be noted, however, that although 35 distinct fuel properties were modeled during development work, fewer fuel properties, corresponding to what would typically be assessed during routine fuel analyses, will be represented in the final FCAST software, at least until it is deemed appropriate to accommodate additional fuel properties. Further, although development work utilized distillation value and FSII prediction models constructed via UVE-PLS, FCAST itself utilizes alternative, compositionally-focused methodologies to predict these fuel properties.

2.3 Mass Channel Analysis

The first iteration of FCAST compositional profiler attempted to implement the full algorithm developed by NIST for AMDIS.[26] Unfortunately, as implemented, very few of the peaks identified had enough masses associated with them to identify the corresponding chemical component properly and would thus return poor results. This resulted in the algorithm spending a great deal of time identifying peaks, only to then disregard these peak identifications in favor of sending the entire raw scan to the NIST/EPA/NIR Mass Spectral database for compound identifications. To accelerate that process in the initial version of FCAST, the mass channel analysis was skipped and a simpler peak finding algorithm was used with the TIC to identify peaks and subsequently send the corresponding data for analysis. Returning to the initial research done with the profiler, the parameters of the algorithm were adjusted to allow more masses to be selected during the analysis. This resulted in a sufficient number of masses to be used during the search to provide good identification results.

The fuels used for comparing the EWFA-MCR algorithm versus AMDIS were used to compare the default compound search using the Agilent ChemStation software provided with in-house GC-MS instrumentation versus the three FCAST search algorithms. All results reported in Table 4 are unique names, with a match factor threshold of 750, and an area > 0.001% of the total sample. All of the FCAST search algorithms reported more compounds identified than the default ChemStation results. Mass Channel Analysis reported more peaks identified then the simple method while approaching the numbers reported by the EWFA-MCR algorithm. The FCAST implementation of the EWFA-MCR algorithm reported fewer results in Table 4 than seen previously due to some constraints placed on the algorithm to reduce the analysis time to a more manageable number. The average time for FCAST analysis of the three methods listed are minutes, tens of minutes, and hours, respectively, based on the recommended GC-MS method.

The Mass Channel Analysis was implemented in the FCAST as a less computationally intensive alternative to EWFA-MCR peak deconvolution that could provide improved chromatographic resolution in less time. EWFA-MCR would be thus used when the most detailed resolution is required, such as when comparing two fuels for minor compositional differences.

Grade	Agilent ChemStation	FCAST Simple	FCAST Mass Channel	FCAST EWFA-MCR
JP-5	57	115	189	224
JP-5	50	118	213	233
JP-8	71	132	238	275
JP-8	84	159	279	372
JetA	82	150	243	289
JetA	69	110	213	222
F-76	78	132	186	157
F-76	57	115	128	98
MGO	106	141	235	237
MGO	104	116	172	222
Alt. Diesel (CHCD)	84	153	251	197
Alt. Diesel (HEFA)	80	122	222	213

Table 4. Number of compounds reported for standard Agilent ChemStation analysis, vs FCAST simple, mass channel and EWFA implementations.

3.0 FCAST Software Automation

The following section reports upon recent FCAST modifications as they are related to the enhancement of the FCAST's automated capabilities.

3.1 Automation of Comparison Sub-Routines

FCAST incorporates two inter-sample comparison sub-routines, which can both be utilized to quickly find the differences between fuels. The deltaCompare sub-routine was designed to quickly and quantitatively compare the magnitudes of area-normalized TICs of two fuel samples, subjecting the higher-magnitude mass spectrum corresponding to any given retention time to a NIST database search for identification purposes if the difference between the two relevant TIC values is greater than the standard deviation of the differences in the two TICs at all retention times multiplied by a constant value. In contrast, the feature selection strategy based on contrasting ANOVA results collected from between-sample and within-sample variances was designed to use the relative differences between larger replicate data populations to isolate more subtle yet still informative data features for further analysis and assessment, providing a more thorough comparative analysis in exchange for the collection of several replicate GC-MS data sets.

deltaCompare. The deltaCompare comparative sub-routine currently offers only a single analysis option, i.e. the ability to adjust the standard deviation multiplication constant associated with its statistical functions. However, previous reporting[12] has already indicated that 2.33 (consistent with a one-tailed z-test at a 99% confidence interval) is a suitable default value. It would thus be a straightforward software change simply to move the standard deviation multiplication constant adjustment capability to a less accessible sub-menu to encourage result uniformity across multiple iterations of the software.

ANOVA sub-routine. The ANOVA sub-routine provides the means to compare the compositions of two fuels using an Analysis of Variance (ANOVA)[27] technique wherein the ratio of the between-sample to within-sample variance is used to determine which data points in the GC-MS total ion chromatogram are statistically different between the two samples. This algorithm requires the informed selection of several critical parameters, most notably an F-ratio threshold value that is used to determine which data features are indicative of categorical change. An F-ratio is calculated for each comparable data point in the GC-MS data collected for multiple samples, and the highest values within these F-ratio collections indicate the most prominent differences between sample categories. A more detailed explanation of F-ratio values and ANOVA can be found in a previous NRL Memorandum Report.[12] *In summary* (see below for details), it was found that a reasonable F-ratio threshold can be automatically selected for any given GC-MS data comparison by first calculating all of the F-ratios for a given class comparison, then taking the natural logarithms of all of these collected values to minimize the effects of overly large F-ratios, then normalizing all

of these collected and transformed values to the maximum possible post-transform value, and then finding the mean and standard deviation of the collected values. The automated F-ratio threshold is thus defined as the mean value plus the standard deviation value multiplied by 1.96.

There are two additional analysis options associated with the ANOVA sub-routine: whether or not to normalize the data, and whether or not to align the data prior to running the actual comparison algorithms. Given that the F-ratio threshold selection methodology itself makes use of normalization, and given that normalization and alignment are prudent preprocessing steps to take in the context of comparing newly collected data sets to pre-existing ones, especially when strict data quality measures might be difficult to implement across multiple laboratories, there is currently no reason to suspect that these two preprocessing options should not be applied as default options in widely distributed versions of FCAST. However, as with all default options, the ability to deselect these two preprocessing steps should be maintained in the final software in at least some capacity for use by advanced software operators.

Details of Fisher-ratio (F-ratio) threshold determination work. Because the goal is to determine which F-ratios in any given F-ratio collection are different from the overall population of F-ratios, one possible course of action is to simply calculate the mean and standard deviation of the entirety of a data comparison's collected F-ratios. One can then add the mean to the standard deviation, multiplied by a constant value associated with the desired confidence interval (as is done in the deltaCompare sub-routine), and use the resulting value as the F-ratio threshold value against which to compare the F-ratios within the original collection. Pursuing this course of action, thus producing a threshold value that appropriately scales with the data without user intervention, requires the determination of a suitably robust constant by which to multiply the standard deviation, one that is as large as possible to eliminate as many undesirable results as possible while still allowing for the reliable identification of relevant compound changes.

Initially, it was believed that the most prominent challenge associated with automatically selecting an F-ratio threshold in this way would be that ANOVA data comparisons can potentially produce extremely large F-ratios, corresponding to extremely prominent changes, which might interfere with the appropriate reporting of smaller yet still realistically significant F-ratios. To accommodate these extremely large values, the natural logarithms of the calculated F-ratios for an entire categorical comparison can be calculated, effectively reducing the relative intensities of extremely large values. Experiments were thus performed to assess what the effects of such a data transformation would be on ANOVA-based comparison results, as opposed to results obtained via untransformed (control) F-ratio collections.

