Autonomous Identification and Quantification of Chemical Species with VCAM for use Onboard the ISS

Benjamin Bornstein, Seungwon Lee, Luke Mandrake, Brian Bue
Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Gove Drive
Pasadena, CA 91109
818-393-5349, 818-393-7720, 818-354-1705
{ben.bornstein, seungwon.lee, luke.mandrake, brian.bue}@jpl.nasa.gov

Abstract—The Vehicle Cabin Atmosphere Monitor (VCAM) instrument is designed to autonomously detect and identify trace organic species in the International Space Station (ISS) cabin air and monitor changes in species concentrations over time after chemical events. The physical instrument is comprised of two subsystems. The first subsystem is a preconcentrator gas chromatograph (PCGC) which separates chemical analytes in time, based on compound specific properties such as molecular weight. The second subsystem is a Mass Spectrometer (MS) which measures the abundance of ionized analytes, separated in the GC phase, at specific mass-to-charge ratios. The VCAM PCGC/MS produces a time-series of mass fractionation patterns, indicative of the chemical compounds present, which is used for subsequent compound detection, identification, and quantification.

In order to autonomously identify and quantify chemical species from the PGGC/MS data, VCAM employs a variant of the de-facto industry standard Automated Mass Spectral Deconvolution and Identification System (AMDIS) algorithm developed by the National Institute of Standards and Technology (NIST). AMDIS was chosen first for its superior performance, when compared to a neural network classifier developed in-house and a proprietary, third-party, commercial algorithm, and second for its reputation within the mass spectrometry community. In this paper we provide an overview of AMDIS, including GC peak identification and spectral matching, as well our variations and additions to the core algorithm for performing mass calibration beforehand and species quantification afterward. We also discuss some of the challenges faced creating an independent implementation of AMDIS for delivery to VCAM flight software. Testing our algorithm, both individual components and in its entirety, was a particularly challenging, as the VCAM instrument was still in development and only periodically able to produce validation datasets[12].

[1] 1-4244-1488-1/08/$25.00 ©2008 IEEE

[2] IEEEAC paper #1327, Version 8, Updated December 14, 2007

TABLE OF CONTENTS

1. INTRODUCTION

The Vehicle Cabin Atmosphere Monitor (VCAM) instrument is designed to autonomously detect and identify trace organic species in the International Space Station (ISS) cabin air and monitor changes in species concentrations over time after chemical events [3]. The physical instrument is comprised of two subsystems. The first subsystem is a preconcentrator gas chromatograph (PCGC) which separates chemical analytes in time, based on compound specific properties such as molecular weight. The second subsystem is a Mass Spectrometer (MS) which measures the abundance of ionized analytes, separated in the GC phase, at specific mass-to-charge ratios [1]. The VCAM PCGC/MS produces a time-series of mass fractionation patterns, indicative of the chemical compounds present, which is used for subsequent compound detection, identification, and quantification.

2. SYSTEM OVERVIEW

The VCAM Data Analysis Software (DASW) transforms a series of raw ion counts from a GC/MS instrument run into a list of chemical compound identifications and quantifications. At a high-level, the data analysis software has four major components: mass calibration, GC peak finding, compound identification, and compound quantification (Figure 1).

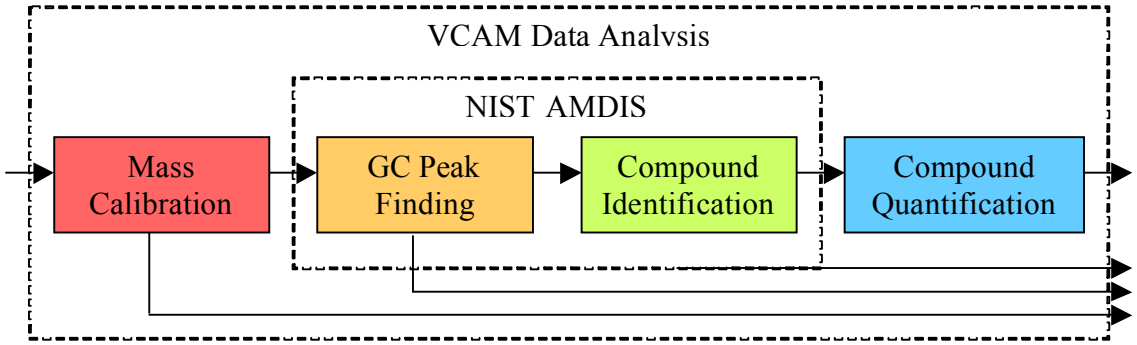


Figure 1: A block diagram illustrating the four major components of the VCAM DASW and information flow among them.

