Naval Research Laboratory

Washington, DC 20375-5320

NRL/MR/6180--18-9776

Application of Chemometric Methods to Devolve Co-Eluting Peaks in GC-MS of Fuels to Improve Compound Identification: Final Report

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

Chemical Sensing & Fuels Technology
Navy Technology Center for Safety & Survivability
Chemistry Division

Robert E. Morris
Kristina M. Myers

Nova Research, Inc.
Alexandria, VA

February 12, 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)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
12-02-2018	Memorandum Report	1 Dec 2016 – 1 Dec 2017

4. TITLE AND SUBTITLE

Application of Chemometric Methods to Devolve Co-Eluting Peaks in GC-MS of Fuels to Improve Compound Identification: Final Report

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Jeffrey A. Cramer, Mark H. Hammond, Thomas N. Loegel, Robert E. Morris* and Kristina M. Myers*

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
61-9251-H-7-5

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

Naval Research Laboratory
4555 Overlook Ave, SW
Washington, DC 20375-5320

8. PERFORMING ORGANIZATION REPORT NUMBER

NRL/MR/6180--18-9776

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

Defense Logistics Agency Energy
8725 John J. Kingman Road
Ft. 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

*NOVA Research, Inc., 1900 Elkin Street, Suite 230, Alexandria, VA 22308

14. ABSTRACT

Fuel stability and performance problems are often due to the presence of trace levels of contaminants or other minor changes in composition. Detailed compositional analyses of suspect fuels are often critical to the determination of the cause(s) of the problem(s) at hand. Sensitive methods to compare fuel compositions via GC-MS methods are available, but the detailed compositional analyses of complex fuels are fundamentally limited by the chromatographic resolutions of these methods. Specifically, insufficient chromatographic resolution negatively impacts the quality of downstream compound identifications generated by the database searches utilized to identify chemical compounds via the Navy's FCAST software. The chemometric techniques of EWFA and MCR, NIST database match quality metric comparisons, and other associated algorithms were thus incorporated into an automated GC-MS peak deconvolution strategy suitable for use in software such as the FCAST, allowing for the more accurate and precise determinations of fuel compositions.

15. SUBJECT TERMS

Gas Chromatography-Mass Spectrometry (GC-MS), Peak Deconvolution, Evolving Factor Analysis (EFA), Evolving Window Factor Analysis (EWFA), Multivariate Curve Resolution (MCR), Fuels

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Jeffrey Cramer
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**	SAR	49	**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 Executive Summary 1
2.0 Objective 1
3.0 Approach 1
4.0 Background 2
5.0 Experimental 4
5.1 Software 4
5.2 GC-MS Data Collections 4
5.3 Synthetic GC-MS Data 5
5.4 EWFA-MCR 6
5.5 Dynamic Window Sizing 8
6.0 Results and Discussion 13
6.1 Trace Component Evaluations 13
6.2 Real Fuel Data 18
6.3 Additional Output Constraints and Data Truncation Operations 23
6.4 Comprehensive Evaluation of Updated Algorithm Constraints 30
6.5 Additional Algorithm Validations 38
7.0 Summary 41
8.0 Conclusions 41
9.0 Recommendations 41
10.0 Acknowledgements 42
11.0 Literature Cited 42

FIGURES

Figure 1. Example synthetic data set TIC (black), consisting of a 25/25/25/25 blend of four constituents and noise 6

Figure 2. Compound identification results after applying EWFA and MCR, using multiple analysis window sizes, to convoluted synthetic data sets 9

Figure 3. Example run times of representative iterations of the deconvolution algorithm, as applied to the Gaussian Width 5 data shown in Figure 2, with the MCR option employed, versus arbitrated window size 10

Figure 4. Compound identification results after applying EWFA and MCR, using a dynamically-adjusting window analysis size, to data sets similar to those shown in Figure 1 12

Figure 5. Example noisy synthetic data set TIC (black), consisting of a 30/10/30/30 blend of four constituents, plotted with no noise 14

Figure 6. Select control and +2 algorithm version results from Table 1 reproduced in a graphical format 17

Figure 7. Summary of the number of dodecane identifications obtained from the GC-MS data collected from a petrochemical jet fuel, and the number of tetradecane identifications obtained from the data collected from a petrochemical diesel fuel and two petrochemical MGO fuels, both without (red) and with (green) the application of the EWFA-MCR deconvolution algorithm 20

Figure 8. First half of a summary of the number of various bioactive molecule identifications obtained from the GC-MS data collected from a petrochemical jet fuel, a petrochemical diesel fuel, and two petrochemical MGO fuels, both without (red) and with (green) the application of the EWFA-MCR deconvolution algorithm 21

Figure 9. Second half of a summary of the number of various bioactive molecule identifications obtained from the GC-MS data collected from a petrochemical jet fuel, a petrochemical diesel fuel, and two petrochemical MGO fuels, both without (red) and with (green) the application of the EWFA-MCR deconvolution algorithm 22

Figure 10. Visualization of estimated individual compound contributions, found via peak deconvolution, to a selected retention time range in a JP-5 GC-MS data set, and the TIC peaks therein 40

Figure 11. Visualization of estimated individual compound contributions, found via peak deconvolution, to a selected retention time range in a Jet A GC-MS data set, and the TIC peak therein 40

TABLES

Table 1. Summary of trace compound identification results before and after applying EWFA-MCR 16

Table 2. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the JP-5 jet fuel…. 26

Table 3. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the F-76 diesel fuel 27

Table 4. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the Miami MGO fuel 28

Table 5. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the Singapore MGO fuel 29

Table 6. EWFA-MCR-augmented, FCAST-type compositional profiling results obtained from the four fuels described previously, before and after changing the area-based thresholding constraint 32

Table 7. EWFA-MCR-augmented, FCAST-type compositional profiling results obtained from the JP-5 jet fuel, before and after changing the significance-based and loading-based thresholding constraints 33

Table 8. EWFA-MCR-augmented, FCAST-type compositional profiling results obtained from four fuels, before and after changing the significance-based and loading-based thresholding constraints 35

Table 9. EWFA-MCR-augmented, FCAST-type heteroatomic profiling results obtained from four fuels, before and after changing the significance-based and loading-based thresholding constraints 37

Table 10. Total numbers of compounds and associated peak areas found across ten replicates, collected for each of four parent diesel fuels, before and after stressing as prescribed in ASTM D5304 39

ABBREVIATIONS

ALS	Alternating Least Squares
ANOVA	Analysis Of Variance
ASTM	American Society of Testing and Materials
DLA Energy	Defense Logistics Agency Energy
DOD	Department Of Defense
EFA	Evolving Factor Analysis
EPA	Environmental Protection Agency
EWFA	Evolving Window Factor Analysis
FCAST	Fuel Composition and Screening Tool
GC	Gas Chromatography
GC-MS	Gas Chromatography with Mass Selective detection (i.e. Mass Spectrometry)
GCxGC-MS	Hyphenated (Two-Dimensional) Gas Chromatography - Mass Spectrometry
LPR	Low Pressure Reactor
LV	Latent Variable
MCR	Multivariate Curve Resolution
MF	Match Factor
MGO	Marine Gas Oil
NIH	National Institute of Health
NIST	National Institute of Standards and Technology
SVD	Singular Value Decomposition
TIC	Total Ion Chromatogram
ULSD	Ultra Low Sulfur Diesel

1.0 Executive Summary

Fuel stability and performance problems are often due to the presence of trace levels of contaminants or other minor changes in composition. Detailed compositional analyses of suspect fuels are often critical to the determination of the cause(s) of the problem(s) at hand. Sensitive methods to compare fuel compositions via GC-MS methods are available, but the detailed compositional analyses of complex fuels are fundamentally limited by the chromatographic resolutions of these methods. Specifically, insufficient chromatographic resolution negatively impacts the quality of downstream compound identifications generated by the database searches utilized to identify chemical compounds via the Navy's FCAST software. The chemometric techniques of EWFA and MCR, NIST database match quality metric comparisons, and other associated algorithms were thus incorporated into an automated GC-MS peak deconvolution strategy suitable for use in software such as the FCAST, allowing for the more accurate and precise determinations of fuel compositions.

2.0 Objective

The objective of the current study is to develop an unsupervised, chemometrics-based strategy, suitable for robust implementation in automated software applications, to enhance and improve fuel composition profiling by GC-MS by leveraging databased mass spectral information to seamlessly and autonomously deconvolve co-eluting fuel components.

3.0 Approach

The objective of this study was met through the following technical approach:

- Develop data sets of synthetic GC-MS spectra, with known numbers and types of co-eluting compounds, by mathematically combining well-characterized mass spectral data from the NIST/EPA/NIH Mass Spectral Library database.
- Develop algorithms for EWFA-based peak component identifications and MCR-based refinements.
- Test and optimize the combined EWFA-MCR algorithm with synthetic data, varying the chromatographic resolution between adjacent peaks and the spectral similarity between co-eluting compounds.
- Continue to test and optimize the overall analysis strategy with real GC-MS data collected from varied fuel samples.
- Implement peak deconvolution algorithms into the FCAST software.

Manuscript approved January 23, 2018.

4.0 Background

GC remains a primary tool for the analysis and compositional characterization of mobility fuels, both inside and outside of the Navy and the DOD. As the requirements for the accurate and precise analytical data required for state-of-the-art applications increase, the performance boundaries of fuel chromatography become a significant limiting factor. This limitation becomes increasingly important in fuel failure investigations, where minor or trace compositional artifacts are often responsible for poor fuel stability or performance.

Fuel chromatography is inherently limited by the high complexity of petroleum fuel compositions. In practice, almost none of the fuel constituents are fully resolved from other components. This is due to a combination of 1) insufficient peak capacity for the large number of individual components within time and chromatographic efficiency constraints and 2) insufficient resolving power of the stationary phase in the gas chromatography column relative to the many structurally similar isomers or homologs present in fuel. Multidimensional approaches, longer columns and slower heating rates can offer some benefits but will not necessarily fully resolve previously co-eluting fuel compounds.

As the mass analyzer samples the effluent from the GC column in GC-MS applications, the mass fragments detected in a particular time slice can be sent to the NIST/EPA/NIH Mass Spectral Library database to determine the identity of the chemical being carried by the parent peak. When two (or more) compounds co-elute, their peaks will overlap and the time-slices sent to the mass analyzer will correspondingly contain contributions from the co-eluting compounds. Depending on the degree of overlap, the NIST database search algorithms will produce one of two results. If the mass spectra are completely inseparable, incorrect assignments will simply be returned from the database search, resulting in errors in downstream information streams. Even in the case of incomplete overlap, however, the mass spectra from one or more compounds may very well be overwhelmed by competing mass spectra, resulting in information loss. This latter phenomenon can be exacerbated by the random elution of siloxanes and other stationary phase hydrocarbon fragments over the course of the GC analysis, often referred to as column bleed. The analyst will, of course, not see the latter obscured compounds in any database search results, but will see the former occurrences of misidentified compounds. Often, the mangled mass spectra resulting from co- elution will return compound identifications that are nonsensical and can be discarded either manually or by means of MF-based quality metrics, but there are also instances where high MF values are obtained from erroneous database matches. These types of co-elution artifacts are also responsible for the reporting of multiple, widely-distributed instances of the same compound throughout a single chromatographic analysis, which have the potential to confuse analysis results.

Any investigation, either *ad hoc* or routine, that depends upon the gas chromatographic analysis of fuel constituency would obviously benefit from improved compound peak resolution, especially if it can be made to occur in a reliable and automated fashion without requiring users to tailor its functionality to any given data peak. This includes fuel failure analysis, compositional database construction, the elucidation of contaminants and the detection of adulterants, and accurate compound-level or compositional-level class profiling. Further, while the present study focuses on GC-MS, the methods that have been developed remain applicable to GCxGC-MS data sets as well, should the enhanced compositional resolution available via this analytical technique be desired. Even GC techniques that do not utilize mass spectral detection can be augmented with those aspects of the present work not directly associated with MF-based quality metrics.

The present study's focus on algorithm development for automated GC-MS data deconvolution is the function of an explicit goal of FCAST implementation. The FCAST is a comprehensive software package that already extracts a wide variety of information from GC-MS data using mathematical, statistical, and chemometric modeling strategies, including detailed compositional assessments, estimates of critical fuel properties, and calculated distillation curves for individual fuel samples, as well as composition-based comparisons of fuel sample pairs.[1,2] These features can individually be very useful when comprehensive fuel analyses are required but 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 previous 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. Implementing automated peak deconvolution into the FCAST software will only enhance the Navy's overall fuel analysis capabilities.

The presently reported study focuses on the end goal of the seamless implementation of EWFA- and MCR-based techniques into the FCAST software and/or a streamlined, standalone software application for use when the full FCAST software is either not available or unnecessary, in a manner that would result in optimal performance while remaining user-friendly. Broadly, the developed approach uses EWFA to estimate the number and location of individual chemical components within local regions containing convolved chromatographic peaks, followed by MCR to estimate the shapes of pure-component mass spectra which will be used to provide more effective database searches and thus more accurate compositional profiles.

Because the peak deconvolution algorithm will be required to function effectively under varied circumstances with minimal input or interpretation required by the end user, a key focus of this work was in developing an automated method to define acceptance/rejection thresholds, as well as other parameters, that will, in turn, be used to define where along the retention time axis individual chemical components appear and disappear within a convoluted chromatographic peak. A data-

driven approach to parameter selection was pursued, allowing the algorithm a measure of adaptability, and by extension greater robustness, without requiring user interaction. In other words, these thresholds and parameters were intelligently set using dynamic information derived from the data itself.

Synthetic GC-MS data were generated for initial algorithm development purposes, and producing these data required a smaller, interconnected research effort to be briefly reported upon herein. This synthetic data allowed for a rigorous determination of the conditions under which automated EWFA-based peak component identifications and MCR decompositions become unreliable for fuel profiling, as well as predictions of critical threshold values. There are two key parameters that were varied for this synthetic data: chromatographic resolution between adjacent peaks and spectral similarity between co-eluting peaks. These parameters together serve to describe an arbitrary level of sample complexity within the context of this work. Effectiveness of peak deconvolution as a function of sample complexity was measured both by comparing the output of the deconvolution algorithm directly against the "ground truth" of the parameters of the simulation as well as by comparing the accuracy of control output results. Once this was demonstrated, real fuel data were employed.