Two different permutations of the logarithmic transformation were utilized during the present work to determine their relative impacts on mean and standard deviation calculations, and thus final comparison results. One permutation leaves negative post-logarithmic transformation values

intact, while the alternative permutation adds, to the entire collection of transformed F-ratios, the minimum constant value necessary to eliminate all negative values in an attempt to improve upon mean and standard deviation determinations. An MF threshold value of 750 is applied to individual compound identifications, as was previously shown to be optimal. It should also be noted that, during preliminary work, it was realized that 1-methynaphthalene was likely to have sometimes been misidentified as 2-methylnaphthalene via NIST database searches, regardless of MF threshold value, due to the similarities between the 1-methylnaphthalene and 2-methylnaphthalene library mass spectra, and these misidentifications were thus accounted for in the results as 1-methylnaphthalene.

The ANOVA technique itself is typically deployed in scenarios involving only two fuels between which more subtle differences are being sought. In these scenarios, provided that the necessary time and resources are available, replicate GC-MS data sets are explicitly collected for the purposes of performing ANOVA-based comparisons. Thus, it was decided to evaluate possible standard deviation multiplication constant values with replicate-based surrogate fuel classes. Pursuant to this decision, Table 5 displays the compositions of the three surrogate fuels, numbered in-house as 29, 30, and 31, from which five suitable replicate data sets each were collected. These three fuels can be variously paired and compared with each other via ANOVA, as two classes of five-replicate sets, to yield three meaningful comparisons upon which to gauge the effectiveness of the automated threshold selection methodology. It should be stated here that these comparisons were performed both by defining the lower-numbered surrogate fuel as class 1 and the higher-numbered surrogate fuel as class 2, and vice versa, but this ordering did not appear to have an impact on the results reported herein.

	Surrogate Fuel 29	Surrogate Fuel 30	Surrogate Fuel 31
n-dodecane	15%	5%	10%
1-dodecene	5%	10%	0%
toluene	40%	40%	0%
1,2,4-trimethylbenzene	20%	0%	30%
tetralin	0%	40%	30%
1-methylnaphthalene	20%	5%	30%

Table 5. Percent compositional contents (v/v) of the three surrogate fuels evaluated to develop the automated F-ratio threshold determination methodology.

Figure 2 shows three sets of results, corresponding to control results that do not make use of logarithmic data transforms (though the other steps required to define the threshold value are carried out in the same manner as described previously), depicted as a red line, and results obtained with the two different permutations of the logarithmic transformation described previously, depicted as the blue and green lines, with the green line showing the results which include the elimination of negative values. To investigate how to define optimal F-ratio threshold values, the potential standard deviation multiplication constant was scaled from 0.5 to 10, in steps of 0.5 in this figure, to determine the most suitable constant value by which to multiply the standard deviations of all variable-specific ANOVA values within any given comparison to help define the overall threshold value. The goal, at least initially, was to find the maximum constant value that would provide as much discriminatory capability as possible (suggesting a relatively high value) while still allowing for even the small differences between classes, as shown in Table 5, to be detected (suggesting a relatively low value). Figure 2 also includes dashed lines indicating the maximum numbers of relevant compounds that might be identified within any given set of results, as also indicated by Table 5. If, as is desired, all relevant compounds are identified within a given set of results, the line corresponding to said results will overlap with the corresponding dashed line.

Interestingly, however, Figure 2 would seem to indicate that the most effective methodology, that of utilizing the logarithmic transform without the additional corrective step (again, as represented by the blue line), provides a great deal of leeway with respect to which constant value might be implemented, as it allows for all possible class differences to be identified along the widest ranges of possible multiplication constants, up to a constant of at least 7 across all three class comparisons.

In order to obtain a more definitive understanding of which multiplication constant would best serve as an optimized default value, one can reinterpret ANOVA comparison results based on whether or not the sums of the original, non-normalized data at the retention times corresponding to the ANOVA results are greater for class 1 or class 2, thereby pinpointing within which class (i.e. which fuel) more of any given compound can be located. These results can further be summed along the mass spectral data axis to produce TIC results, up to and including entire TIC spectra, corresponding to all data increases found in one class compared to another class, as well as individual compound increases and decreases across classes as determined via NIST/EPA/NIH database searches. Such class-specific TIC results can be found plotted in Figures 3a through 4c for all three surrogate fuel comparisons, when utilizing the multiplication constant values of 0.5 (for the sub-figures of Figure 3) and 6.0 (for the sub-figures of Figure 4). What is made apparent by cross-referencing these figures with the fuel compositions shown in Table 5 is that the larger compound-specific peaks within each class pair always correspond to larger amounts of the corresponding compounds known to be within any given class comparison, thus further indicating the fundamental effectiveness of the ANOVA class comparison strategy.

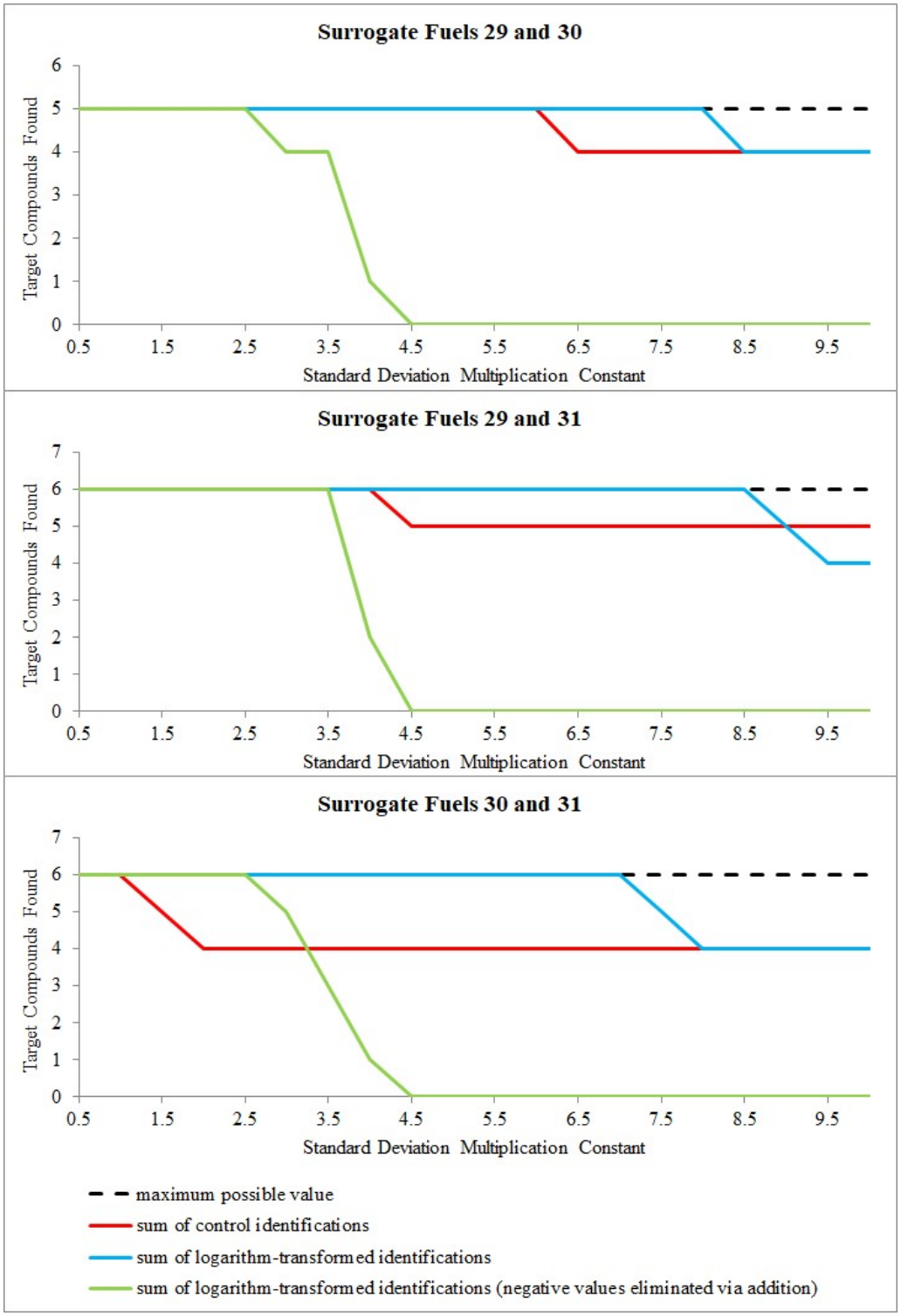

Figure 2. Summaries of the total numbers of compounds found that would have been expected to change between pairings of three surrogate fuels, utilizing multiple standard deviation multiplication constants and three different F-ratio transformations, including the control, which skipped logarithmic transformation.