3. Mass Calibration

The mass calibration routine maps raw instrument channel numbers to corresponding atomic mass units for later use in GC peak finding and MS spectral matching (compound identification). The algorithm aligns ion counts integrated over time with the two earliest high-intensity channels with the expected mass positions of major constituents nitrogen (N_2 at 28 AMU) and oxygen (O_2 at 32 AMU), and later peaks with calibration gases acetone (CH_3COCH_3) and fluorobenzene (C_6H_5F) (when available) and performs a mass-channel least squares fit. As a sanity check, an approximate mass-channel mapping is derived from the Mathieu equation [4] by using parameter values obtained from the VCAM RF ramp hardware registers. A detailed description of this procedure is follows.

We compute a total ion chromatogram (TIC) from the mass spectra by summing ion counts over mass channels for each scan time:

$$TIC(scan) = \sum_{channel} I(scan, channel)$$

Then we smooth the TIC by averaging over adjacent neighbor scans using a window of size three:

$$J(scan) = \frac{I(scan-1) + I(scan) + I(scan+1)}{3}$$

From the smoothed TIC, we calculate the overall noise factor (N_f) of the data. The noise factor is identical to that performed by NIST AMDIS [7]. Scans are divided into 12 channel segments and the mean of each segment is computed. Within each segment, the number of times the data, taken in pairs, "cross" the mean is counted. If the number of crossings is larger than six, the segment is marked as nominal background. The median of the difference between a segment's ion counts and the average ion count present in the nominal background segments is taken. Call the median of the difference md(segment). The median of md(segment) is the noise factor of the total ion chromatogram.

If a noise factor cannot be computed, the whole total ion chromatogram is treated as a single elution peak. Once the noise factor is found, elution peaks are located using J(scan) and N_f.

Candidate mass peaks above background are chosen such that their width is at least seven channels and the preliminary peak height is greater than the threshold given by the noise factor [2, 7].

$$S > N,$$

where

$$S = J(peak) - J(background)$$

$$N = 5N_f\sqrt{J(peak)}$$

If no elution peaks are found, the entire total ion chromatogram is treated as a single elution peak.

When only one elution peak is present (including the case that the total chromatogram is treated as one elution peak), the peak is assumed to be air. When multiple elution peaks are found, the first peak is treated as an air peak and the remaining peaks are treated as potential acetone peaks.

Next we find channel peaks of each TIC elution peak. Channel peaks are found as follows. First compute the mass spectrum of one elution peak.

$$I(channel) = \sum_{scan=\text{left limit of the peak}}^{scan=\text{right limit of the peak}} I(scan, channel)$$

Second, find the width of the largest channel peak. The largest channel peak's width is defined as the channel range within which the ion count keeps descending as it goes away from the largest peak. Half of the largest channel peak width is used as the threshold of the peak width to find other channel peaks. If an ion count of a given channel is

the largest within the threshold range (channel – threshold_width, channel + threshold_width), the channel is considered a channel peak.

For the air (N_2 and O_2) elution peak, we then identify the two largest channel peaks (channel A and channel B) and assign them masses 28 and 32. The smaller channel number is assigned mass 28 and the larger channel number is assigned mass 32. Using the two channel-mass pairs (channel A, 28) and (channel B, 32), we find the linear equation for the relationship between channel number and mass charge (m/z).

We refine this initial fit by applying the mass-channel linear equation to other channel peaks. In particular we find the channel peaks (channel C and channel D) that are the closest to atmosphere major constituent calibration masses for argon (40 AMU) and CO_2 (44 AMU). The four channel-mass pairs (channel A, 28), (channel B, 32), (channel C, 40), (channel D, 44) are fit to a linear equation for the relationship between channel number and mass charge (m/z):

$$mass = a \times channel + b$$

With a four-point channel-mass calibration, we search for acetone. For each of the potential acetone elution peaks the two largest channels are identified (channel E, channel F) and assigned calibration mass 43 and 58. We record two channel-mass pairs: (channel E, 43) and (channel F, 58). By combining the two channel-mass pairs from the acetone peak with the four channel-mass pairs from the air peak we obtain a six channel-mass pair linear fit. Finally, we asses the quality of the fit.

If only an air peak was found, we use the linear equation from the air elution peak to estimate the error of each calibration mass (28, 32, 40, and 44). If both an air peak and a potential elution peak are used in the mass calibration, use the linear equation from the two elution peaks to estimate the error of each calibration mass (28, 32, 40, 43, 44, 58).

$$Error(m) = m - (a \times channel + b)$$

If the maximum error is smaller than 0.5 AMU, the mass calibration is considered successful. For compound identification, masses are binned to a resolution of 1 AMU to allow matching of mass fractionation patterns to those in the NIST Spectral Library [6, 10]. A mass calibration error greater than 0.5 AMU results in incorrect mass patterns and results poor identification performance.