Because of the data-driven methodology used to develop the overall peak deconvolution strategy, the initial development of both synthetic data production methods and the fundamental operations of the combined EWFA-MCR algorithm will be described in the Experimental section, with further testing steps that may or may not have resulted in further algorithm optimizations being described later in the Results and Discussion section.

5.0 Experimental

5.1. Software

Synthetic data production and data analysis algorithms were developed using MATLAB (MathWorks, Inc., Natick, MA), the PLS_Toolbox for MATLAB (Eigenvector Research, Inc.), and the MATLAB based Calibration and Standard Toolboxes (FABI/ChemoAC Consortium). When necessary and appropriate, both mass spectral data and derived mass spectral loadings were submitted to the NIST Mass Spectral Search Program for the NIST/EPA/NIH Mass Spectral Library (NIST 11, version 2.0g) for identification purposes. All data subsets were mean-centered prior to each individual SVD operation, but were otherwise not preprocessed.

5.2 GC-MS Data Collections

GC-MS data sets collected from the real-world fuel samples presented herein were acquired on an Agilent 5890 GC with an Agilent 5971 mass selective detector. Injections (1.0 μL) of the neat fuels

were made with an autoinjector into a split/splitless inductor at a split ratio of 200:1. An AT-1 cross-linked polysiloxane capillary column (50 m x 0.25 mm ID, 0.20 μm film thickness) was used with an oven temperature program that initiated data collection at a temperature of 40 ˚C and ramped the temperature at 10 ˚C/min to 290 ˚C, holding this temperature for the remainder of the data collection. Data were acquired at column retention times from 5.1 to 70.0 minutes, at a rate of about two mass spectra per second, from 35 to 400 m/z.

5.3 Synthetic GC-MS Data

Work was performed to produce useful (i.e. both realistic and challenging) synthetic data sets for the purposes of algorithm development. It had previously been known internally that phenols tend to convolute with one another during the GC-MS analysis of fuel samples, so work initially focused on convoluting, in various proportions, the spectra of four phenols possessing similar chemical structures: p-tert-butylphenol, 2-(1,1-dimethylethyl)phenol, 2-(1-methylethyl)phenol, and 2-propylphenol. The initial, unblended mass spectra for these compounds were located in the NIST mass spectral database, normalized to unit area, and adjusted by the most appropriate mass spectral response factors determined during previous internal work on our own equipment. This last step was undertaken simply to ensure that arbitrated blend ratios would accurately translate to what one would see if an actual blend possessing these ratios were to be analyzed using GC MS in our laboratories.

After the mass spectral data vectors are thus preprocessed, they can then be individually cross-multiplied by vectors containing Gaussian curves to produce data matrices consisting of mass spectra and retention time axes. The use of Gaussian curves accurately represents the normally distributed manner in which mass spectral data can appear across multiple retention times.[3] Because chromatographic columns can vary a great deal in separation characteristics, the locations of the Gaussian curves in the chromatographic vector can be arbitrated by the end-user. This, in turn, very easily allows end-users to arbitrate that the two-dimensional data sets for individual compounds will possess identical or nearly-identical synthetic column retention times to result in co-elution. Once the appropriate two-dimensional data sets are constructed, they are multiplied by the desired blend ratios, then added to a separate noise matrix, itself being a convolution of random noise (which can be represented by a normalized distribution of random numbers[3]) and a Gaussian distribution that simulates the hump-like background seen in realistic fuel samples. This initial development work only requires the production of data sets possessing a single convoluted peak of interest, such as are represented by the example TICs seen in Figure 1 (which shows synthetic data consisting of a blend of the four phenols described previously, at a relative concentration of 25% each), but multi-peak data sets can easily be produced by adjusting production parameters.

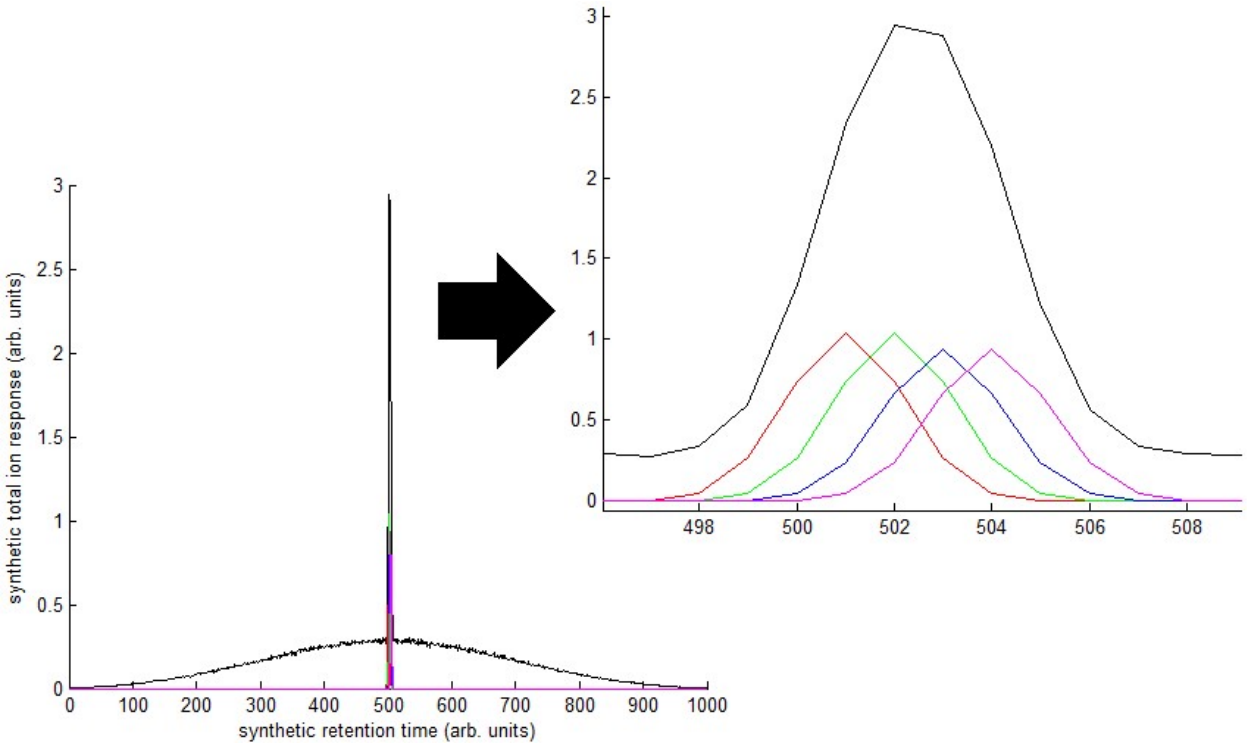

Figure 1. Example synthetic data set TIC (black), consisting of a 25/25/25/25 blend of four constituents and noise. The TIC of the p-tert-butylphenol without noise is plotted in red, the TIC of the 2-(1,1-dimethylethyl)phenol without noise is plotted in green, the TIC of the 2-(1-methylethyl)phenol without noise is plotted in blue, and the TIC of the 2-propylphenol without noise is plotted in magenta. The lack of noise results in a complete overlap of the individual component plots with the x-axis for the majority of retention times, hence the overwhelming magenta x-axis coloration in the non-inset portion of the figure. Some components are higher in magnitude than others due to the instrument response factors described in the text. All components have a Gaussian width of 7 variables, and the individual components were staggered by a single variable along the retention time axis.

5.4 EWFA-MCR

EWFA is a modified version of EFA[4] that basically functions by performing EFA, in the form of repeated SVD operations, on subsets of the overall data set. These data subsets, in the case of GC-MS data, are subdivided along the retention time axis because the decomposed results

corresponding to the mass spectra can be used in NIST database searches to identify the chemical sources of underlying linear data factors.

SVD mathematically breaks a given data matrix down into its underlying linear factors, which are typically designated as scores, loadings, and singular values. The singular values in particular can be used to indicate how many of the factors in the decomposed results can be considered prominent, which is important to know in most applications of EFA. Specifically, in EFA, SVD is 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.

The theory behind EFA is that the number of prominent singular values found through SVD will increase when the ever-increasing subsets of data being analyzed possess more underlying factors than previous subsets. By going in both a forward and reverse direction, one can thus pinpoint where factors within a data set appear, disappear, and overlap, all of which constitute important information when assessing convoluted GC-MS data. In addition, the loadings collected along with these singular values are also valuable information, as they will (at least ideally) closely correspond to the actual shapes of the underlying factors themselves, i.e. the mass spectra required to identify individual components by means of a NIST database search. The use of EWFA further refines EFA results by moving a window within which to perform EFA across a larger data set, allowing for a more nuanced understanding of these appearances, disappearances, and overlaps, especially in the context of discrete data peaks.

It must be clarified here that the EWFA algorithm, based on repeated iterations of SVD, is capable of producing multiple mass spectrum-like loadings per iteration, each corresponding to a different LV and a different relative prominence in the parent data set. This, of course, allows for the deconvolution of multiple compounds from even closely overlapping data, which is part of EWFA's core functionality. However, loading shapes might not be well-resolved from the parent data, at least in part due to their relative prominences in said data. To compensate, MCR[5] can be included in the use of EWFA, functioning as a means by which to refine the shapes of the loadings produced using EWFA, typically by means of an iterative ALS algorithm. This allows the sub-optimal shapes of some loadings, distorted by factors such as their lower relative prominences as found via SVD operations, to more accurately reflect underlying chemical phenomena, thus allowing more fully obfuscated chemical components to be effectively identified via NIST database searches.

Although the aforementioned singular values would most likely find substantial use in applications not utilizing mass spectra, including those existing in-house, the combined EWFA-MCR algorithm being produced in the present work has been designed to utilize the information produced via NIST database searches in a novel, alternative manner. This was done because NIST database searches provide an internal metric that can be used to interrogate loadings in a manner more directly related to composition than that which would be provided by their corresponding singular values. Specifically, because NIST database searches provide MF values to quantify identification quality, the loadings can effectively be ranked by how accurately they represent the mass spectra contained within convoluted data, under the assumptions that 1) high-quality matches are unlikely to be primarily superfluous data artifacts, and 2) the mass spectra found in the NIST database and those collected in-house are similar enough that comparisons between the two populations will provide meaningful MF values. Although the manner in which underlying factors evolve within the data is thus not explicitly being modeled thanks to this disregard of singular values, the fact of their evolution remains critical to the analysis, justifying this novel technique's continued designation as an EFA variation.

5.5 Dynamic Window Sizing

The phenol-based synthetic data already shown in Figure 1 are considered very convoluted data because the maximum values of the four individual component peaks all fall within at least some portion of each of the other three component peaks, all within the space of ten total retention time variables. Additionally, the four convoluted phenols to be evaluated all possess similar chemical structures, which should provide a non-trivial analysis challenge upon which to evaluate potential algorithms. A preliminary investigation was thus undertaken to deconvolve not only data with these parameters, but also similar data produced using peak widths of 3 and 5 (expected to be representative of more routine convoluted data). This portion of the investigation focused upon which EWFA window size would be most appropriate for the present work, which at least initially seemed like a straightforward early decision to make. The window sizes were chosen to all be odd numbers to ensure that a central variable is available for each window to allow the results collected within said window to be easily correlated to a distinct central retention time.

Interestingly, however, as can be seen in Figure 2, assigning the peak width parameter is not as straightforward as it might initially seem because larger peaks require larger window sizes to be more effectively deconvolved. While this might mean that the safest and most thorough course of action would be to simply apply the maximum possible window size to the entire EWFA analysis area, no small amount of consideration should be given to calculation times, which, as can be seen in Figure 3, increase significantly when larger window sizes are employed.

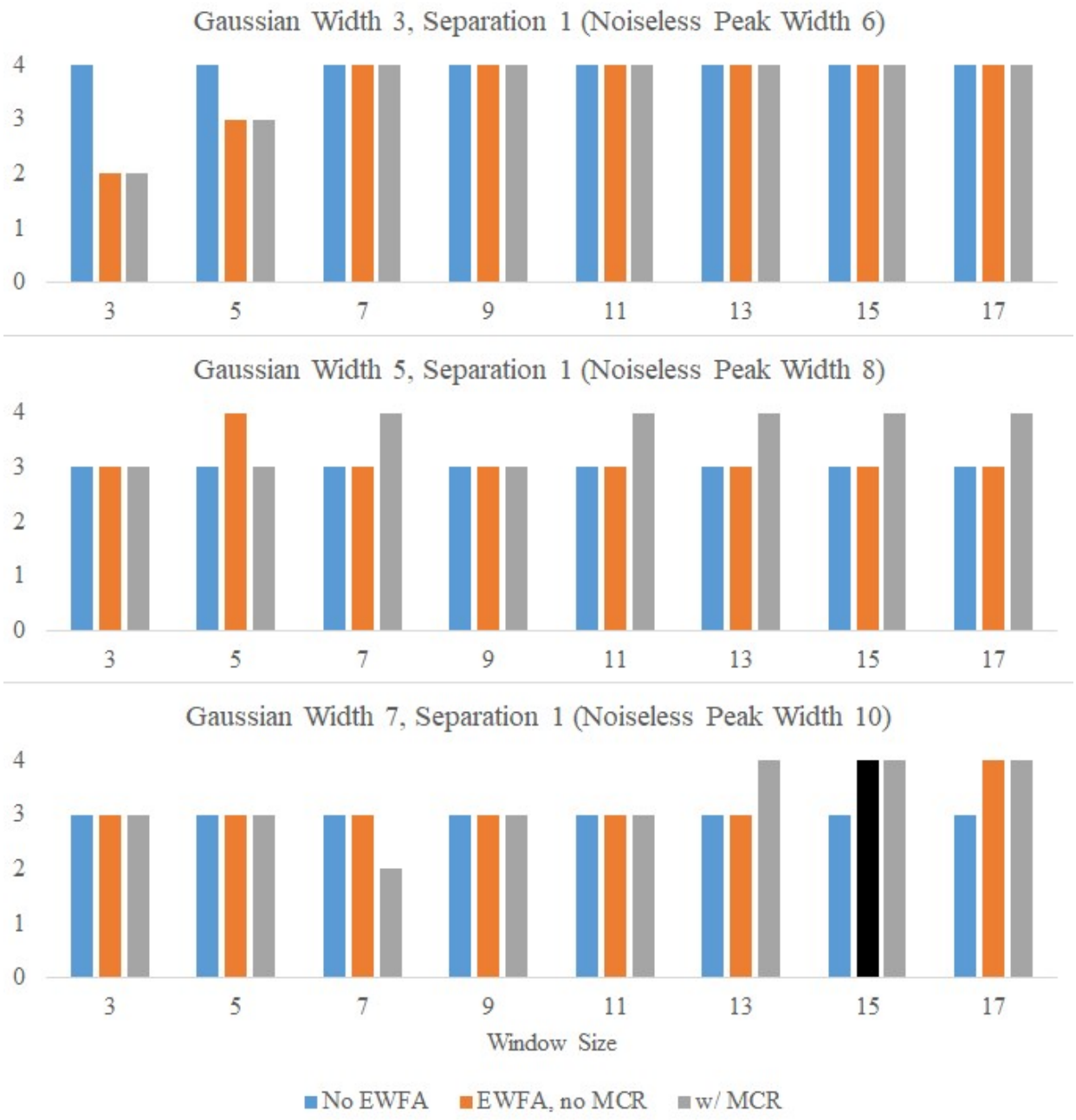

Figure 2. Compound identification results after applying EWFA and MCR, using multiple analysis window sizes, to convoluted synthetic data sets. These results represent the number of phenols detected out of four possible. *Black column additional information*: The compound 4-(1-methylethyl)phenol was detected in lieu of 2-(1-methylethyl)phenol, which is considered acceptable in the present circumstances given how similar the mass spectra of these two compounds would be to one another even in non-convoluted data.