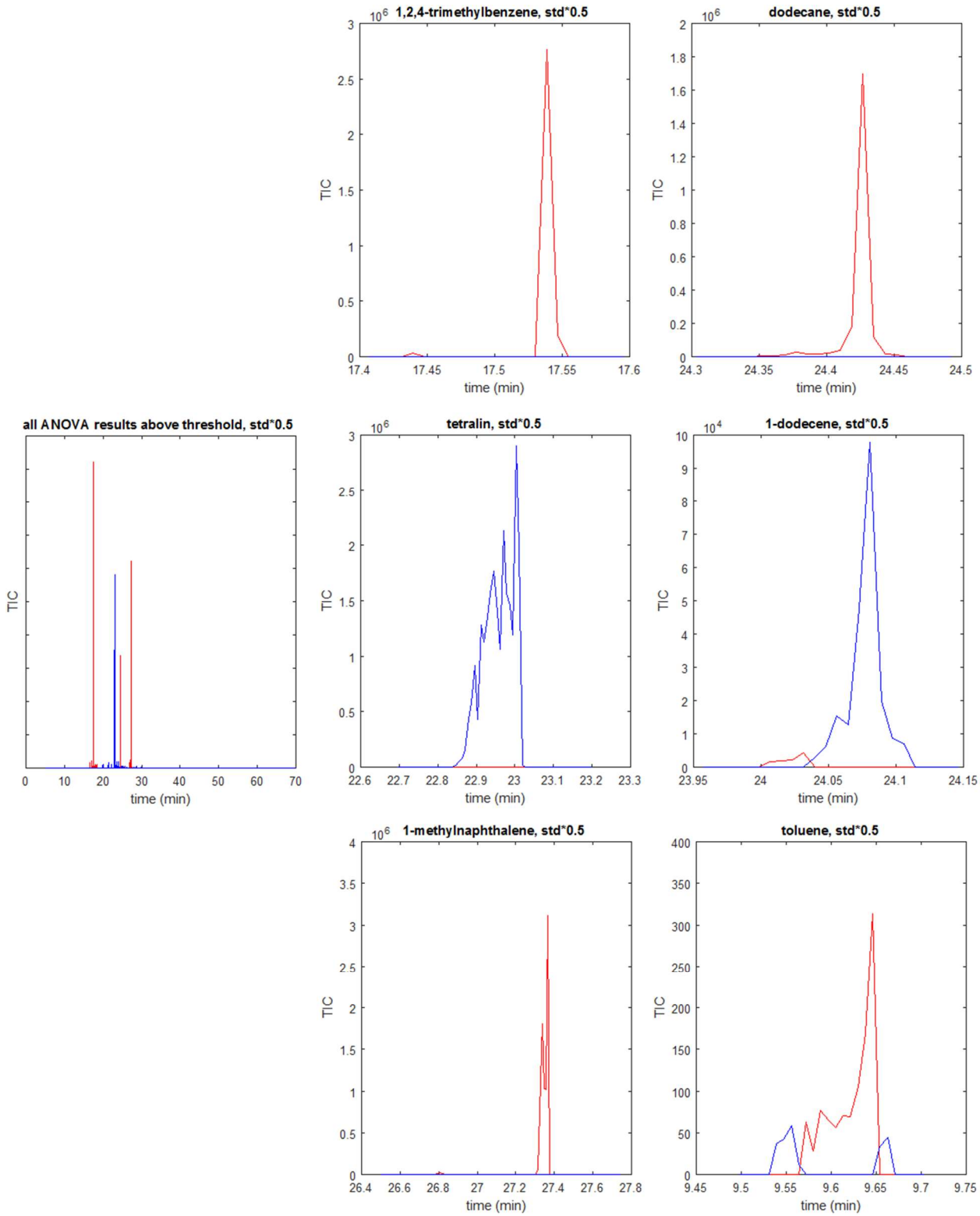

Figure 3a. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 30 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 0.5.

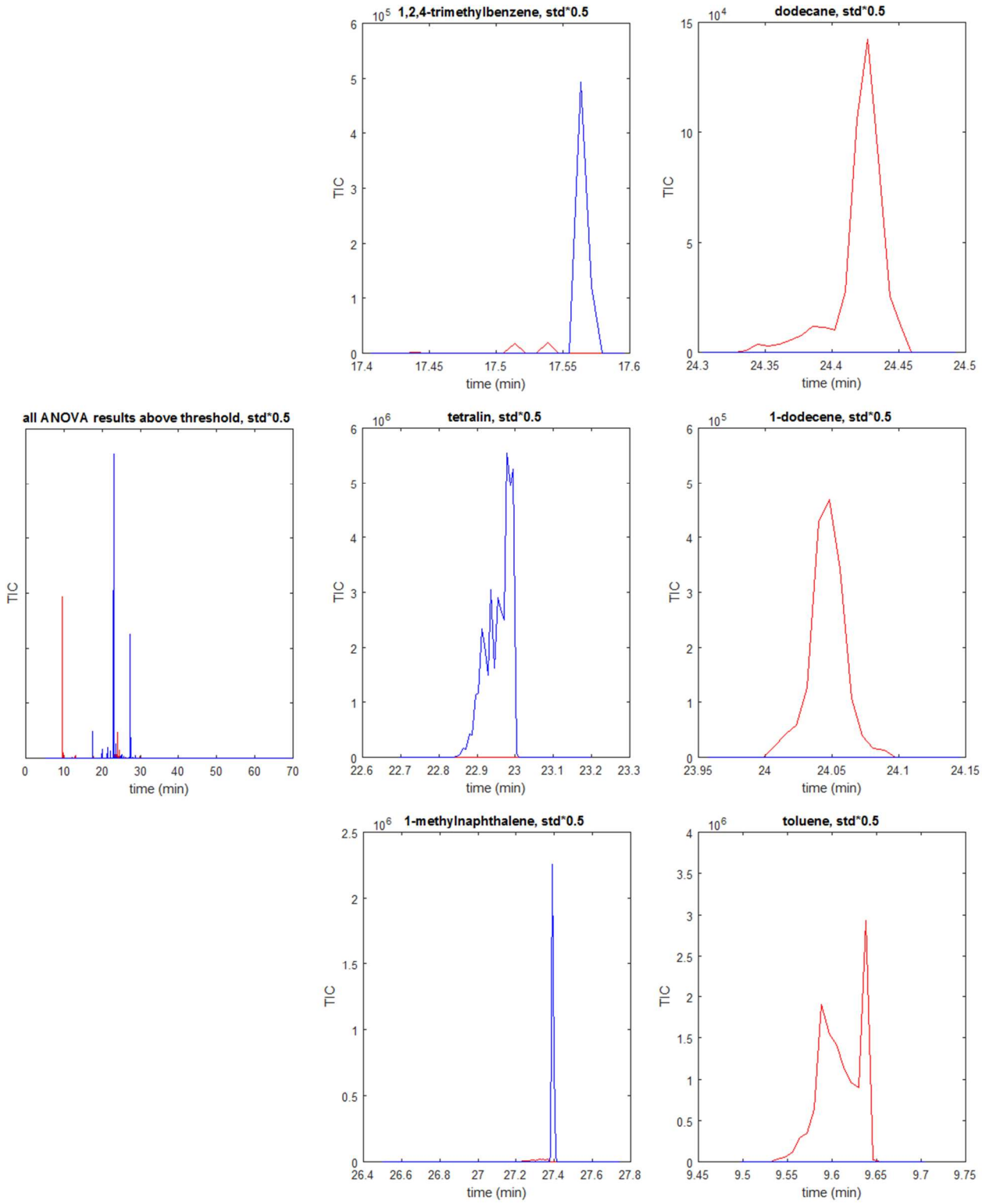

Figure 3b. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 0.5.