4. GC PEAK FINDING

Gas Chromatograph (GC) peak finding is performed according to the NIST Automated Mass Spectral Deconvolution and Identification System (AMDIS) algorithm [2, 7].

The VCAM GC/MS unit produces data as a 2D grid of ion counts with axes representing discretized elution time vs. mass to charge ratio in units of AMU per fundamental electron charge (Figure 2).

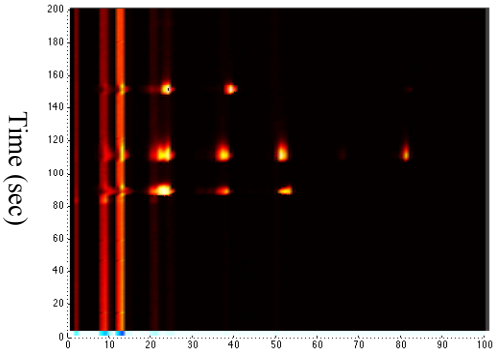


Figure 2: An overhead view of a VCAM GC/MS image, color-coded according to ion count. Notice the persistent nitrogen and oxygen mass lines present throughout the GC/MS run.

It is customary to refer to the m/z ratio simply as "mass" and assume units of AMU despite this not being precisely correct if multiple ionizations are possible. Multiple ionizations permit apparent fractional "masses" to appear, and the scanning resolution the VCAM MS sensor effectively detects sub-AMU accuracy. However, as the identification library we intend to use [6, 10] has only been recorded to integer AMU/z accuracy, our first step is to sum all bins such that our resolution along the mass axis only measures integer values of AMU/z. In our case, this results in mass channels ranging from 20 to 400. We have considered, as a means of increasing identification accuracy, using the full mass/charge resolution of the device; however, this would entail the manual creation of a compound identification library either in very controlled, reliable conditions using highly standard equipment. In our case, the labor required for such an approach was beyond the scope of our project, and thus we instead sacrificed sub-integer mass accuracy. With our data in this format, the AMDIS identification method [7, 5, 6] is then broken into four main algorithms: noise analysis, time peak identification, processing identified peaks into potential compound spectra, and finally matching these potential compounds against the library standards. In the coming description, we will be working individually with mass channel data as well as TIC, which is merely the summation of all ion counts across all channels (Figure 3).

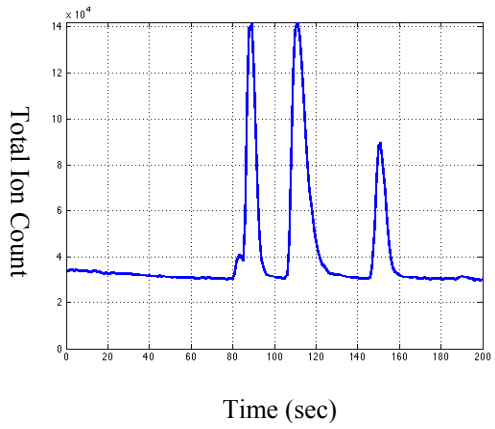


Figure 3: Ion counts vs. time summed over all mass channels (TIC) with time sample # on the horizontal. This plot can be viewed as a vertical slice of Figure 2.

Where TIC may pick up weak elution events present across many channels but relatively weak in each, it may miss activity in only one or two channels that would be quite apparent in a channel-by-channel approach. Thus, both 380 individual mass channels and the TIC signal are used in all analysis below. The original AMDIS method includes details to estimate a "lowest sensitivity" for GC./MS instruments, such as anything less than five ion counts will be truncated to zero; however, the VCAM MS sensor accurately measures even single ion counts, and thus no such logic was required for our AMDIS adaptation.

Noise Analysis

AMDIS depends crucially on the concept of accurate recognition of peaks of ion counts in time. Determination of a peak's veracity, however, depends critically on the relative size of the potential peak versus the ambient noise in the measured signal. Thus we require an estimate of this noise in some meaningful unit. We define a "noise factor" (N_f) as the mean fluctuation of a "calm" (no elution event) region of our grid divided by the square root of the mean signal. To calculate this value, we break each channel's signal (including TIC as one such channel) into segments of 12 time samples. Each segment is then examined to count the number of times its signal crosses its mean value. If that number is less than 4, the segment is rejected as likely having an elution event or instrument instability occurring. For each accepted segment, we then compute:

$$\frac{\text{mean}(|\, signal - \text{mean}(signal)\,|)}{\sqrt{\text{mean}(signal)}}$$

This yields a distribution of candidate noise estimates. The median of these noise estimates is then calculated, yielding our final noise factor (N_f). We are now ready to define the significance of any particular peak relative to this noise estimate. This procedure must be done once per dataset and may fail if there are no regions of sufficiently "flat" signal such as if there is insufficient lead or lag time surrounding elution events.