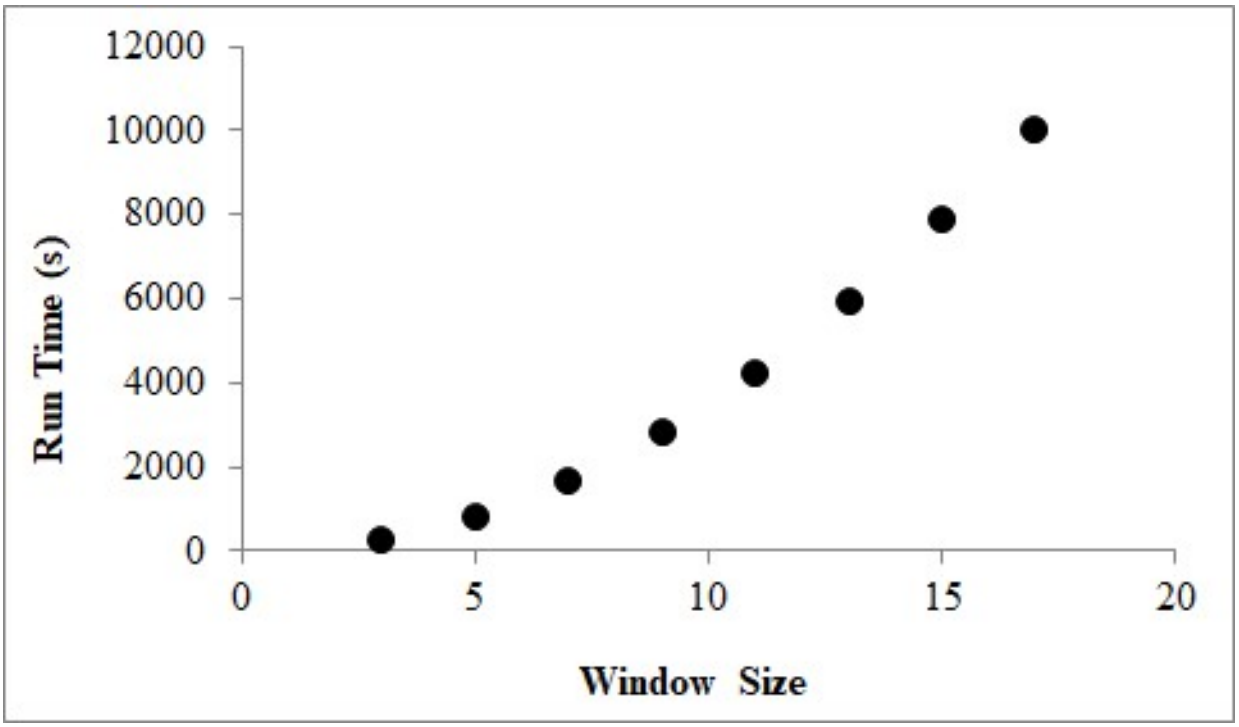

Figure 3. Example run times of representative iterations of the deconvolution algorithm, as applied to the Gaussian Width 5 data shown in Figure 2, with the MCR option employed, versus arbitrated window size. The same computer was used for all run time measurements.

Given the balance that must therefore be struck between thorough window sizes and increased calculation times, the decision was thus made to develop a strategy to dynamically resize analysis windows depending upon the most obvious data structures to be assessed: the peaks in the TICs of the GC-MS data. Fortunately, the FCAST software already quantifies peak lengths along the retention time axis. This pre-existing information can be leveraged to dynamically adjust window sizes to each peak, which will, in turn, provide for an adaptive strategy suitable for automation.

Consider again the results shown in Figure 2, obtained from synthetic data consisting of blends of the four phenols described previously, at relative concentrations of 25%, utilizing peak widths of 3, 5, and 7 variables, in addition to a peak separation of a single variable, along the retention time axis, resulting in three data sets. When applying MCR, a minimum window size of 11 is necessary to acceptably analyze the four-component peak with a total width of 8 variables (Gaussian Width 5 data, 11-8=3), and a window size of 13 is necessary to acceptably analyze the four-component peak with a total width of 10 variables (Gaussian Width 7 data, 13-10=3). The Gaussian Width 3 data analysis seems to be a bit more forgiving overall and is thus not being considered at present. Although adding 3 to the width of dynamically-selected window sizes would thus seem to be in order, peak lengths would necessarily be defined as odd numbers within the algorithm (rounding even numbers up to odd numbers if necessary) due to the use of central variables for indexing purposes. Because adding an odd number to such a length would eliminate these central variables, dynamically adjusting window sizes of both 2 and 4 plus the underlying TIC peak length should be evaluated here to compare their relative merits in the context of this automated analysis.

The +2 and +4 versions of the dynamic-window EWFA-MCR algorithm were thus applied to synthetic data sets similar to those just described, aside from the fact that the individual chemical component peaks began at a distance of 5 variables from each other and were brought closer together in increments of 1 variable to determine if and when the algorithm would become ineffective at uncovering underlying chemical information. These updated compound identification results, found in Figure 4, do not evaluate EWFA without MCR because Figure 2 indicates that applying MCR reliably improves upon EWFA results in the present work. Concurrently, EWFA alone will not be revisited in the remainder of this report, and the EWFA-MCR designation will be used to refer to the combined algorithm.

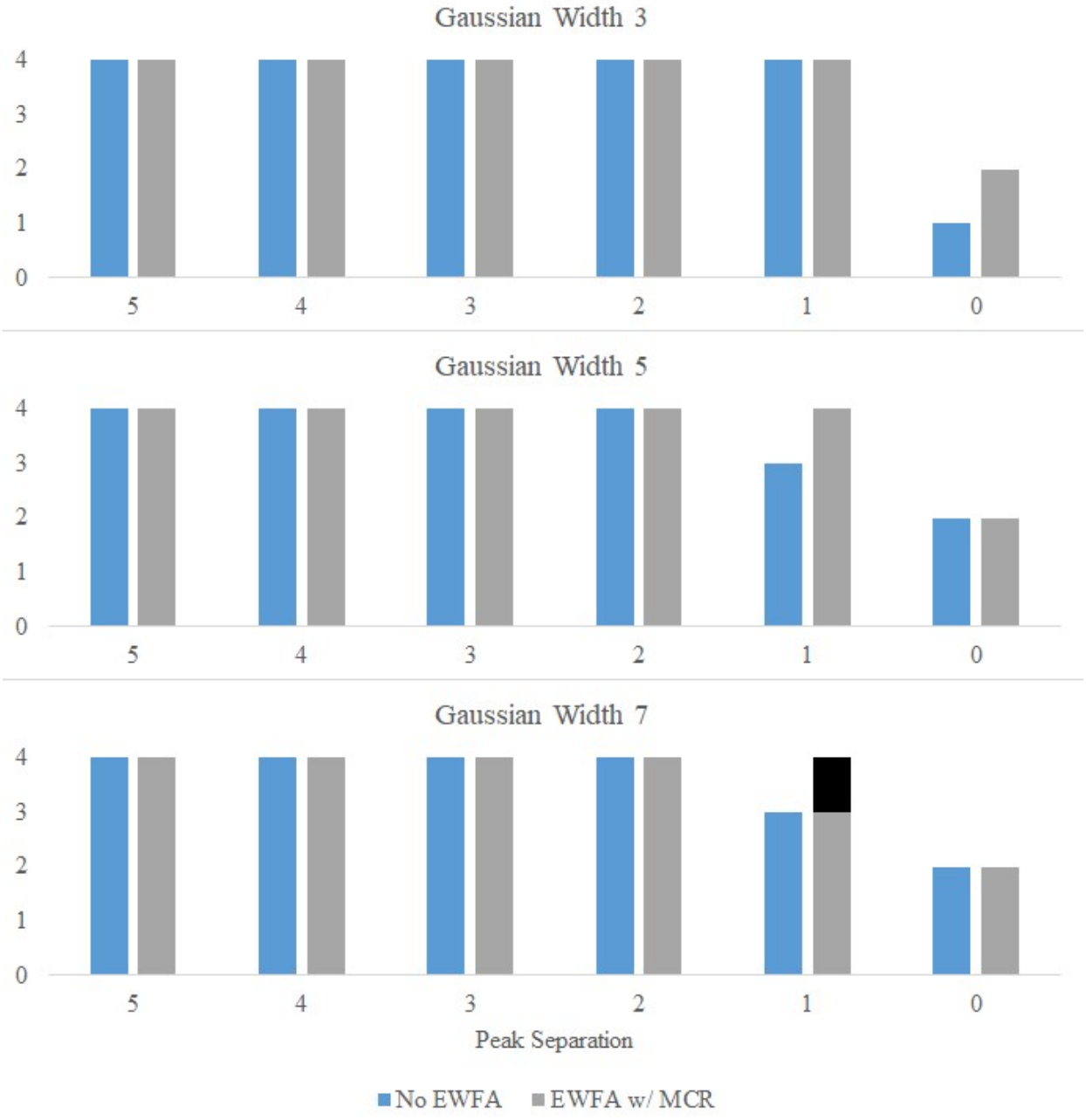

Figure 4. Compound identification results after applying EWFA and MCR, using a dynamically-adjusting window analysis size, to data sets similar to those shown in Figure 1. These results represent the number of phenols detected out of four possible. The only difference between using a size addition of +2 and +4 is indicated by the partially blackened column, which indicates where the +2 size addition identifies 4 phenols and the +4 size addition identifies 3 phenols.

As one might expect, unacceptable degrees of data convolution do not truly appear until the individual component peaks are quite close to one another, though the results seen when using peak separations of 2 through 5 are still of some use insofar as they show that the deconvolution algorithm is not inadvertently causing compound identifications to decrease in quality through data over-processing. Interestingly, the use of either +2 or +4 window size additions produces almost identical results, though the use of the +2 version of the algorithm does uncover one additional compound identity and is thus considered superior in the context of dynamically-adjusting EWFA-MCR. The fact that smaller window sizes would also result in accelerated analyses is also considered a non-trivial side benefit.

Although Figure 4 does show that the algorithm cannot reliably deconvolve peaks that are completely overlapping with one another (i.e. Peak Separation 0), the +2 EWFA-MCR dynamic-window algorithm is capable of extracting three chemical identities that would have remained unavailable otherwise and will thus be further optimized in the remainder of this report. This optimization will take the form of additional data-driven trials that will be used to modify the EWFA-MCR algorithm thus far established.

6.0 Results and Discussion

6.1 Trace Component Evaluations

The combined EWFA-MCR algorithm thus far developed, which dynamically defines analysis window sizes as the full span of any given underlying peak plus two variables, has been trained upon a single, albeit difficult, family of analysis challenges involving multiple convoluted phenols. In order to thoroughly test, evaluate, and optimize the algorithm, the first decision that was made was to apply it to a much wider range of convoluted data sets. In addition to simply varying the compounds to be convoluted with one another, this wide-ranging synthetic data-based evaluation work was also reconfigured to focus upon the algorithm's abilities to ascertain the existence of trace components from larger parent peaks, an important capability for the deconvolution algorithm to possess.

Synthetic data sets were thus produced each consisting of a four-component blend of various organic chemical compounds at a combined blend ratio of 30/10/30/30, with the 10% component being defined as the trace component to explicitly be detected. To further convolute the trace component, the two 30% component peaks immediately surrounding the trace component peak were given a width of 5 and the fourth component peak was given a width of 7, thereby allowing all three non-trace components to interfere with the width-3 trace peak. Figure 5 shows an example of what such a four component TIC peak would look like.