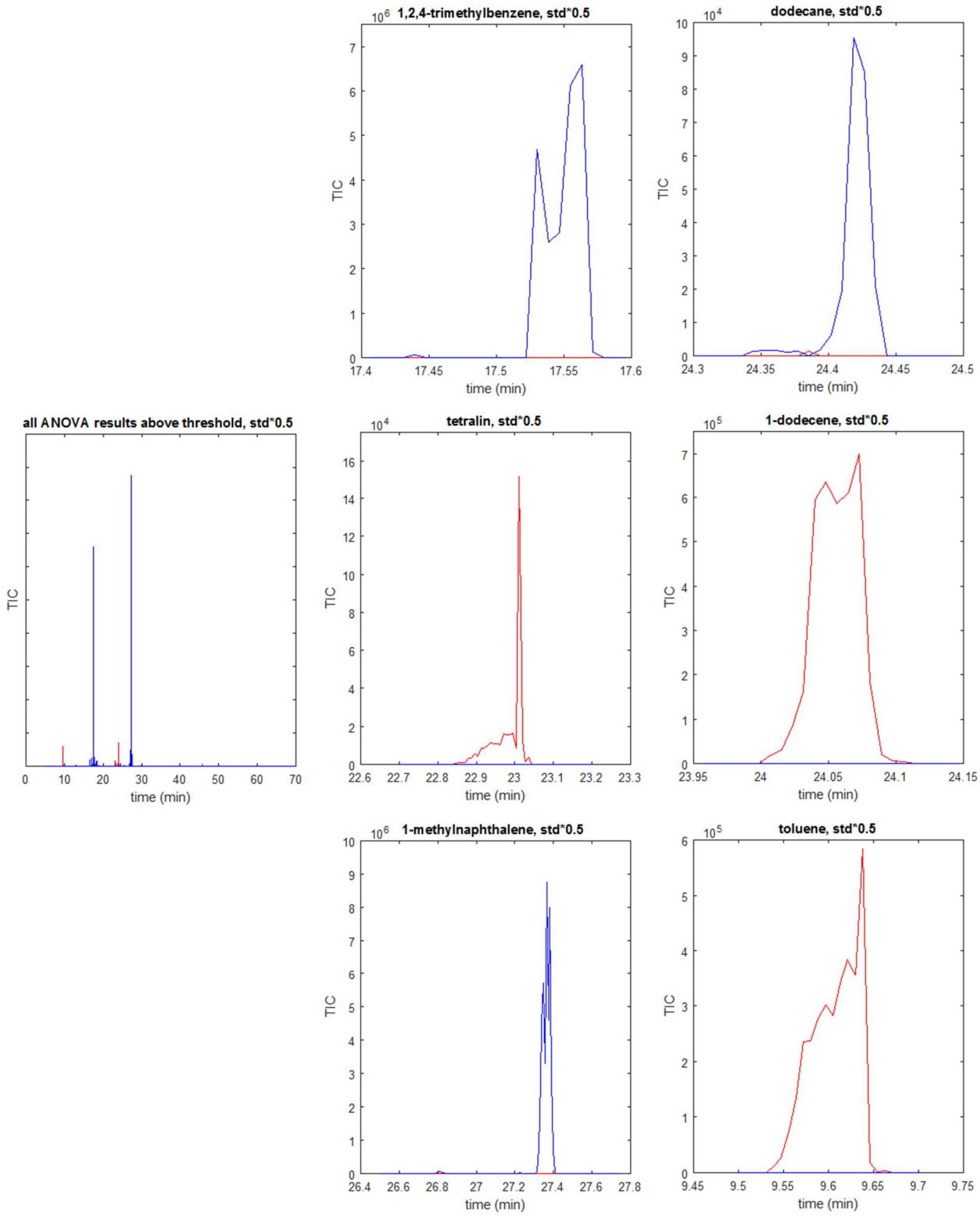

Figure 3c. Class-sorted TIC results, obtained for the Surrogate Fuel 30 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 0.5.

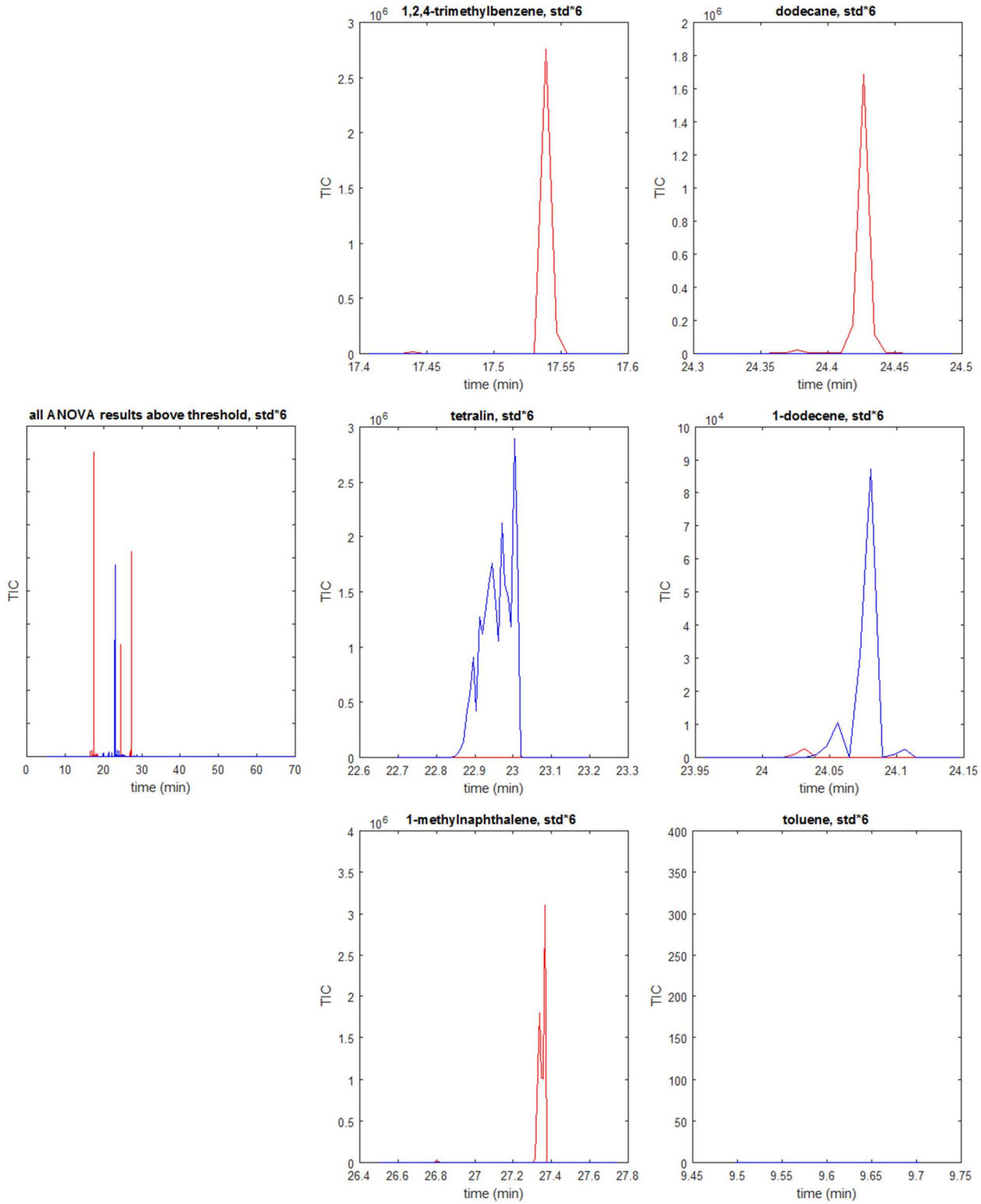

Figure 4a. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 30 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 6.0.

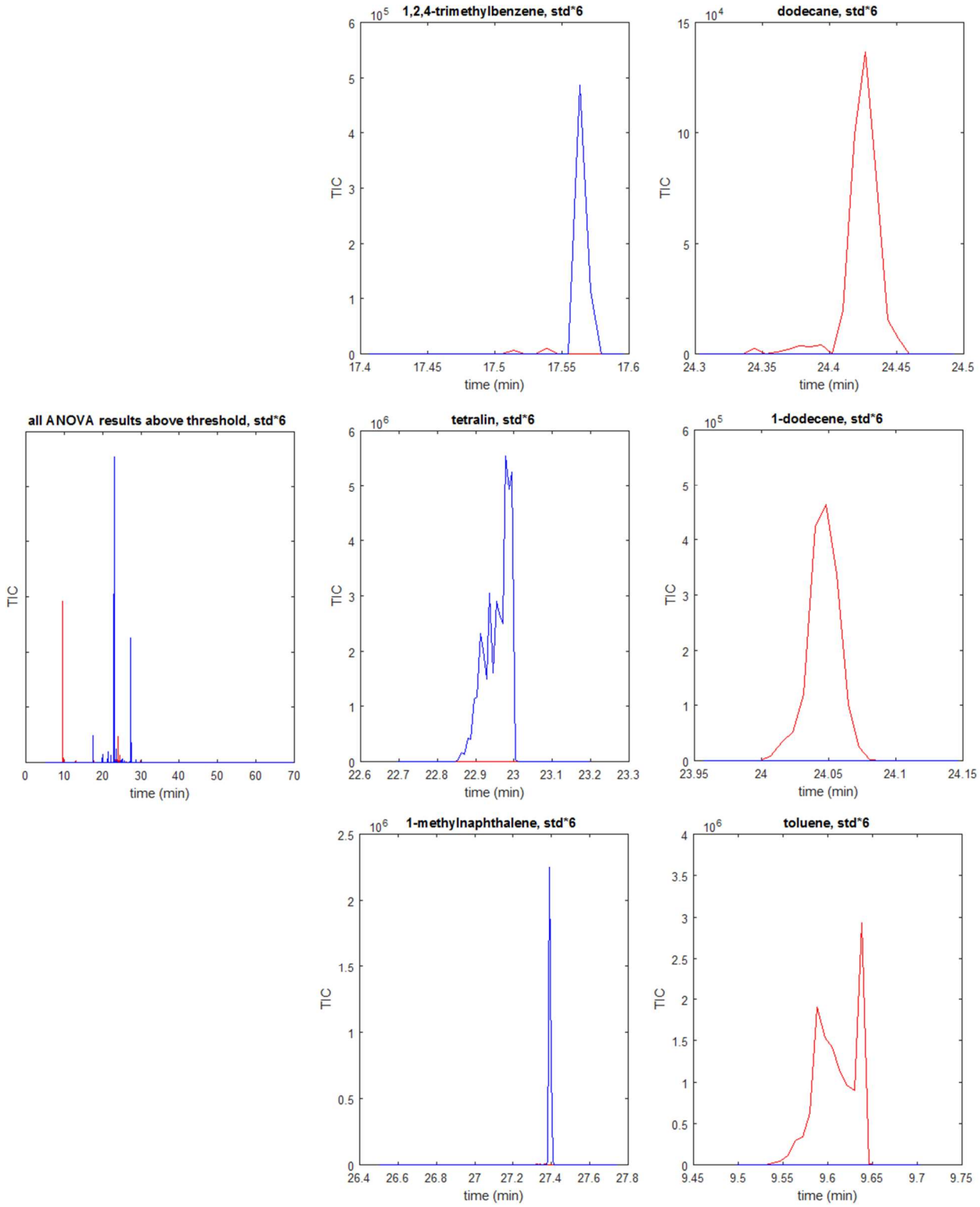

Figure 4b. Class-sorted TIC results, obtained for the Surrogate Fuel 29 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 6.0.

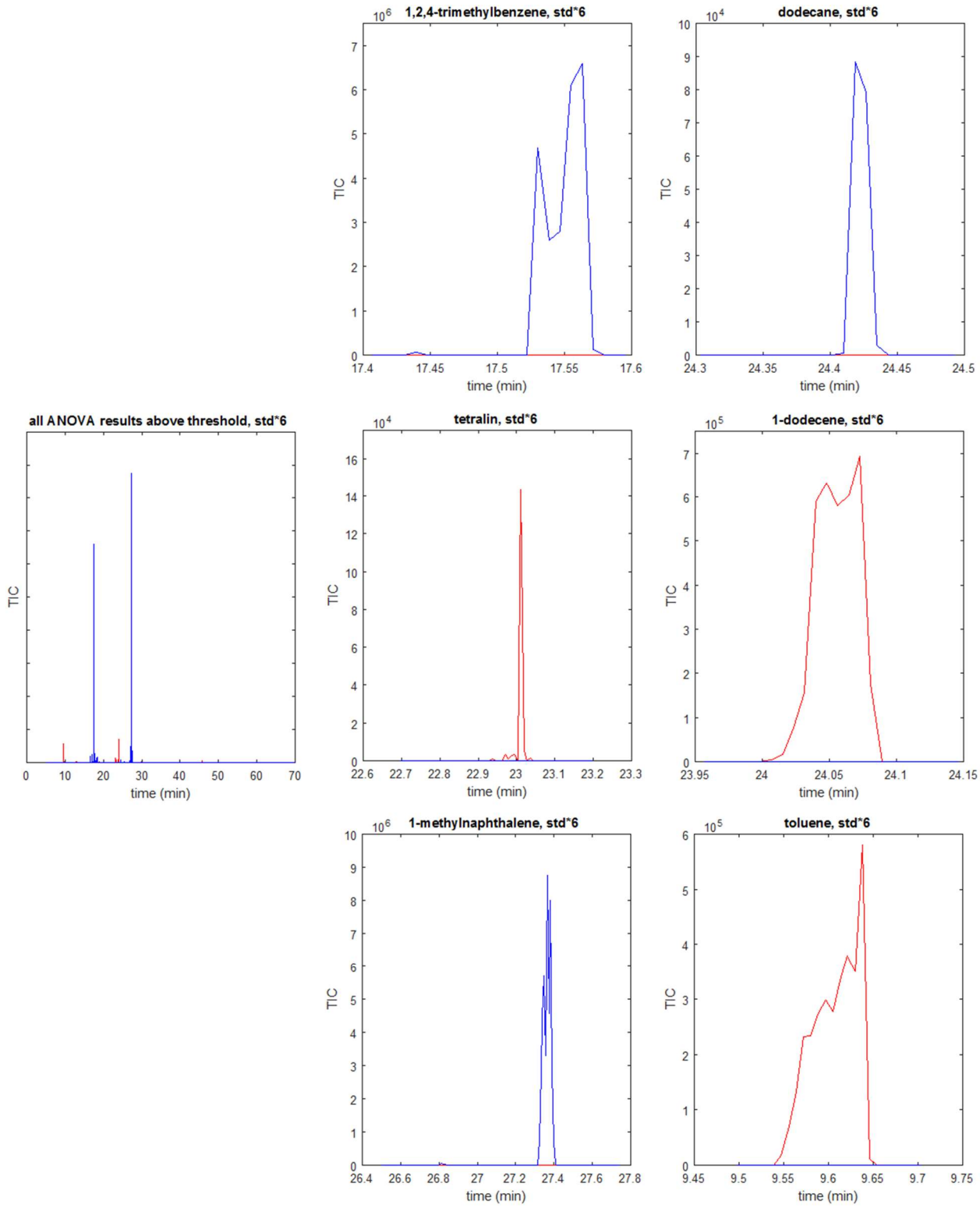

Figure 4c. Class-sorted TIC results, obtained for the Surrogate Fuel 30 (red) / Surrogate Fuel 31 (blue) comparison, whose corresponding logarithmic ANOVA values exceed the threshold logarithmic value defined for a multiplication constant value of 6.0.