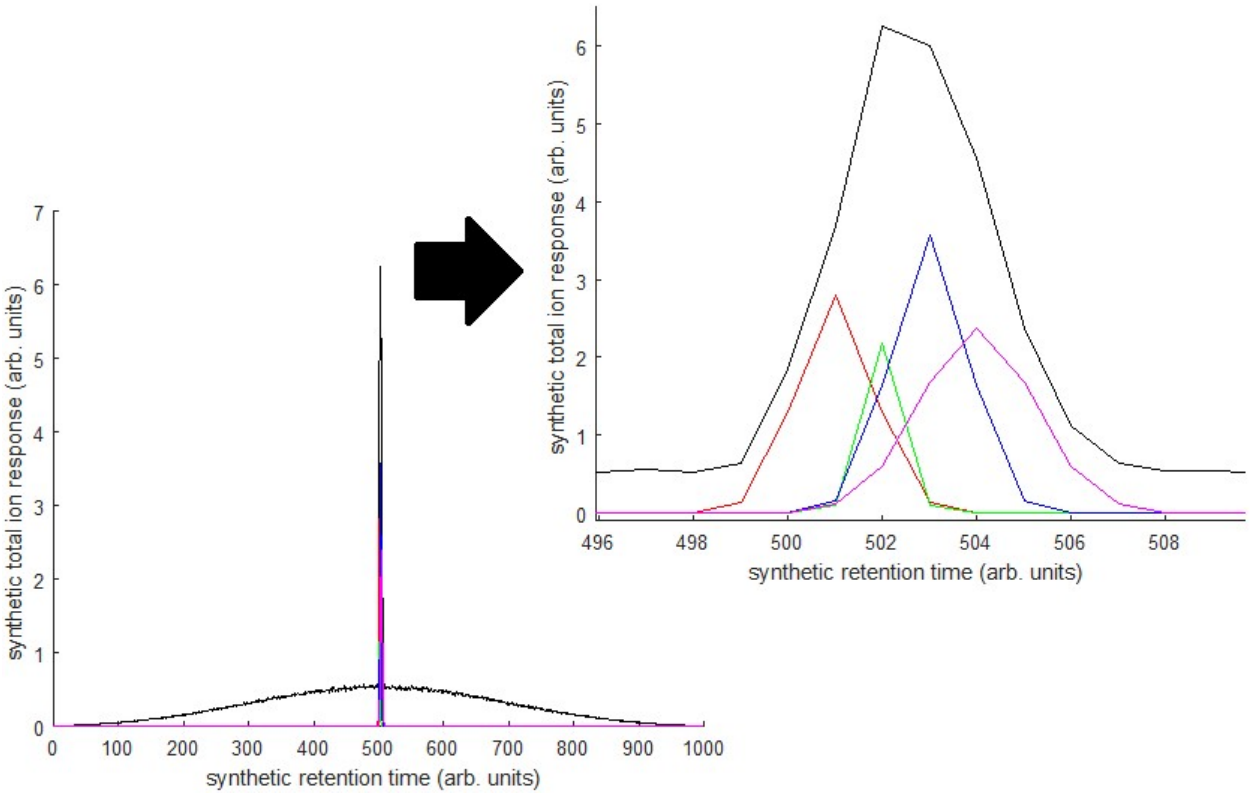

Figure 5. Example noisy synthetic data set TIC (black), consisting of a 30/10/30/30 blend of four constituents, plotted with no noise. The lack of noise results in a complete overlap of the individual component plots with the x-axis for the majority of retention times, hence the overwhelming magenta x-axis coloration in the non-inset portion of the figure. The TIC of the trace component is plotted in green.

Although considerations were given to producing data sets with multiple peaks along the full lengths of their TICs for simultaneous analyses, it was quickly determined that peaks appearing at locations further from the middle of the major noise hump, such as that which is apparent in both Figure 5 and Figure 1, simply possessed lower amounts of noise, with no additional analysis challenges that required overcoming. Therefore, each data set shown in the present reporting was made to possess only a single, centrally-located TIC peak requiring deconvolution in order to ensure that interferences from noise were not mitigated.

The specific chemical compounds upon which to base these convoluted data sets were determined, for each data set, by first randomly selecting an initial compound from the NIST spectral database itself. These randomly selected compounds were restricted to those composed only of any number of carbon atoms, any number of hydrogen atoms, one oxygen atom, one nitrogen atom, and/or one sulfur atom, as chemical compounds possessing other potential components, such as azo compounds and halogenated compounds, were deemed to be outside of the primary scope of the present fuel analysis-based work. The chemical formula for the initial compound thus selected was then used to randomly select three additional isomers of the initial compound from the same database. These four compounds were then used in subsequent convoluted data constructions, with the actual trace component to be detected being randomly selected from the four compounds. Once 100 data sets were created under these conditions, the combined EWFA-MCR algorithm thus far produced in the present work was applied to them to determine its effectiveness at identifying the presence of the trace component. Furthermore, multiple match factor (MF) acceptance/rejection thresholds were applied to determine their effects on the overall analysis strategy, with the best (i.e. highest) MF value being selected if multiple identifications were made. In addition, as a further validation of the work detailed previously, a +4 version of the program was evaluated alongside to the primary +2 version during this operation. A full summary of the results of this operation can be found in Table 1, and Figure 6 presents selected control and +2 results in a graphical format.

Table 1. Summary of trace compound identification results before and after applying EWFA-MCR. In those cases in which the +2 and +4 versions of the combined EWFA-MCR algorithm produce different results using the same MF threshold, the result indicating a greater trace detection capability is highlighted in bold italics.

	Best MF > 600			Best MF > 700			Best MF > 800		
Trace Component	**w/o EWFA-MCR**	**w/ EWFA-MCR (+2 ver.)**	**w/ EWFA-MCR (+4 ver.)**	**w/o EWFA-MCR**	**w/ EWFA-MCR (+2 ver.)**	**w/ EWFA-MCR (+4 ver.)**	**w/o EWFA-MCR**	**w/ EWFA-MCR (+2 ver.)**	**w/ EWFA-MCR (+4 ver.)**
only carbon/hydrogen									
Total Identified (out of 28):	2 (7%)	17 (61%)	***18 (64%)***	2 (7%)	17 (61%)	17 (61%)	1 (4%)	15 (54%)	***16 (57%)***
with nitrogen									
Total Identified (out of 17):	4 (24%)	11 (65%)	11 (65%)	4 (24%)	***11 (65%)***	10 (59%)	2 (12%)	***9 (53%)***	8 (47%)
with oxygen									
Total Identified (out of 38):	11 (29%)	29 (76%)	***30 (79%)***	10 (26%)	29 (76%)	***30 (79%)***	4 (11%)	24 (63%)	***28 (74%)***
with sulfur									
Total Identified (out of 2):	1 (50%)	2 (100%)	2 (100%)	1 (50%)	2 (100%)	2 (100%)	1 (50%)	2 (100%)	2 (100%)
with nitrogen and oxygen									
Total Identified (out of 8):	3 (38%)	***8 (100%)***	7 (88%)	2 (25%)	***8 (100%)***	7 (88%)	0 (0%)	***6 (75%)***	4 (50%)
with nitrogen and sulfur									
Total Identified (out of 2):	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)	2 (100%)
with oxygen and sulfur									
Total Identified (out of 1):	1 (100%)	1 (100%)	1 (100%)	0 (0%)	**1 (100%)**	0 (0%)	0 (0%)	***1 (100%)***	0 (0%)
with nitrogen, oxygen, and sulfur									
Total Identified (out of 4):	2 (50%)	3 (75%)	3 (75%)	2 (50%)	3 (75%)	3 (75%)	0 (0%)	***3 (75%)***	2 (50%)

Figure 6. Select control and +2 algorithm version results from Table 1 reproduced in a graphical format. Unselected results primarily excluded because graphically comparing their percent identification rates across multiple MF values would not be overly informative.

It should, of course, first be noted that the EWFA-MCR algorithm is not completely effective at uncovering the presence of all 100 trace components, as one might expect given the challenging nature of the parent data sets. Nonetheless, the algorithm allows for significant improvements in trace component detection. This is particularly striking in the case of non-heteroatomic compounds (i.e. those containing only carbon and hydrogen atoms), where an almost trivial ability to detect trace components (7% detection rate) is improved to the point of becoming decidedly non-trivial (61% detection rate). Also, as would be consistent with previous results, employing the +4 version of the algorithm results in no clear improvements in trace detection in the present evaluation. As the +4 version of the algorithm additionally increases calculation times by roughly 50% per data set over the +2 version of the algorithm, the decision arrived at previously to proceed with the +2 version as the primary algorithm appears to be well-founded.

Interestingly, the use of either 600 or 700 as an MF threshold value produces identical trace detection results when applying the +2 algorithm. It is only when the value of 800 is employed that a higher level of discrimination becomes apparent, though a threshold of 800 would already be a rather stringent threshold to employ when using software such as the FCAST.

6.2 Real Fuel Data

The combined EWFA-MCR deconvolution algorithm, thus far in the reporting, has only been trained and evaluated with synthetic data sets. Though these data sets have clearly presented the algorithm with non-trivial challenges, both during fundamental development and in the post-development testing previously performed, it was decided that experimental GC-MS data collected from real-world fuel samples were required for further evaluations. To this end, GC-MS data sets were retrieved, from our internal fuel archives, for four real-world fuel samples: a petrochemical JP-5 jet fuel, a petrochemical F-76 diesel fuel, and two petrochemical MGO fuels, one obtained from a location associated with Miami, and one obtained from a location associated with Singapore.

Determining whether or not the previous synthetic data sets were productively deconvolved was relatively straightforward, as the synthetic data were constructed using known quantities of known compounds. However, real-world fuels are typically never interrogated for similarly precise compositional information, simply because this information is rarely required during the course of routine fuel handling and use. Although preliminary internal work was also performed on simple surrogate fuels, the compositions of which were known because they were created in-house, it quickly became apparent that the GC-MS data collected from such surrogates were insufficiently complex for the purposes of proper algorithm testing.

Two indirect methodologies were thus developed to assess algorithm effectiveness with real-world fuel samples:

1. It can be reasonably assumed that petrochemical jet fuels contain dodecane and petrochemical diesel and MGO fuels contain tetradecane. This assumption is consistent with both their known prevalence in these types of fuels and the fact that, at least in certain circumstances, dodecane can function as a surrogate jet fuel and tetradecane can function as a surrogate diesel fuel, and these compounds are commonly found in these types of fuels. It can thus be determined whether or not the deconvolution algorithm allows these compound identities to be assigned to a larger proportion of the overall TIC area in their respective data sets and/or causes these assignments to be more accurate. In the latter case, accuracy can be quantified by applying MF thresholds and subsequently rejecting results if they are accompanied by insufficiently high MF values.

2. The NIST/EPA/NIH Mass Spectral Library contains several mass spectra correlated to bioactive molecules that are most likely not present in petrochemical fuels to any significant extent. However, during the course of previous in-house work, these bioactive molecules were identified as being present in fuel samples to a non-trivial degree. It is speculated that these dubious identifications were made because GC-MS data convolution results in mangled mass spectra that happen to look something like the Mass Spectral Library's reference mass spectra for these bioactive molecules. Data deconvolution should thus reduce the number of such identifications.

It will again be clarified that the combined EWFA-MCR algorithm, based on repeated iterations of SVD, is capable of producing multiple mass spectrum-like loadings per iteration, each corresponding to a different LV. This, of course, allows for the deconvolution of multiple compounds from even closely overlapping data, which is part of the algorithm's core functionality. The results from multiple LVs will thus be added together for comprehensive reporting purposes.

Figure 7 shows that the EWFA-MCR algorithm allows more of the parent GC-MS data to be identified as representing a compound that is likely to be present in the underlying fuel. This is promising, as is the fact that this improvement is based on deconvolved results yielding sufficiently high MF values that the overall trend remains consistent when utilizing the various MF threshold values. Although the use of EWFA-MCR is not required to find dodecane in the jet fuel and tetradecane in the MGO fuels, the increased number of identifications would seem to indicate that convoluted data, perhaps existing at the edges of convoluted dodecane-heavy and tetradecane-heavy chromatographic peaks, are effectively deconvolved using the algorithm, allowing for a more accurate and thorough compositional assessment of these fuel samples.

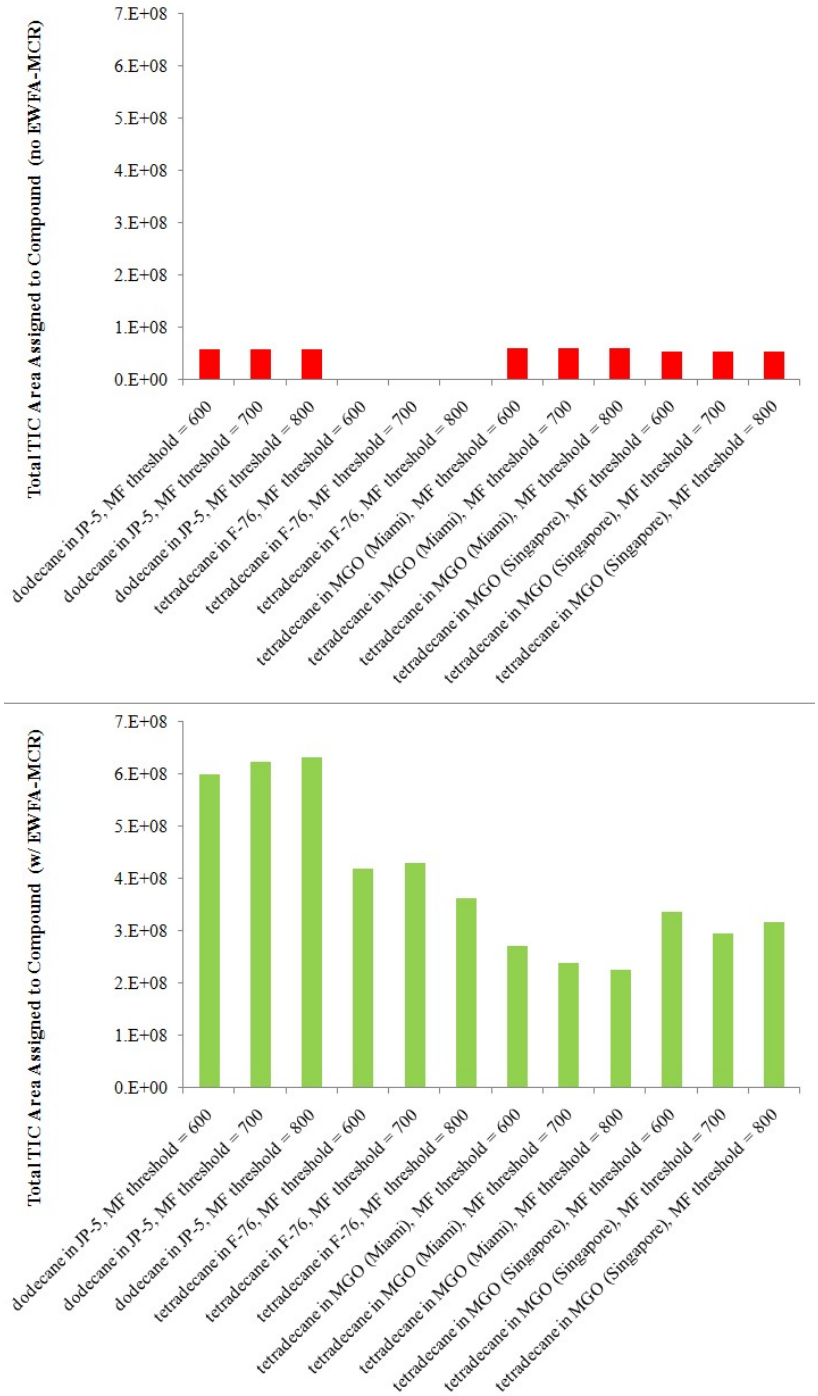

Figure 7. Summary of the number of dodecane identifications obtained from the GC-MS data collected from a petrochemical jet fuel, and the number of tetradecane identifications obtained from the data collected from a petrochemical diesel fuel and two petrochemical MGO fuels, both without (red) and with (green) the application of the EWFA-MCR deconvolution algorithm.