What is also apparent in Figures 3a through 4c, however, is that the TIC spectra do not look overly different when comparing the results identified when employing the two disparate multiplication constant values of 0.5 and 6.0. Note in particular that the scales of corresponding figures do not even need to be modified when considering the results obtained using the two different values. The most obvious apparent change when using these two multiplication constant values is that the toluene-based differences found between Surrogate Fuels 29 and 30 are not apparent when using a value of 6.0, as can be seen in Figure 4a, whereas they are apparent when using a value of 0.5, as can be seen in Figure 3a. However, even when detected, this difference is orders of magnitude below the other differences found between the two surrogate fuel replicates, as might be expected given that both of these surrogate fuels should contain nearly identical amounts of toluene according to Table 5. This toluene difference is thus not considered a sufficient metric upon which base the selection of a default multiplication constant value.

Despite these visual similarities, however, there are small differences to be found when employing different multiplication constants, and these differences can be further analyzed to select a default multiplication constant. In order to more ably consider these small differences in aggregate, and because the goal of this analysis is to assess unambiguous discriminatory capability, one can first subtract the compound-specific ANOVA-selected class 1 TIC values from the compound-specific ANOVA-selected class 2 TIC values seen in Figures 3 and 4, then find these differences' absolute values. The sum of these differences can then be obtained for the smallest tested multiplication constant of 0.5. Similar sums can then be produced while utilizing other multiplication constants. These sums would be expected to decrease in magnitude as multiplication constants increase because fewer individual ANOVA comparison results would be able to exceed the thresholds thus defined. For a visual representation, one can thus subtract these sums from the sum obtainable when using the smallest tested multiplication constant (i.e. 0.5) to determine how much discriminatory capability is lost when larger multiplication constants are employed. This operation produces results that are more negative as more discriminatory capability is lost when higher multiplication constants are employed, as can be seen in Figure 5 for all three class comparisons.

Figure 5 shows that the TIC differences that can be unambiguously associated with the expected changes between the compounds known to be different between the two classes decrease exponentially with increases in multiplication constant. This would seem to indicate that erring on the side of caution (i.e. lower multiplication constants) would be advisable in selecting a default constant value, thus helping to ensure that meaningful class differences continue to be identified, even if those class differences are subtle.

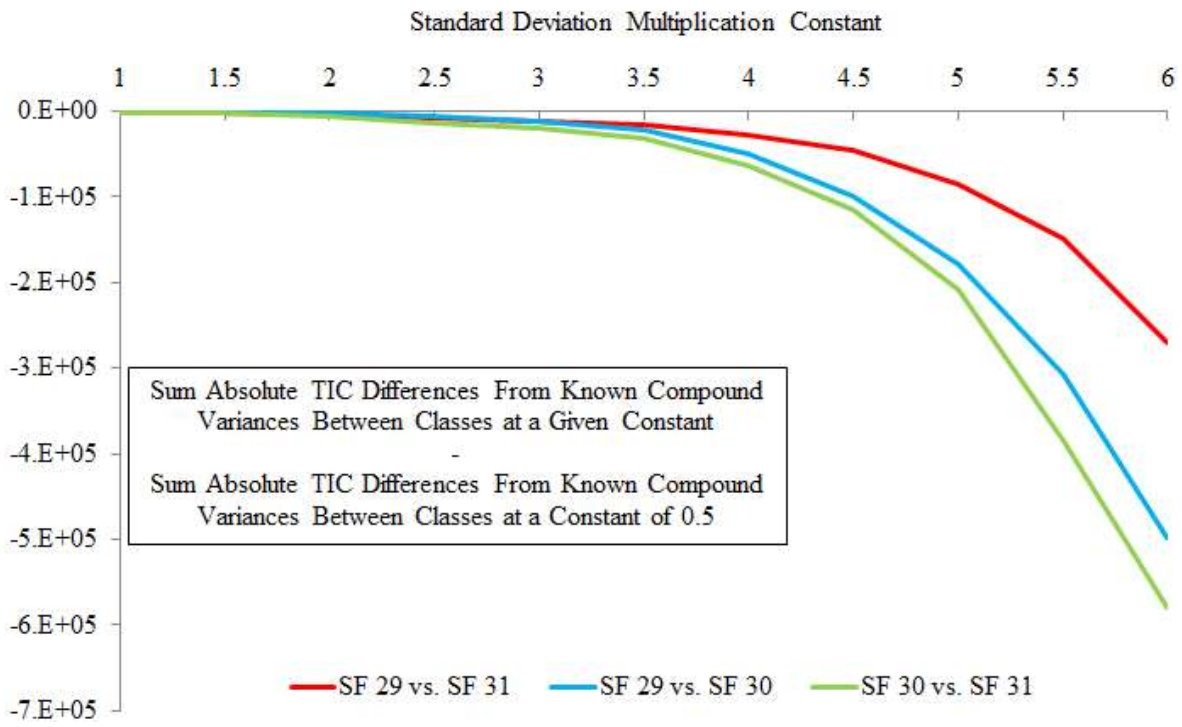

Figure 5. Sum absolute TIC differences between classes, for compounds known to be different between classes, with losses being reported relative to the results obtained when a constant of 0.5 is employed.

The TIC differences seen in Figure 5 when employing a constant value of 2 remain very close to zero, which means that the overall TIC difference results obtained when using a multiplication value of 2 do not differ a great deal from the results obtained when using a value of 0.5. A multiplication constant value of 2 can be altered slightly to the proximal value of 1.96 in order to conform the constant to a one-tailed z-test at a 97.5% confidence interval, which is a fairly standard constant employed in statistical analyses. Given available evidence, then, a standard deviation multiplication constant value of 1.96 would seem to be a reasonable default value to employ in the context of the ANOVA comparison strategy.

3.2 Default Peak Area Threshold Selection

Because of the need for reliably optimized analysis parameters in automated software applications, and because data deconvolution is being implemented into FCAST specifically, work was undertaken during the present work to determine whether or not FCAST's pre-existing default peak area threshold, 0.001%, should be redefined to perform more optimally with the most up-to-

date version of the software. It was initially suspected that adjusting this threshold between values of similarly small magnitudes would not overly affect peak identification and modeling results when the relatively restrictive MF threshold of 750 is employed. Thus, the peak area thresholds tested covered a somewhat more expansive logarithmic progression of 0.1%, 0.01%, 0.001% (previously determined default), 0.0001%, and 0.00001%. While utilizing the four threshold values not previously investigated, deconvolved compositional results from the eighty fuel samples required to model five different compositional properties (naphthalene content via ASTM D1840, aromatics content via ASTM D6379, and aromatics, olefins and saturates contents via ASTM D1319) were recollected and subjected to the same UVE-PLS modeling shown previously, for three replicate calculations. RMSEP and R^2 values were collected, as well as how many of the ten most prominent compounds in each model actually represent the targeted compound class, and the results from the three replicates indicating the highest modeling accuracy were determined. These modeling results, alongside the original 0.001% results shown in Table 2, are shown in Table 6.

		0.10%	0.01%	0.001%	0.0001%	0.00001%
Aromatics, FIA (ASTM D1319)	*RMSEP*	0.5275	0.7847	0.7881	0.791	0.7737
	R^2	0.95	0.89	0.89	0.89	0.89
	Aromatics (of 1st 10)	5	8	8	8	8
Aromatics, HPLC (ASTM D6379)	*RMSEP*	0.4789	0.3873	0.3803	0.3794	0.3858
	R^2	0.99	0.99	1.00	1.00	0.99
	Aromatics (of 1st 10)	2	4	4	4	4
Naphthalene Content (ASTM D1840)	*RMSEP*	0.245	0.1808	0.1813	0.2567	0.1811
	R^2	0.89	0.94	0.94	0.88	0.94
	Naphthalenes (of 1st 10)	4	2	2	2	2
Olefins, FIA (ASTM D1319)	*RMSEP*	0.6865	0.708	0.6430	0.7117	0.6451
	R^2	0.88	0.87	0.89	0.87	0.89
	Olefins (of 1st 10)	4	2	4	2	4
Saturates, FIA (ASTM D1319)	*RMSEP*	0.7419	0.7833	0.7592	0.7642	0.7769
	R^2	0.98	0.97	0.97	0.97	0.97
	Saturates (of 1st 10)	5	4	7	6	4

Table 6. UVE-PLS prediction results obtained for five fuel properties while utilizing various peak area thresholds. Most favorable results of three replicates reported for each value. The modeling metric(s) in each row indicative of the most accurate modeling results are highlighted in green.

Table 6 shows that the threshold value of 0.1%, despite providing the most accurate quantitative modeling results, provides the least compositionally accurate modeling results, and is thus not an optimal threshold value to employ within FCAST for reasons already indicated in section 2.2. With respect to the other potential threshold values, as was initially suspected, the choice between them would seem to have little systematic impact on overall fuel property model quality, at least in the context of accurate quantitative modeling predictions. Although different peak area thresholds do indeed provide slightly different prediction results for each of the fuel properties, it cannot be said that any given threshold value provides a clear analytical advantage over the others. However, because the original threshold value of 0.001% provides the highest degree of compositional fidelity to the fuel property being modeled, it was deemed prudent to leave the peak area threshold at this value.

3.3 Other Automation-Friendly Optimizations

FCAST possesses additional minor analysis parameters that should presently be addressed in the context of automation.

Solvent delay. While a solvent delay can be manually entered into the software to minimize adverse effects on distillation curve calculations, this value should remain at a default value of zero. If solvent delays must be accommodated in any given analysis, either the raw data itself can be manipulated using the software used to collect the data in the first place, or the parameter can be changed in FCAST via a less accessible option screen.

Mass Range. Mass spectral mass ranges, to be utilized for NIST/EPA/NIH Mass Spectral Library searches and correlated compound identifications, can also be manually adjusted, but a default option of reading the mass range used from the method file itself, if available, was added. The 35-400 m/z range is left as the default values since those are the recommendations for the preferred method for data collection.

MS Mass Factors. Mass factor corrections were determined empirically in earlier work[10] to account for different mass analyzer ionization efficiencies to convert peak areas to mass percent values. This conversion is necessary to achieve accurate quantitative results and thus remains as a default option in the software.

Multiple Compound Identifications. FCAST allows for the detection of multiple iterations of the same compound across multiple retention times throughout a given data set to allow for the compound-level characterization of specific peaks, if as much were deemed necessary. However, despite the fact that multiple compound identity iterations can be a consequence of column bleed, which may become significant as a column ages or degrades with use, said iterations are not to be expected in most realistic and routine scenarios, simply by virtue of how chromatography

functions. Because multiple identified iterations of the same compound are almost invariably the result of library misidentifications, the default protections currently in place to disallow multiple such compound identifications will remain in place in the software. As the majority of the mismatches are usually due to closely-related isomers, combining these compounds results in limited downstream analysis issues.

Column Bleed and Peak Overloading. FCAST currently has algorithms in place to accommodate both column bleed and overloaded peaks. As with the solvent delay, data collection should be done that limits the need for these options. These options will remain unselected as defaults to allow for the accommodation of as diverse a set of future fuel data sets as possible.

False Compound Identifications. FCAST currently has the option to automatically exclude selected names and partial names from possible compound identities as a default setting. At present, the compounds to ignore are those possessing the following names or partial names: "methylene chloride," "siloxane," "silane," "silicic," "silyl," "trifluoroacetate," and "TMS derivative." As might be expected, the majority of these exclusions explicitly target silicon-containing compounds that are the result of chromatographic column bleed, while methylene chloride is a solvent typically employed in GC-MS analysis work. "Trifluoroacetate" and "TMS derivative," meanwhile, are partial compound names that should be impossible to find in realistic fuel populations. These compounds thus discriminated against provide virtually no meaningful compositional information with respect to the actual fuel being analyzed. Although the option to re-include these compounds in potential FCAST-based analyses will be maintained, to accommodate possible circumstances in which non-fuel samples might be analyzed, their exclusion will remain the default setting.

NIST/EPA/NIH Mass Spectral Library database searches are based on individual mass spectra from parent GC-MS data sets. However, it is possible to obtain individual mass spectra that, due to lack of analyte signal, appear to consist only of a single peak. Such mass spectra produce database search results indicating that some material that also possesses a single peak, such as argon, has been identified. However, because these types of materials are typically nonsensical in the context of fuel analyses, a parameter in the software can be adjusted to disallow database searches utilizing anything less than three distinct m/z peaks. At present, it remains prudent to retain this as a default analysis option.

Distillation Curve Calculations. An alternative automated n-alkane calibration procedure is available in FCAST in order to properly model the distillation curves of samples containing few if any of the n-alkane compounds typically utilized as markers. This alternative procedure will be disabled as a default setting because the end user would be expected to be analyzing samples with sufficient n-alkane compounds to define the retention times of most of these C_6-C_{24} markers.

3.4 Software Optimizations

In addition to the analysis-specific determinations made throughout the present report, several smaller changes were also made to FCAST regarding its overall usability, and some of these are reported upon here. A fuller reporting of such changes, especially as they relate to actual software use, can be found in a concurrent NRL Memorandum Report.[13]

The automated EWFA-MCR algorithm requires that SVD be performed repeatedly, over overlapping data sections, across the entirety of any given GC-MS data set. While SVD does not, in and of itself, require a great deal of time to reach completion on most modern computers, the sheer number of times that SVD must be performed to achieve a comprehensive level of data deconvolution greatly increases overall computational requirements, to the point that the analysis of a GC-MS data set for a single fuel might require hours of calculation time. During the present work, however, an alternative SVD sub-routine was implemented into the deconvolution algorithm that is faster than the previous SVD sub-routine used in FCAST. This alternative SVD sub-routine could also be used to accelerate non-deconvolution FCAST functionalities if an accelerated version of SVD is deemed necessary. It should be noted here that, despite this acceleration step, the deconvolution algorithm still requires hours to fully deconvolve a typical GC-MS data set. However, sample data can be fully deconvolved autonomously on a computer that would otherwise be relatively idle, such as on a workstation running overnight. This time restriction is thus not considered a critical shortcoming for FCAST-based data deconvolution.

Also, to provide end-users with complete information, the masses identified by deconvolution are explicitly highlighted within the visual mass spectral display of the full scan on the FCAST screen. This allows the end user to see which masses are part of the compound and which masses were not used, and most likely part of an overlapping compound.

4.0 Conclusions

The compositional characterization of fuels has always been limited by the inherent limitations of gas chromatography to resolve similar compounds or isomers. One of the primary goals of the present work was thus to overcome this limitation to the extent possible in FCAST by implementing improved algorithms to deconvolve co-eluting peaks. Two deconvolution methods were developed and implemented for the NRL FCAST application, thus providing three options for peak deconvolution: a simple peak detection, a mass channel analysis that provides improved resolution, and an automated EWFA-MCR analysis to provide the highest chromatographic resolution.