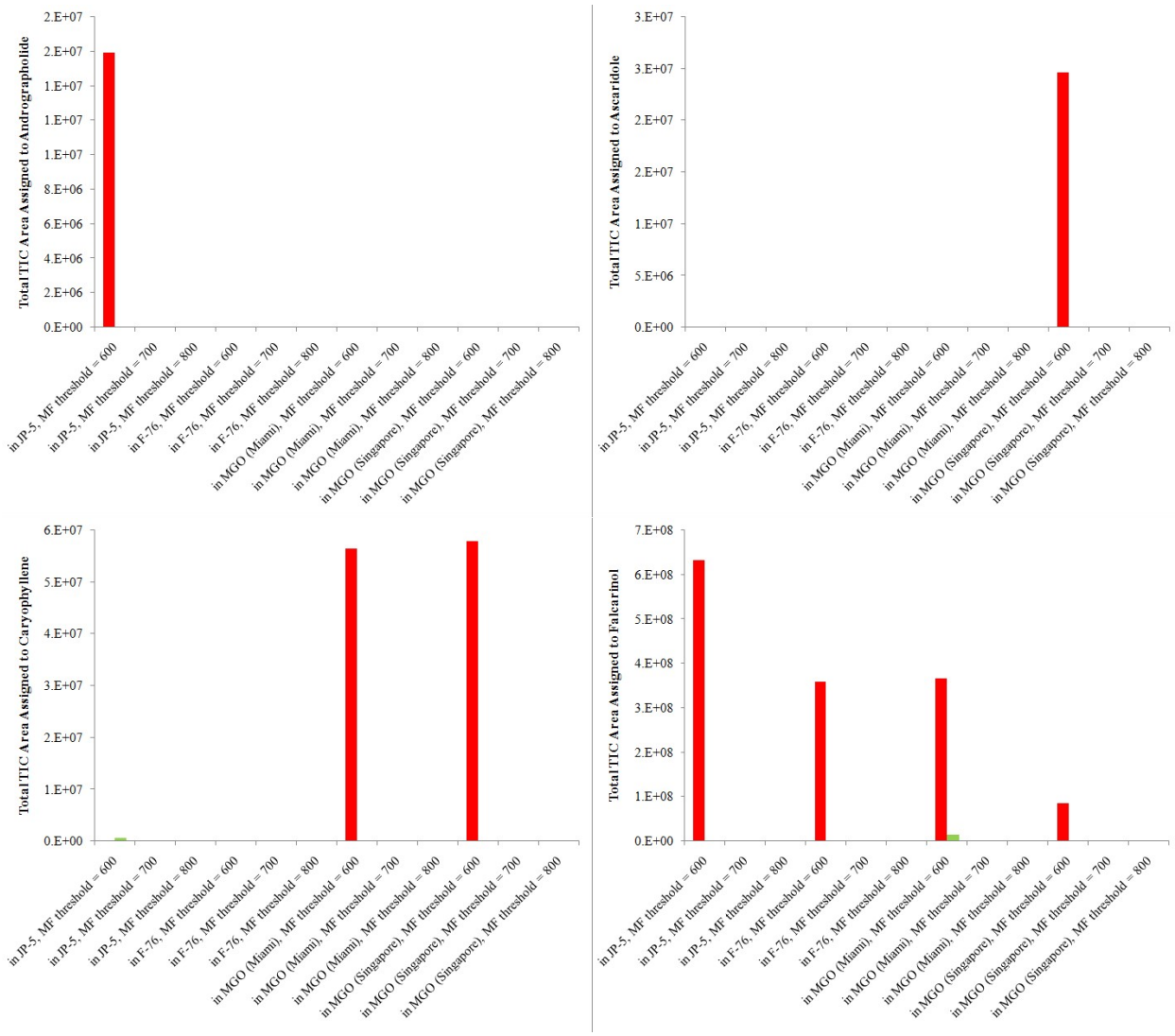

Figure 8. First half of a summary of the number of various bioactive molecule identifications obtained from the GC-MS data collected from a petrochemical jet fuel, a petrochemical diesel fuel, and two petrochemical MGO fuels, both without (red) and with (green) the application of the EWFA-MCR deconvolution algorithm. TIC area results obtained with the EWFA-MCR algorithm (green) have been multiplied by 10 simply to make them visible.

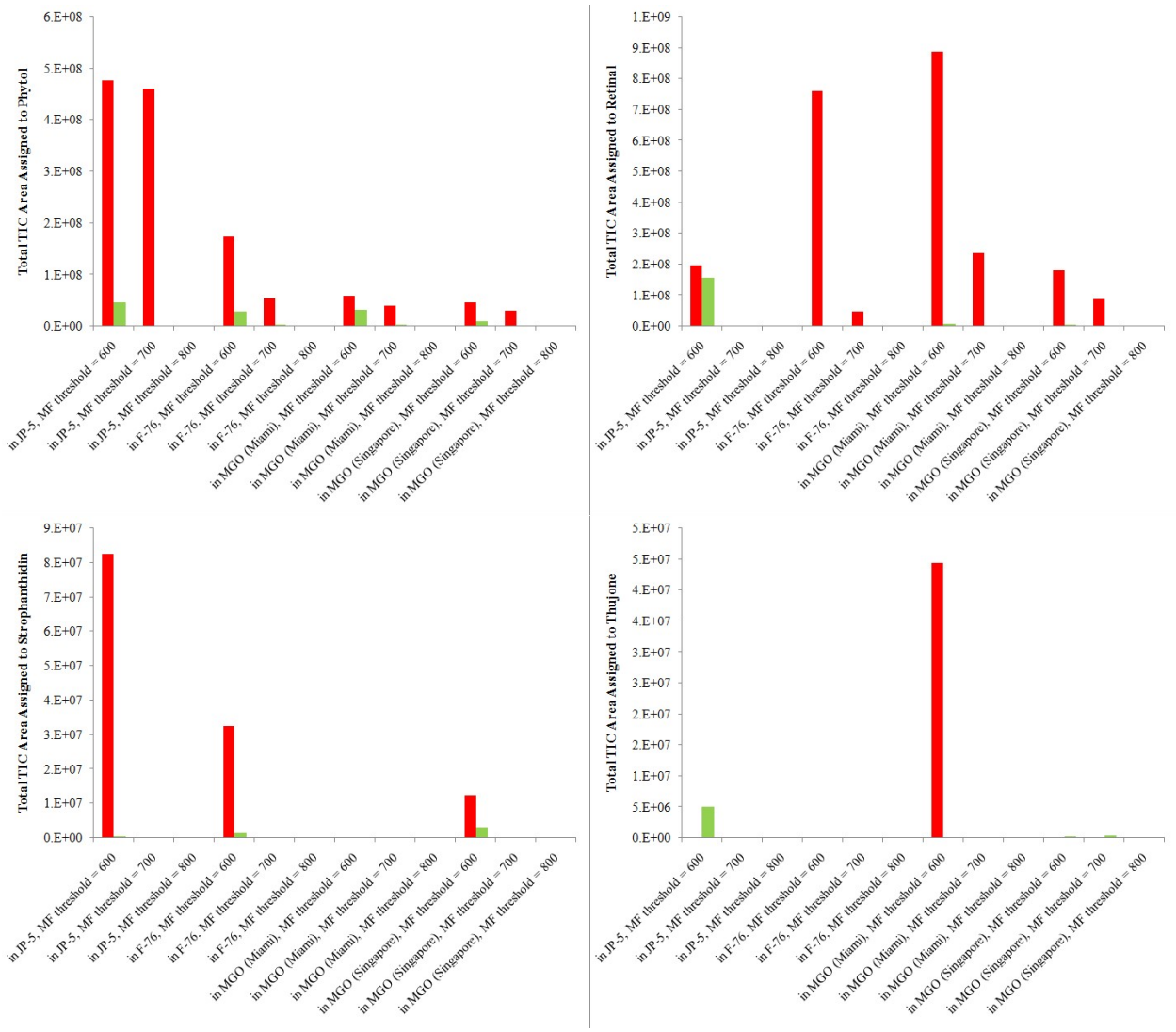

Figure 9. Second half of a summary of the number of various bioactive molecule identifications obtained from the GC-MS data collected from a petrochemical jet fuel, a petrochemical diesel fuel, and two petrochemical MGO fuels, both without (red) and with (green) the application of the EWFA-MCR deconvolution algorithm. TIC area results obtained with the EWFA-MCR algorithm (green) have been multiplied by 10 simply to make them visible.

Figures 8 and 9, meanwhile, show similar results for the TIC areas that can be assigned to eight bioactive molecules (andrographolide, ascaridole, caryophyllene, and falcarinol in Figure 8; phytol, retinal, strophanthidin, and thujone in Figure 9). After the use of EWFA-MCR, it would be expected that these bioactive molecules would cease to be identified in fuels, and, to a large extent, this expectation is met, regardless of the choice of MF threshold. In fact, due to the scales imposed by the control TIC area results, the TIC area results obtained with EWFA-MCR usage are multiplied by 10 in these figures, simply to make them visible.

6.3 Additional Output Constraints and Data Truncation Operations

Given available evidence, the EWFA-MCR algorithm fundamentally seems to be effective at improving the accuracy of GC-MS fuel analysis. During initial FCAST algorithm implementation, however, it became apparent that this data deconvolution procedure still produced an unacceptably high number of superfluous, questionable compound identifications despite the implementation of a standard MF-based thresholding. This is problematic in the context of automated compositional profiling, especially in those cases in which the same compound might be identified at multiple retention times. While a certain amount of compound redundancy at adjacent retention times would be consistent with non-ideal chromatographic resolution and/or peak widths exceeding one variable along retention time axes, identical compound identifications found at more widely dispersed retention times are considered unacceptable in the context of automated software applications. Although a certain level of expertise in GC-MS data interpretation might allow some end-users to ignore spurious identifications without discarding compositionally significant information, such a level of expertise cannot realistically be relied upon in the context of widely distributed software applications.

In order to more thoroughly correct for these superfluous compound identifications, and in particular to correct for multiple identifications of the same compound at different retention times, the MF threshold was first raised to 750, which is higher than the threshold range of about 600-700 typically employed in-house when analyzing GC-MS data sets. This was done to account for the higher overall quality of the mass spectral data that becomes available after deconvolution. In addition to this relatively straightforward analysis change, three additional output constraints and two additional data truncation operations were built into the original algorithm.

Output constraint 1: significance-based thresholding. EWFA relies upon multiple executions of SVD, the core decomposition of which can be represented by the following equation:

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

In this equation, **R** is the original data set (in this case, the portion of the GC-MS data being

analyzed), $\mathbf{V}^T$ 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. This **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.

Output constraint 2: loading number evaluation limit. 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 were subjected to further evaluations in earlier versions of the EWFA-MCR deconvolution algorithm. However, not only was this time consuming, but it is not likely that loadings associated with lower significances, as described above, will yield useful chemical information. While earlier versions of the deconvolution algorithm simply allowed less significant loadings to be deselected by means of MF values, practical implementation work made it apparent that this approach was insufficient. Thus, an output constraint was implemented 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.

Output constraint 3: area-based thresholding. The FCAST software prevents compound identifications that do not represent a certain percentage of the data's total ion chromatograph (TIC) area. This threshold is typically set at 0.001%, but deconvoluted results were initially allowed to possess areas less than this in order to more thoroughly assess trace components. However, it became apparent that allowing for any percent area concurrently allowed for nonsensical compound identifications, particularly in the form of incongruous heteroatomic compounds, to slip into the final analysis results, so the 0.001% threshold was re-applied to near-final compositional profile results as well. It should be noted, of course, that these percentages must be recalculated after this thresholding itself is applied to allow for the percentages to equal to 100% once again.

Data truncation 1: adjacent redundancies. Identical compound identifications that remain after these output constraints are applied are then initially added together when they appear at retention times that possess no intervening compound identifications. This addition occurs by first selecting the individual compound identification that contributed the most to the overall TIC area of the GC-MS data, which is then considered the primary identification at a primary retention time. All relevant TIC area contributions at this primary retention time and nearby retention times are then

summed into a single value and associated with the primary identification, with non-primary compound identifications deleted afterwards. This maintains the quantitative results obtained from multiple compound identifications while consolidating them in a single identification. This operation is iterated ten times across the list of identified compounds to safeguard against new adjacencies that might arise due to previous iterations of the operation.

Data truncation 2: non-adjacent redundancies. Identical compound identifications remaining after the first data truncation operation are then simply truncated by selecting a primary compound identification, in the same manner as described previously, and deleting all other non-primary identifications. This differs from the first operation because non-primary area contributions are not maintained as contributions to overall compound identifications. Adding area contributions together is defensible in the case of the adjacent compound identifications described previously, but it is much more likely that widely disparate compound identifications along a given retention time axis are due to lingering compound misidentifications and should thus be eliminated accordingly. While this second operation is particularly heavy-handed in its approach to obtaining the desired results, its positioning after the first data truncation operation helps to ensure that a minimal amount of meaningful chemical information is subject to loss.

Once these new constraints and data truncation operations were put into place, FCAST-type compositional profiles were obtained from the same GC-MS data acquired from the JP-5 jet fuel, F-76 diesel fuel, and two MGOs, associated with Miami and Singapore, utilized previously. Compositional profiles were used as the means of algorithm evaluation because it is within these profiles that the presence of redundant compound identifications first became apparent during FCAST implementation work. To simplify results, both the control profiles and post-deconvolution profiles were not allowed to have individual compounds appear in more than one hydrocarbon category within a profile, an option that can already be adjusted in the FCAST software. For the purposes of comparison, control (i.e. non-deconvolved) results and results obtained by using an earlier version of the EWFA-MCR algorithm that does not possess the new constraints and truncation operations were also collected. Tables 2 through 5 present the most immediately relevant results that can be derived from each fuel's compositional profiles: the numbers of unique compounds identified overall for saturates, olefins, aromatics, and heteroatomics, as well as each of twenty-five compound sub-categories, and the percentages of the total TIC areas that can be ascribed to these compound categories.

Table 2. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the JP-5 jet fuel.

	Control		**Old EWFA-MCR**		**Updated EWFA-MCR**	
	number	*area%*	*number*	*area%*	*number*	*area%*
Total Saturates	73	69.78%	868	56.64%	144	64.00%
Total Olefins	5	3.33%	39	1.18%	18	2.50%
Total Aromatics	29	18.05%	519	34.08%	95	27.53%
Total Heteroatomics	25	8.84%	853	8.11%	71	5.97%
Normal Alkanes	15	44.12%	336	35.22%	13	32.70%
Iso Alkanes	26	18.26%	325	16.50%	69	22.46%
Monocyclo Alkanes	2	1.16%	8	0.05%	2	0.22%
Alkyl Monocyclo Alkanes	29	5.85%	185	4.36%	54	8.02%
Dicyclo Alkanes	1	0.38%	10	0.29%	4	0.23%
Alkyl Dicyclo Alkanes	0	0.00%	4	0.21%	2	0.37%
Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Alkyl Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Acyclic Alkenes	5	3.33%	27	1.00%	12	2.36%
Cyclo Alkenes	0	0.00%	12	0.18%	6	0.14%
Alkyl Benzenes	20	12.39%	311	13.48%	58	13.39%
Indans and Tetralins	0	0.00%	51	7.69%	21	9.15%
Indenes	0	0.00%	0	0.00%	0	0.00%
Naphthalene	0	0.00%	2	0.86%	1	1.17%
Branched Naphthalenes	9	5.66%	132	10.56%	15	3.83%
Acenaphthenes	0	0.00%	0	0.00%	0	0.00%
Acenaphthylenes	0	0.00%	0	0.00%	0	0.00%
Tricycloaromatics	0	0.00%	23	1.48%	0	0.00%
Methyl Esters	0	0.00%	10	0.02%	1	0.02%
Sulfur-Bound	0	0.00%	76	1.23%	3	0.12%
Nitrogen-Bound	0	0.00%	357	2.89%	15	0.12%
Oxygen-Bound	18	6.61%	336	3.41%	38	4.70%
Chlorine-Bound	2	0.03%	12	0.01%	4	0.03%
Other Halogen-Bound	0	0.00%	15	0.03%	3	0.13%
Other	5	2.21%	47	0.51%	7	0.85%
Total # Compounds	132		2279		328	

Table 3. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the F-76 diesel fuel.

	Control		**Old EWFA-MCR**		**Updated EWFA-MCR**	
	number	*area%*	*number*	*area%*	*number*	*area%*
Total Saturates	90	74.31%	859	57.09%	138	59.99%
Total Olefins	7	3.68%	19	0.19%	11	1.73%
Total Aromatics	33	19.16%	491	36.78%	119	34.50%
Total Heteroatomics	20	2.85%	910	5.93%	63	3.77%
Normal Alkanes	21	58.70%	273	39.11%	20	36.83%
Iso Alkanes	39	12.93%	379	14.99%	68	17.97%
Monocyclo Alkanes	2	0.35%	16	0.00%	3	0.20%
Alkyl Monocyclo Alkanes	27	2.22%	182	2.93%	44	4.95%
Dicyclo Alkanes	1	0.10%	6	0.06%	2	0.03%
Alkyl Dicyclo Alkanes	0	0.00%	3	0.01%	1	0.02%
Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Alkyl Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Acyclic Alkenes	6	1.91%	17	0.04%	8	0.32%
Cyclo Alkenes	1	1.77%	2	0.16%	3	1.41%
Alkyl Benzenes	26	9.32%	318	10.95%	66	10.67%
Indans and Tetralins	2	1.59%	61	6.81%	22	7.37%
Indenes	0	0.00%	0	0.00%	0	0.00%
Naphthalene	0	0.00%	2	0.04%	1	0.17%
Branched Naphthalenes	5	8.25%	86	16.12%	18	13.18%
Acenaphthenes	0	0.00%	1	0.00%	1	0.01%
Acenaphthylenes	0	0.00%	0	0.00%	0	0.00%
Tricycloaromatics	0	0.00%	23	2.86%	11	3.11%
Methyl Esters	0	0.00%	2	0.00%	0	0.00%
Sulfur-Bound	0	0.00%	87	2.04%	11	2.33%
Nitrogen-Bound	0	0.00%	434	2.67%	12	0.10%
Oxygen-Bound	14	2.40%	302	1.00%	31	0.98%
Chlorine-Bound	3	0.05%	9	0.01%	2	0.03%
Other Halogen-Bound	0	0.00%	20	0.15%	2	0.20%
Other	3	0.40%	56	0.06%	5	0.13%
Total # Compounds	150		2279		331	

Table 4. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the Miami MGO fuel.

	Control		Old EWFA-MCR		Updated EWFA-MCR	
	number	*area%*	*number*	*area%*	*number*	*area%*
Total Saturates	91	58.73%	1178	53.35%	165	51.88%
Total Olefins	8	3.52%	33	0.20%	11	0.99%
Total Aromatics	38	31.23%	548	42.52%	121	44.41%
Total Heteroatomics	21	6.52%	773	3.93%	52	2.72%
Normal Alkanes	19	43.89%	536	35.91%	21	27.51%
Iso Alkanes	31	11.63%	348	13.96%	74	17.26%
Monocyclo Alkanes	2	0.15%	13	0.00%	3	0.04%
Alkyl Monocyclo Alkanes	38	2.61%	245	2.22%	57	3.97%
Dicyclo Alkanes	1	0.46%	15	0.11%	3	0.30%
Alkyl Dicyclo Alkanes	0	0.00%	21	1.15%	7	2.80%
Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Alkyl Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Acyclic Alkenes	8	3.52%	24	0.04%	8	0.51%
Cyclo Alkenes	0	0.00%	9	0.16%	3	0.48%
Alkyl Benzenes	26	10.94%	360	14.27%	70	15.86%
Indans and Tetralins	12	20.29%	119	18.69%	33	19.68%
Indenes	0	0.00%	0	0.00%	0	0.00%
Naphthalene	0	0.00%	1	0.23%	1	0.51%
Branched Naphthalenes	0	0.00%	44	6.28%	12	5.51%
Acenaphthenes	0	0.00%	0	0.00%	0	0.00%
Acenaphthylenes	0	0.00%	0	0.00%	0	0.00%
Tricycloaromatics	0	0.00%	24	3.05%	5	2.86%
Methyl Esters	0	0.00%	1	0.00%	0	0.00%
Sulfur-Bound	0	0.00%	63	1.04%	1	0.01%
Nitrogen-Bound	0	0.00%	315	1.15%	7	0.17%
Oxygen-Bound	11	3.32%	272	1.09%	32	1.33%
Chlorine-Bound	2	0.03%	24	0.01%	3	0.03%
Other Halogen-Bound	0	0.00%	24	0.13%	2	0.46%
Other	8	3.17%	74	0.51%	7	0.71%
Total # Compounds	158		2532		349	

Table 5. Profiles obtained before and after applying the previous and updated EWFA-MCR algorithms to the FCAST-type compositional profiling of the Singapore MGO fuel.

	Control		Old EWFA-MCR		Updated EWFA-MCR	
	number	*area%*	*number*	*area%*	*number*	*area%*
Total Saturates	109	82.07%	1305	68.26%	213	67.39%
Total Olefins	7	2.20%	32	0.30%	13	1.45%
Total Aromatics	33	12.51%	533	23.44%	122	24.55%
Total Heteroatomics	24	3.23%	662	8.00%	69	6.61%
Normal Alkanes	30	67.53%	561	45.75%	21	38.37%
Iso Alkanes	37	11.37%	447	20.06%	114	24.62%
Monocyclo Alkanes	3	0.25%	12	0.02%	5	0.11%
Alkyl Monocyclo Alkanes	38	2.82%	271	2.25%	66	4.08%
Dicyclo Alkanes	1	0.09%	11	0.11%	5	0.07%
Alkyl Dicyclo Alkanes	0	0.00%	3	0.06%	2	0.14%
Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Alkyl Tricyclo Alkanes	0	0.00%	0	0.00%	0	0.00%
Acyclic Alkenes	7	2.20%	26	0.07%	7	0.31%
Cyclo Alkenes	0	0.00%	6	0.23%	6	1.14%
Alkyl Benzenes	27	6.85%	368	9.13%	74	10.87%
Indans and Tetralins	1	0.68%	75	4.91%	24	5.10%
Indenes	0	0.00%	1	0.00%	0	0.00%
Naphthalene	0	0.00%	2	0.17%	1	0.24%
Branched Naphthalenes	5	4.97%	72	8.25%	14	6.92%
Acenaphthenes	0	0.00%	1	0.03%	0	0.00%
Acenaphthylenes	0	0.00%	0	0.00%	0	0.00%
Tricycloaromatics	0	0.00%	14	0.93%	9	1.42%
Methyl Esters	0	0.00%	3	0.00%	0	0.00%
Sulfur-Bound	0	0.00%	54	3.81%	14	4.21%
Nitrogen-Bound	0	0.00%	229	2.07%	3	0.01%
Oxygen-Bound	15	2.47%	268	1.54%	35	1.48%
Chlorine-Bound	3	0.04%	15	0.01%	6	0.04%
Other Halogen-Bound	0	0.00%	11	0.32%	4	0.47%
Other	6	0.72%	82	0.26%	7	0.41%
Total # Compounds	173		2532		417	

Table 2 shows the results obtained from the JP-5 jet fuel. The older data deconvolution algorithm did indeed allow for many more compound identifications across almost all compound categories and sub-categories, but note that the number of heteroatomics identified increased far more in proportion to the saturate, olefin, and aromatic increases. Given the sheer volume of these heteroatomic identifications, it is likely that a non-trivial number of these are analysis artifacts, as opposed to true representations of the jet fuel's composition. The updated algorithm corrects for this over-identification trend, and assigns less of the overall TIC area to the presence of heteroatomics as well, both developments that are consistent with the composition of a standard petrochemical fuel. Though the total TIC areas that are ascribed to the presence of saturates and aromatics still seem improperly balanced after applying the updated EWFA-MCR algorithm, especially for a petrochemical jet fuel, clear compound identification advantages are still apparent when using the updated EWFA-MCR algorithm.

Table 3 shows the results obtained from the F-76 fuel. Again, a seemingly overaggressive identification of heteroatomic compounds, in the case of the older deconvolution algorithm, becomes a more reasonable overall improvement after the algorithmic updates are applied. Also note here, and is the case with the other three tested fuels, that the total TIC areas identified with olefin content decrease relative to the control (non-deconvolved) analysis results. This remains consistent with what would be expected from petrochemical fuel compositions.

Tables 4 and 5 show the results obtained from the MGO fuels, in which similar heteroatomic identification improvements to those shown previously can again be seen. These MGO results also most clearly indicate the reductions in the numbers of saturates that are identified after the updates are applied to the deconvolution algorithm. Note in particular the normal alkane results, which are far more consistent with the numbers of normal alkanes that would be expected from a petrochemical fuel after the updates are applied than beforehand, most likely due to the data truncation operations described previously.

Overall, Tables 2 through 5 show that the deconvolution algorithm, after the employed updates, still allows for the identification of substantially more unique compounds than without its use, while also more accurately representing expected compositional profiles.

6.4 Comprehensive Evaluation of Updated Algorithm Constraints

In the profiles shown in the previous section, the total TIC areas assigned to saturates and aromatics remained, at least seemingly, out of balance (i.e. too many aromatics and too few saturates). This is particularly apparent in the case of the jet fuel sample, for which ASTM D1319 saturates data (78.6%) and ASTM D6379 aromatics data (20%) are actually available, though an 80% saturates/20% aromatics ratio would not be out of the ordinary for diesel fuels as well. It was thus

decided that the specific parameters of the three new output constraints employed in the updated EWFA-MCR algorithm should be more thoroughly evaluated to determine if further improvements in compositional results can be obtained. As before, the MF threshold was kept at a constant value of 750 throughout the present evaluations.

Early in these evaluations, it was decided that manipulating the area-based thresholding would most likely not be of much use in the present work. Initial results for all four previously evaluated fuels, which show the consequences of lowering this threshold from 0.001% (consistent with the FCAST's default settings) to 0.0001% (i.e. one-tenth the previous value) are shown in Table 6. While lowering this threshold does, as would be expected, result in the identification of more compounds, there is very little in the way of corresponding peak area changes. This would seem to indicate that the new compound identities, deconvolved from the same parent chromatographic peaks as the older compound identities, still tend to represent the same broad categorical chemical information. Significant changes to something like a saturates/aromatics ratio, then, would likely be more productively pursued by varying the two other output constraints.

The significance-based thresholding was considered very straightforward with respect to how to augment it. In the previous section, the thresholding value was set to an almost trivial value of $1x10^{-15}$, simply to avoid the further propagation of the most insignificant results produced via SVD. It was thus hypothesized that increasing the magnitude of this number might very well result in an increased discriminatory capability, though a smaller value ($1x10^{-17}$) was evaluated as well for the sake of completeness.

The numbers of loadings to further evaluate after each individual SVD operation was restricted, in the previous section, to the three most significant loadings. While this appeared to produce a good balance of analysis speed and accuracy during preliminary internal work, the present work will focus on comparing this threshold value of three to values of two and four as well, to determine if the consideration of more or fewer loadings will significantly impact profiled compositional results.