A second primary goal was to improve FCAST's automated analysis capabilities by selecting initial default analysis parameters for various operations, including the detailed comparisons of

two fuel populations by the Fisher Ratio ANOVA method and the more streamlined deltaCompare statistical analysis of two GC-MS TICs. These default parameters were specifically selected to allow for a robust automated GC-MS analysis without the need for extensive user training. The peak deconvolution and automation upgrades implemented during the course of this work will improve the results obtained by FCAST to characterize the compositions of fuels in the DoD fuel library, which includes the fuel forensics library.

The improved FCAST software can consequently be provided to DLA Energy for routine use to provide a more detailed and comprehensive analysis of fuels by standard Agilent GC-MS instrumentation and can potentially allow field laboratories to transmit data instead of sending samples to DLA during remediation efforts associated with real-world fuel quality issues.

5.0 Acknowledgements

This work was supported by DLA Energy, program manager Philip Chang.

6.0 Literature Cited

1. Pierce, K.M.; Schale, S.P. Predicting Percent Composition of Blends of Biodiesel and Conventional Diesel Using Gas Chromatography–Mass Spectrometry, Comprehensive Two-Dimensional Gas Chromatography–Mass Spectrometry, and Partial Least Squares Analysis. *Talanta* **2011**, *83 (4)*, 1254-1259.
2. Flood, M.E.; Connolly, M.P.; Comiskey, M.C.; Hupp, A.M. Evaluation of Single and Multi-Feedstock Biodiesel – Diesel Blends Using GCMS and Chemometric Methods. *Fuel* **2016**, *186 (15)*, 58-67.
3. de Carvalho Rocha, W.F.; Schantz, M.M.; Sheen, D.A.; Chu, P.M.; Lippa, K.A. Unsupervised Classification of Petroleum Certified Reference Materials and Other Fuels by Chemometric Analysis of Gas Chromatography-Mass Spectrometry Data. *Fuel* **2017**, *197 (1)*, 248-258.
4. Begue, N.J.; Cramer, J.A.; Bargen, C.V.; Myers, K. M.; Johnson, K.J.; Morris, R.E. Automated Method for Determining Hydrocarbon Distributions in Mobility Fuels. *Energy and Fuels* **2011**, *25 (4)*, 1617−1623.
5. Cramer, J.A.; Begue, N.J.; Morris, R.E. Developing Fundamental Relationships Between Fuel Composition And Performance. *Proceedings of the 12th International Symposium on Stability, Handling, and Use of Liquid Fuels (IASH)*, **2011**.
6. Cramer, J.A.; Hammond, M.H.; Myers, K.M.; Loegel, T.N.; Morris, R.E. Novel Data Abstraction Strategy Utilizing Gas Chromatography−Mass Spectrometry Data for Fuel Property Modeling. *Energy and Fuels* **2014**, *28 (3)*, 1781−1791.

7. Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C. Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry* **1996**, *68 (21)*, 3851−3858.

8. https://www.nist.gov/srd/nist-standard-reference-database-1a-v17 (last accessed August 2018).

9. Morris, R.E.; Hammond, M.H.; Cramer, J.A.; Myers, K.M.; Loegel, T.N. A Fit-For-Purpose Screening Tool Based on Statistical Modeling of Fuel Composition. *Proceedings of the 13th International Symposium on Stability, Handling, and Use of Liquid Fuels (IASH)*, **2013**.

10. Morris, R.E.; Hammond, M.H.; Cramer, J.A.; Myers, K.M.; Loegel, T.N.; Johnson, K.J.; Leska, I.A. The Fuel Composition and Screening Tool (FCAST): A GC-MS Modeling Application for Fuel Characterization. *Proceedings of the 14th International Symposium on Stability, Handling, and Use of Liquid Fuels (IASH)*, **2015**.

11. Hammond, M.H.; Morris, R.E.; Cramer, J.A.; Loegel, T.N.; Johnson, K.J.; Myers, K.M. "Navy Fuel Composition and Screening Tool (FCAST) v.2.5." *NRL Memorandum Report No. NRL/MR/6180—14-9551*, July 18, 2014.

12. Hammond, M.H.; Morris, R.E.; Cramer, J.A.; Loegel, T.N.; Johnson, K.J.; Myers, K.M. "Navy Fuel Composition and Screening Tool (FCAST) v.2.8." *NRL Memorandum Report No. NRL/MR/6180—16-9685*, May 10, 2016.

13. Hammond, M.H.; Morris, R.E.; Cramer, J.A.; Loegel, T.N.; Johnson, K.J.; Myers, K.M. "Navy Fuel Composition and Screening Tool (FCAST) v.3.0." *NRL Memorandum Report*, in preparation.

14. Cramer, J.A.; Begue, N.J.; Morris, R.E. Improved Peak Selection Strategy for Automatically Determining Minute Compositional Changes in Fuels by Gas Chromatography–Mass Spectrometry. *Journal of Chromatography A* **2011**, *1218 (6)*, 824−832.

15. https://chemdata.nist.gov/dokuwiki/doku.php?id=start (last accessed August 2018).

16. Dunkerley, S.; Brereton, R.G.; Crosby, J. A Comparison of Deconvolution Methods as Applied to High Performance Liquid Chromatography–Diode Array Detector–Electrospray Mass Spectrometry of 2- and 3-Hydroxypyridine at Varying pH in the Presence of Severely Tailing Peak Shapes. *Chemometrics and Intelligent Laboratory Systems* **1999**, *48 (1)*, 99–119.

17. Kind, T.; Fiehn, O. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews* **2010**, *2 (1-4)*, 23-60.

18. Manne, R.; Grande, B.V. Resolution of two-way data from hyphenated chromatography by means of elementary matrix transformations. *Chemometrics and Intelligent Laboratory Systems* **2000**, *50 (1)*, 35-46.

19. Xu, C.J.; Jiang, J.H.; Liang, Y.Z. Evolving window orthogonal projections method for two-way data resolution. *Analyst* **1999**, *124 (10)*, 1471–1476.

20. Li, P.; Cai, W.; Shao, X. Generalized window factor analysis for selective analysis of the target component in real samples with complex matrices. *Journal of Chromatography A* **2015**, *1407*, 203-207.

21. Setarehdan, S.K. Modified Evolving Window Factor Analysis for Process Monitoring. *Journal of Chemometrics* **2004**, *18 (9)*, 414-421.

22. Parastar, H.; Tauler, R. Multivariate Curve Resolution of Hyphenated and Multidimensional Chromatographic Measurements: A New Insight to Address Current Chromatographic Challenges. *Analytical Chemistry* **2014**, *86 (1)*, 286−297.

23. Cramer, J.A.; Hammond, M.H.; Loegel, T.N.; Morris, R.E.; Myers, K.M. "Application of Chemometric Methods to Devolve Co-Eluting Peaks in GC-MS of Fuels to Improve Compound Identification: Final Report." *NRL Memorandum Report No. NRL/MR/6180—18-9776*, February 12, 2018.

24. Cramer, J.A.; Hammond, M.H.; Loegel, T.N.; Morris R.E. Evolving Window Factor Analysis-Multivariate Curve Resolution with Automated Library Matching for Enhanced Peak Deconvolution in Gas Chromatography-Mass Spectrometry Fuel Data. *Journal of Chromatography A*, submitted, **2018**.

25. Cramer, J.A.; Hammond, M.H.; Loegel, T.N. Application of Chemometric Methods to Devolve Co-Eluting Peaks in GC-MS Data. *Proceedings of the 15th International Symposium on Stability, Handling, and Use of Liquid Fuels (IASH)*, **2017**.

26. Stein, S.E. An Integrated Method for Spectrum Extraction and Compound Identification from Gas Chromatography/Mass Spectrometry Data. *Journal of the American Society of Mass Spectrometry* **1999**, *10*, 770-781.

27. Johnson, K. J.; Synovec, R. E. Pattern Recognition of Jet Fuels: Comprehensive GC × GC with ANOVA-Based Feature Selection and Principal Component Analysis. *Chemometrics and Intelligent Laboratory Systems* **2002**, *60*, 225-237.