Table 6. EWFA-MCR-augmented, FCAST-type compositional profiling results obtained from the four fuels described previously, before and after changing the area-based thresholding constraint.

	0.001%		0.0001%	
	compound #	area %	compound #	area %
JP-5 jet				
Total Saturates	144	64.00%	163	64.00%
Total Olefins	18	2.50%	20	2.50%
Total Aromatics	95	27.53%	96	27.52%
Total Heteroatomics	71	5.97%	88	5.98%
F-76 diesel				
Total Saturates	138	59.99%	157	59.99%
Total Olefins	11	1.73%	12	1.73%
Total Aromatics	119	34.50%	121	34.49%
Total Heteroatomics	63	3.77%	90	3.78%
MGO (Miami)				
Total Saturates	165	51.88%	189	51.87%
Total Olefins	11	0.99%	15	0.99%
Total Aromatics	121	44.41%	123	44.41%
Total Heteroatomics	52	2.72%	76	2.73%
MGO (Singapore)				
Total Saturates	213	67.39%	240	67.38%
Total Olefins	13	1.45%	19	1.45%
Total Aromatics	122	24.55%	126	24.54%
Total Heteroatomics	69	6.61%	85	6.62%

Table 7. EWFA-MCR-augmented, FCAST-type compositional profiling results obtained from the JP-5 jet fuel, before and after changing the significance-based and loading-based thresholding constraints.

Fuel: C419 (JP-5 jet)	**Max. Eval. Loadings=2**		**Max. Eval. Loadings=3**		**Max. Eval. Loadings=4**	
sig. threshold: $1x10^{-17}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.53%	148	65.05%	155	64.32%
Total Olefins	17	2.19%	19	2.06%	19	2.02%
Total Aromatics	91	27.46%	95	26.98%	96	27.65%
Total Heteroatomics	56	5.82%	72	5.91%	81	6.01%
sig. threshold: $1x10^{-15}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.53%	148	65.05%	155	64.32%
Total Olefins	17	2.19%	19	2.06%	19	2.02%
Total Aromatics	91	27.46%	95	26.98%	96	27.65%
Total Heteroatomics	56	5.82%	72	5.91%	81	6.01%
sig. threshold: $1x10^{-11}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.53%	148	65.05%	155	64.32%
Total Olefins	17	2.19%	19	2.06%	19	2.02%
Total Aromatics	91	27.46%	95	26.98%	96	27.65%
Total Heteroatomics	56	5.82%	72	5.91%	81	6.01%
sig. threshold: $1x10^{-7}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.53%	148	65.05%	155	64.32%
Total Olefins	17	2.19%	19	2.06%	19	2.02%
Total Aromatics	91	27.46%	95	26.98%	96	27.65%
Total Heteroatomics	56	5.82%	71	5.91%	81	6.01%
sig. threshold: $1x10^{3}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.53%	148	65.05%	155	64.32%
Total Olefins	17	2.19%	19	2.06%	19	2.02%
Total Aromatics	91	27.46%	95	26.98%	96	27.66%
Total Heteroatomics	56	5.82%	71	5.91%	81	6.01%
sig. threshold: $1x10^{4}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.54%	148	65.08%	155	64.36%
Total Olefins	17	2.19%	19	2.06%	19	2.02%
Total Aromatics	91	27.47%	95	26.99%	96	27.67%
Total Heteroatomics	55	5.80%	69	5.86%	77	5.95%
sig. threshold: $1x10^{5}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	141	64.60%	151	65.17%	157	64.46%
Total Olefins	17	2.20%	19	2.06%	19	2.02%
Total Aromatics	91	27.49%	95	27.02%	96	27.70%
Total Heteroatomics	40	5.72%	52	5.74%	56	5.82%
sig. threshold: $1x10^{6}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	135	64.52%	140	65.19%	144	64.46%
Total Olefins	16	2.36%	16	2.32%	16	2.30%
Total Aromatics	87	27.40%	91	26.81%	93	27.45%
Total Heteroatomics	34	5.73%	40	5.68%	42	5.79%
sig. threshold: $1x10^{7}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	80	68.59%	80	68.24%	80	67.96%
Total Olefins	7	1.88%	7	1.95%	7	2.00%
Total Aromatics	52	21.71%	52	21.94%	52	22.11%
Total Heteroatomics	20	7.81%	21	7.87%	21	7.94%
sig. threshold: $1x10^{8}$	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	31	73.25%	31	73.25%	31	73.25%
Total Olefins	2	1.47%	2	1.47%	2	1.47%
Total Aromatics	19	18.69%	19	18.69%	19	18.69%
Total Heteroatomics	9	6.59%	9	6.59%	9	6.59%

Table 7 shows the consequences of varying both the significance-based thresholding and the numbers of evaluated loadings on the composition summaries obtained from the JP-5 jet fuel. This JP-5 jet fuel was solely focused upon in this table because, as indicated previously, a specific saturates/aromatics ratio can be targeted due to the existence of relevant specification data for the sample. It should first be noted that the results found in Table 7 do not present any clear evidence that either increasing or decreasing the maximum number of loadings to be evaluated per SVD operation from three would provide a meaningful analytical benefit. Otherwise, most immediately apparent in this table is the lack of anything of impact to report when considering significance-based thresholds from $1x10^{-17}$ to $1x10^{3}$. While the previous selection of the value of $1x10^{-15}$ was somewhat arbitrary, this value apparently has several orders' of magnitude worth of leeway before it performs a function more wide-ranging than the basic gatekeeping ability for which it was initially designed. Conversely, however, even superficial improvements to the calculated saturates/aromatics ratio do not make themselves apparent through significance-based thresholding until the rather large significance thresholds of $1x10^{7}$ are $1x10^{8}$ are employed. When these thresholds are used, however, the actual numbers of deconvolved compounds identified become substantially diminished, thus defeating the purpose of employing peak deconvolution in the first place.

Interestingly, however, it would also at least initially appear that increasing the significance threshold to a more modest value of $1x10^{5}$ might provide an analytical benefit by decreasing the total number of heteroatomic compounds reported to the end-user. As indicated previously, petrochemical fuels typically to not possess large amounts of heteroatomic compounds, which implies that larger detected amounts of these compounds may be the result of spurious identifications. Follow-up work was thus performed to determine the reliability of this potential benefit across multiple fuel types. Table 8 shows the results of adjusting the significance threshold from $1x10^{-15}$ to $1x10^{5}$ when deconvolving the GC-MS data from all four petrochemical fuels described previously. For the purposes of obtaining a thorough evaluation, results were collected using not only the current default of a maximum number of three loadings to evaluate per SVD operation, but also under alternative maximum values of two and four loadings per SVD operation.

Table 8. EWFA-MCR-augmented, FCAST-type compositional profiling results obtained from four fuels, before and after changing the significance-based and loading-based thresholding constraints.

	JP-5 jet				*F-76 diesel*			
sing. val. threshold:	$1x10^{-15}$		$1x10^{5}$		$1x10^{-15}$		$1x10^{5}$	
Max. Eval. Loadings=2	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	138	64.53%	141	64.60%	143	61.11%	144	61.21%
Total Olefins	17	2.19%	17	2.20%	10	1.86%	9	1.85%
Total Aromatics	91	27.46%	91	27.49%	118	33.14%	118	33.19%
Total Heteroatomics	56	5.82%	40	5.72%	58	3.90%	44	3.75%
Max. Eval. Loadings=3	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	148	65.05%	151	65.17%	144	60.35%	143	60.52%
Total Olefins	19	2.06%	19	2.06%	12	1.84%	10	1.83%
Total Aromatics	95	26.98%	95	27.02%	122	33.84%	122	33.84%
Total Heteroatomics	72	5.91%	52	5.74%	68	3.96%	54	3.81%
Max. Eval. Loadings=4	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	155	64.32%	157	64.46%	143	60.03%	145	60.21%
Total Olefins	19	2.02%	19	2.02%	13	1.85%	11	1.84%
Total Aromatics	96	27.65%	96	27.70%	125	33.89%	125	33.89%
Total Heteroatomics	81	6.01%	56	5.82%	75	4.23%	59	4.07%
	MGO (Miami)				*MGO (Singapore)*			
sing. val. threshold:	$1x10^{-15}$		$1x10^{5}$		$1x10^{-15}$		$1x10^{5}$	
Max. Eval. Loadings=2	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	172	51.81%	171	51.86%	213	68.84%	211	68.89%
Total Olefins	9	0.82%	9	0.82%	10	1.28%	10	1.28%
Total Aromatics	117	44.84%	117	44.80%	118	22.94%	118	22.90%
Total Heteroatomics	43	2.53%	40	2.53%	63	6.94%	60	6.93%
Max. Eval. Loadings=3	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	179	52.22%	175	52.25%	218	67.37%	217	67.40%
Total Olefins	11	1.01%	11	1.01%	11	1.25%	12	1.26%
Total Aromatics	123	44.27%	122	44.24%	126	24.08%	126	24.07%
Total Heteroatomics	51	2.51%	49	2.51%	72	7.29%	68	7.27%
Max. Eval. Loadings=4	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Total Saturates	180	51.88%	177	51.89%	220	67.27%	219	67.29%
Total Olefins	11	1.02%	11	1.02%	12	1.35%	13	1.36%
Total Aromatics	129	44.68%	128	44.67%	128	24.02%	128	24.01%
Total Heteroatomics	57	2.41%	54	2.42%	81	7.36%	77	7.34%

As was indicated previously, the JP-5 fuel shows a marked decrease in the number of heteroatomics detected when raising the significance threshold (72 to 52 when considering a maximum loading threshold of 3), and a somewhat less substantial but still significant decrease in this same value could also be found when considering the F-76 diesel fuel as well (68 to 54 when considering a maximum loading threshold of 3). However, the MGO samples display much less in the way of heteroatomic detection decreases when the significance threshold is increased. While this could possibly indicate that the parameter change leading to an analysis improvement for the former two fuels is not useful but still benign for the latter two fuels, changing a universally-applied parameter in light of non-universal analysis improvements remains imprudent without the collection of additional evidence.

Given the lack of a clearly identifiable trend in Table 8, the heteroatomics data were investigated in more detail, as can be seen in Table 9. Notably, in the JP-5 and F-76 fuels, about half of the overall significance-based decreases in detected heteroatomics can be attributed to decreases in the numbers of nitrogen-containing compounds that were detected. This loss of nitrogen-containing compound identification capabilities might become problematic in realistic fuel investigations, as at least some nitrogen-containing compounds are known to be responsible for fuel quality and performance issues, even at the trace levels indicated by the area results found in Table 9. Ultimately, no clear evidence exists that increasing the singular value threshold results in analysis improvements suitable for automated applications. It is thus not presently considered necessary to permanently raise the singular value threshold beyond the marginal value of $1x10^{-15}$ deemed appropriate during previous development work.

It must, of course, also be noted that the results shown in Tables 8 and 9 do not clearly indicate a need to change the default maximum number of evaluated loadings from a value of 3. This value still appears to provide a significantly more thorough investigative capability than a value of 2, and raising the value to 4, while providing an added compound identification capability, is not deemed necessary in the context of commensurately increased calculation times. This is roughly consistent with observation made previously.

Table 9. EWFA-MCR-augmented, FCAST-type heteroatomic profiling results obtained from four fuels, before and after changing the significance-based and loading-based thresholding constraints.

	JP-5 jet				*F-76 diesel*			
sing. val. threshold:	$1x10^{-15}$		$1x10^{5}$		$1x10^{-15}$		$1x10^{5}$	
Max. Eval. Loadings=2	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Methyl Esters	1	0.00%	0	0.00%	0	0.00%	0	0.00%
Sulfur-Bound	2	0.02%	1	0.01%	7	2.29%	6	2.29%
Nitrogen-Bound	10	0.05%	2	0.01%	10	0.03%	2	0.00%
Oxygen-Bound	28	4.42%	25	4.37%	31	1.16%	28	1.05%
Chlorine-Bound	4	0.03%	2	0.02%	4	0.03%	3	0.02%
Other Halogen-Bound	3	0.08%	3	0.08%	2	0.26%	2	0.26%
Other	8	1.22%	7	1.22%	4	0.12%	3	0.12%
Max. Eval. Loadings=3	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Methyl Esters	1	0.02%	0	0.00%	0	0.00%	0	0.00%
Sulfur-Bound	3	0.04%	2	0.04%	10	2.35%	9	2.34%
Nitrogen-Bound	11	0.08%	3	0.02%	11	0.05%	2	0.00%
Oxygen-Bound	42	4.45%	35	4.36%	36	1.15%	33	1.06%
Chlorine-Bound	4	0.03%	2	0.02%	4	0.03%	3	0.02%
Other Halogen-Bound	3	0.10%	3	0.10%	2	0.26%	3	0.27%
Other	8	1.20%	7	1.20%	5	0.12%	4	0.12%
Max. Eval. Loadings=4	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Methyl Esters	1	0.02%	0	0.00%	0	0.00%	0	0.00%
Sulfur-Bound	3	0.05%	2	0.05%	10	2.37%	9	2.37%
Nitrogen-Bound	15	0.08%	3	0.02%	12	0.08%	4	0.03%
Oxygen-Bound	47	4.53%	39	4.43%	42	1.37%	36	1.27%
Chlorine-Bound	5	0.03%	2	0.02%	4	0.03%	3	0.02%
Other Halogen-Bound	2	0.10%	3	0.10%	2	0.26%	2	0.27%
Other	8	1.20%	7	1.20%	5	0.12%	5	0.12%

	MGO (Miami)				*MGO (Singapore)*			
sing. val. threshold:	$1x10^{-15}$		$1x10^{5}$		$1x10^{-15}$		$1x10^{5}$	
Max. Eval. Loadings=2	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Methyl Esters	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Sulfur-Bound	0	0.00%	0	0.00%	13	4.27%	13	4.27%
Nitrogen-Bound	4	0.06%	4	0.06%	4	0.01%	2	0.01%
Oxygen-Bound	27	1.42%	26	1.42%	30	1.65%	31	1.65%
Chlorine-Bound	4	0.03%	3	0.03%	5	0.04%	3	0.03%
Other Halogen-Bound	1	0.22%	1	0.22%	3	0.59%	3	0.59%
Other	7	0.79%	6	0.79%	8	0.39%	8	0.39%
Max. Eval. Loadings=3	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Methyl Esters	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Sulfur-Bound	0	0.00%	0	0.00%	14	4.49%	13	4.48%
Nitrogen-Bound	5	0.07%	4	0.07%	4	0.01%	2	0.01%
Oxygen-Bound	32	1.42%	33	1.42%	37	1.71%	38	1.71%
Chlorine-Bound	4	0.03%	3	0.03%	5	0.04%	3	0.02%
Other Halogen-Bound	2	0.21%	2	0.21%	4	0.65%	4	0.65%
Other	8	0.78%	7	0.78%	8	0.40%	8	0.40%
Max. Eval. Loadings=4	*number*	*area*	*number*	*area*	*number*	*area*	*number*	*area*
Methyl Esters	0	0.00%	0	0.00%	0	0.00%	0	0.00%
Sulfur-Bound	1	0.00%	0	0.00%	14	4.48%	13	4.47%
Nitrogen-Bound	6	0.08%	5	0.08%	3	0.01%	2	0.01%
Oxygen-Bound	36	1.38%	37	1.39%	46	1.77%	46	1.76%
Chlorine-Bound	4	0.03%	3	0.03%	5	0.04%	3	0.02%
Other Halogen-Bound	2	0.21%	2	0.21%	4	0.65%	4	0.65%
Other	8	0.72%	7	0.71%	9	0.42%	9	0.42%

6.5 Additional Algorithm Validations

Algorithm validation via oxygenate discovery. Previous work in our laboratory has indicated that stressing diesel fuels, including F-76 and ULSD fuel grades, in an oxygenated LPR for 16 hours as prescribed in ASTM D5304 results in an elevated presence of oxygenated products in the stressed samples. However, GC-MS analysis of ten stressed and unstressed fuel sample pairs, collected recently in our laboratory for each of four diesel fuels, were not initially found to possess significantly increased amounts of oxygenated compounds in the stressed samples. In order to elucidate the potential presence of these oxygenated compounds, and to determine whether or not their amounts are elevated after the application of the described stressing, the EWFA-MCR deconvolution algorithm was applied to the GC-MS data collected from ten stressed samples and ten unstressed samples of these four different diesel fuels. To focus subsequent results, this GC-MS data were initially truncated to only those portions of the data whose TIC traces were found, via the data analysis technique of ANOVA, to possess an f-statistic value greater than or equal to 70% of the maximum f-statistic value found for the total TIC. The results of this focused EWFA-MCR analysis can be found in Table 10, wherein it is shown that a greater number of oxygenates were found in the stressed fuels, across all ten replicates, for all four diesel fuels (both absolutely and proportionally). In addition, a greater total TIC peak area was assigned to oxygenates after stress, across all ten replicates, for three out of four of the diesel fuels (both absolutely and proportionally). These results illustrate the potential of the deconvolution algorithm to uncover obscured fuel components. It is noteworthy that the analysis suggests that the ULSD contained fewer oxygenates than the high sulfur F-76 fuel, both before and after D53204 testing. This is consistent with the fact that this particular ULSD fuel was found to be stable during D5304 testing and didn't develop significant levels of hydroperoxides.

Visual representations of peak deconvolution. The current iteration of the EWFA-MCR algorithm does not, as a matter of course, produce estimated visualizations of distinct, deconvolved component peaks from convoluted TIC peaks. This is at least in part because the number of individual co-eluting isomers that can be found within any given convoluted chromatographic TIC peak obtained from a fuel sample has the potential to be quite large. Nonetheless, these estimated visualizations can be produced by measuring the contributions of each deconvolved compound to the shape of a parent TIC peak, and then normalizing these contributions to the area of said parent peak. The contributions can be plotted versus retention time to give some idea as to what these deconvolved peaks would look like, as can be seen in Figures 10 and 11 for selected retention time ranges within the data collected from a JP-5 and Jet A fuel, respectively. This operation was performed to prove the concept behind a post-deconvolution visualization capacity suitable for reporting purposes, should as much be required.

Table 10: Total numbers of compounds and associated peak areas found across ten replicates, collected for each of four parent diesel fuels, before and after stressing as prescribed in ASTM D5304.

ULSD #1	*# oxygenates identified*	*total # compounds identified*	***% identifications are oxygenates***	*peak areas of oxygenates identified*	*total peak areas of compounds identified*	***% identified peak areas are oxygenates***
unstressed	66	260	**25.38%**	98349287	1259259706	**7.81%**
stressed	75	234	**32.05%**	67703729	968031444	**6.99%**
ULSD #1 after copper doping	*# oxygenates identified*	*total # compounds identified*	***% identifications are oxygenates***	*peak areas of oxygenates identified*	*total peak areas of compounds identified*	***% identified peak areas are oxygenates***
unstressed	51	405	**12.59%**	2213266	368706061	**0.60%**
stressed	126	530	**23.77%**	4495141	308438915	**1.46%**
F-76 #1	*# oxygenates identified*	*total # compounds identified*	***% identifications are oxygenates***	*peak areas of oxygenates identified*	*total peak areas of compounds identified*	***% identified peak areas are oxygenates***
unstressed	15	94	**15.96%**	410494	13014826	**3.15%**
stressed	50	150	**33.33%**	1055612	8119033	**13.00%**
F-76 #1 / F-76 #2 blend	*# oxygenates identified*	*total # compounds identified*	***% identifications are oxygenates***	*peak areas of oxygenates identified*	*total peak areas of compounds identified*	***% identified peak areas are oxygenates***
unstressed	186	774	**24.03%**	290660138	1782688865	**16.30%**
stressed	242	785	**30.83%**	287270246	1641791030	**17.50%**

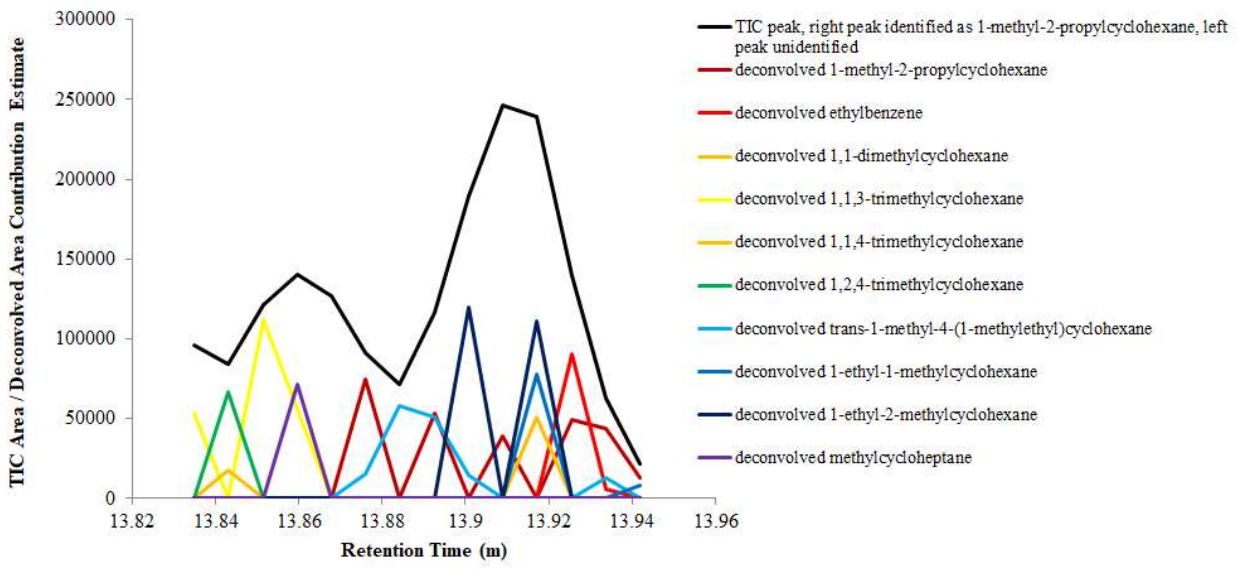

Figure 10. Visualization of estimated individual compound contributions, found via peak deconvolution, to a selected retention time range in a JP-5 GC-MS data set, and the TIC peaks therein.

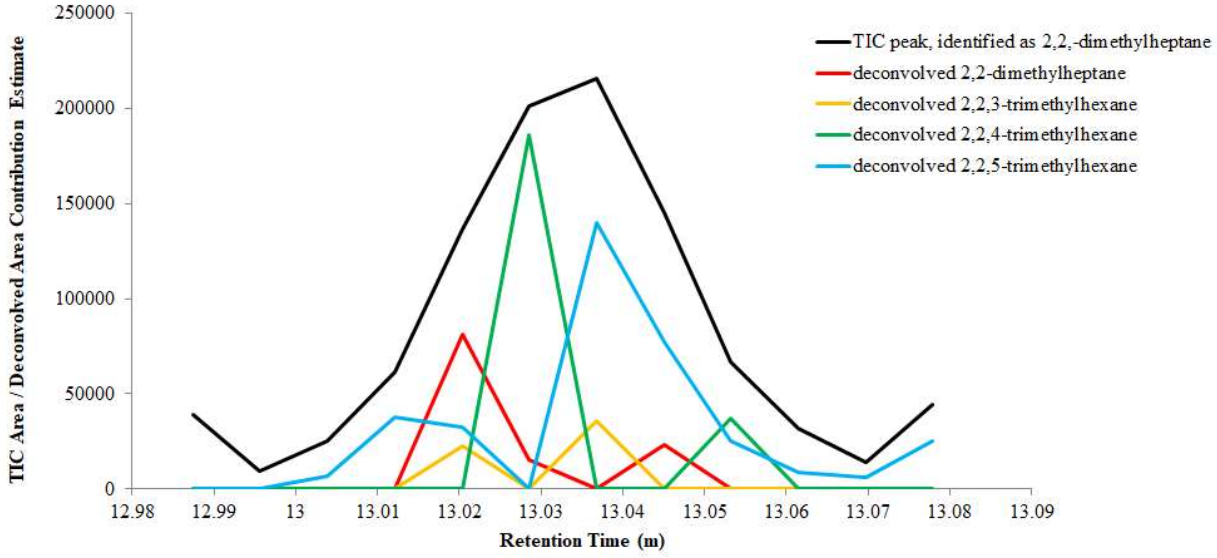

Figure 11. Visualization of estimated individual compound contributions, found via peak deconvolution, to a selected retention time range in a Jet A GC-MS data set, and the TIC peak therein.

7.0 Summary

A wide range of automated GC-MS data deconvolution parameters were addressed during the course of developing the combined EWFA-MCR algorithm presented herein. The broad scope of this research effort was necessary due to the broad scope of the analytical challenge at hand. Developing practical methodologies by which to reliably interrogate fuel composition by means of GC-MS in an automated fashion requires an accommodation of realistic fuel complexity during algorithm design work and similarly varied testing and evaluation challenges. Fuel analysis requirements that are more limited in scope might very well be ably addressed using more straightforward analysis methodologies, whether or not they would also be well-served by incorporating autonomous data deconvolution. Care should always be taken to consider the scope of any given analysis challenge, not only to develop the necessary solutions more efficiently, but also to minimize the unintended consequences that could potentially arise from the interactions of multiple individual solutions.

8.0 Conclusions

It has been shown that EWFA, MCR, and MF values resulting from NIST database searches can be intelligently leveraged into an automated GC-MS peak deconvolution strategy suitable for implementation into the FCAST software, and there is no reason to suspect that other software packages requiring similarly autonomous GC-MS data deconvolution capabilities would not benefit from the present work as well. Because fuel stability and performance problems are often due to the presence of trace levels of contaminants or other minor changes in composition, detailed compositional analyses of suspect fuels are critical for effective investigations, and said analyses would be enhanced through the implementation of the combined EWFA-MCR data deconvolution algorithm developed herein, as it allows for the more accurate and precise determinations of fuel compositions.

9.0 Recommendations

The updated data deconvolution algorithms presented herein have already been implemented into a standalone software architecture, with further implementation into the FCAST itself being an immediate in-house goal. Some of the challenges associated with this practical implementation work were addressed earlier in this report during discussions of algorithm optimization.

One challenge that remains, however, is that, at present, performing the EWFA-MCR data deconvolution across the whole of an average fuel's GC-MS data set results in calculation times that are best measured in hours or perhaps even days, as opposed to minutes or seconds. Therefore, an additional speed enhancing modification to the overall data deconvolution strategy is being

considered, one which would focus deconvolution operations only upon those TIC peaks that, based upon existing FCAST-derived information, are already suspected to consist of more than one co-eluted chemical compound. Work is currently being pursued in order to test, evaluate, and optimize such an enhancement. More testing with standard mixtures will also be conducted to more thoroughly assess the accuracy of the EFWA-MCR deconvolution algorithms, especially in the context of algorithm speed enhancements.

Otherwise, although the robustness of the developed data modeling strategy has been established through the use of multiple data sets and types of fuel samples, minor optimizations to the strategy's operational parameters, similar to the minor optimizations already described in this report, will be investigated and implemented as necessary.

This peak deconvolution strategy also has the potential to provide significant improvements in GCxGC chromatographic resolution for comprehensive fuel characterizations, and it is thus anticipated that future in-house GCxGC fuel modeling work will, in whole or in part, make use of the strategy.

10.0 Acknowledgements

The authors thank DLA Energy for their funding and support. Philip Chang was the technical monitor for this program.

11.0 Literature Cited

[1] Morris, RE; Hammond, MH; Cramer, JA; Myers, KM; Loegel, TN; Johnson, KJ; Leska, IA. *Proceedings of the 14th International Symposium on Stability, Handling, and Use of Liquid Fuels (IASH)*, 2015.
[2] Hammond, MH; Morris, RE; Cramer, JA; Loegel, TN; Johnson, KJ; Myers, KM. "Navy Fuel Composition and Screening Tool (FCAST) v2.8", NRL Memorandum Report NRL/MR/6180—16-9685, May 10, 2016.
[3] Dixon, SJ; Brereton, RG; Soini, HA; Novotny, MV; Penn, DJ. *Journal of Chemometrics* 2006; **20**:325-340.
[4] Maeder, M. *Analytical Chemistry* 1987; **59**:527-530.
[5] Parastar, H; Tauler, R. *Analytical Chemistry* 2014; **86**:286−297.