Temporal Peak Identification

For each channel, we now scan for any significant peaks. For each local maxima, we begin expanding a surrounding local window. This expansion occurs to the left (earlier time) and right (later time) independently. A maximum size of 12 time samples is permitted in either direction. Our estimate for the noise of this particular window is N_f x sqrt(smallest_signal), where the smallest signal is for this window only. We then begin to expand, sample by sample, our window about the central peak. If the signal rises five times the previous noise estimate above the lowest signal thus seen within the expanding window, we determine we have left the influence of the current peak and stop expansion. If the difference between the current signal and the peak signal has become more than 95% of the peak's height relative to the window's lowest signal, we determine we have captured the entire peak and stop expansion. Finally, if we encounter two sequential zero signal samples, we conclude the elution event must be over and stop expansion. Thus, utilizing these three conditions, we now have a window bracketing the current local signal maxima. Note that these windows can and usually do overlap one another in noisy samples. Typical GC/MS time profiles have sharper onset rise and a smoother, longer decay tail, and thus these windows are usually asymmetric about the central peak.

Within this window, we now calculate three basic peak parameters: the relative peak height, a precision estimate of the true peak height in time, and the maximum sharpness of the peak. Relative peak height is a somewhat ornate calculation wherein we first fit a line to the lowest sample left and right of the peak in the window and subtract this out as our first initial guess at bias and instrument drift. We then fit a second line through all of the lowest half points within the window. The relative height is then determined by the absolute peak height minus this second baseline plus the mean relative height of the two samples to the immediate left and right of the main peak, divided by the noise estimate of the peak given simply as N_f x sqrt(peak_absolute_signal). The second peak parameter of precise location in time is determined by performing a parabolic fit to the peak maxima and its two nearest neighbors. Finally, the maximum sharpness is the maximum of the following quantity calculated for all t within the window of interest:

$$sharpness[t] = \left| \frac{signal[t_{peak}] - signal[t]}{(t - t_{peak})(N_f \sqrt{signal[t_{peak}]})} \right|$$

This sharpness maxima is found independently for the left and right windows relative to the central peak, and the final reported sharpness is the sum of both. Armed with these

three parameters, we may now judge the peak to be of interest or not. If the relative peak sharpness is not at least two for the TIC or three for individual channels, the peak is rejected as too broad. Likewise, the peak is rejected if the relative height fails the following where min_height is 12 for TIC and three for individual channels:

$$relative\_peak\_height > \frac{total\_peak\_width + 4.0}{total\_window\_size \cdot min\_height}$$

This process generates hundreds to thousands of validated peaks for a typical dataset. We are now faced with packaging these peaks into coherent groupings of candidate compounds for comparison using their elution times and peak shapes as guides.

Compound Packaging

The problem of identifying a group of peaks as part of a single candidate compound is twofold: first we must determine which peaks go together, and second whether two extremely close potential compounds are in fact two overlapping coelutions or a single noisy elution. To do this, we define a grid ten times as dense in time as our sample data. Running through all peaks previously validated from individual channels or the TIC, we add to each of these bins any peak's sharpness whose precise maximizing time (estimated by the parabolic fit above) falls within said bin. We also keep track of the "dominant" peak within each bin, namely that peak which possesses the greatest sharpness. Once this dense grid is formed with sums of sharpness within, we walk the grid until we find a bin with nonzero value. If this bin's value is not larger than bins within two to the left and right, the bin is rejected as a potential compound elution as not being robust enough compared to nearby potential elutions. Next we form the nearby_sum of all bins within two to the left and right. From this value, we calculate an estimated width over which this bin must be the dominant event to be considered a compound elution. This time width (in bin number) is given by the max(2, int(150/nearby_sum)). If this bin's value is greater than any other bin within this time width to the left and right, it is accepted as a compound elution at the precise time represented by this bin.

For each compound elution time accepted above, we walk the entire list of peaks found in the peakfinding step. Any peak which a) is within the time width window given above of the current elution time, and b) has an area normalized height at least 75% of the dominant peak for this bin, we add as belonging to this elution event. Once our list of peaks is complete for each compound elution, we add together the original signals of every contributing peak within the window of this elution event to obtain a representative "model" of this particular elution. This model, baseline removed, will be then used to determine the mass spectrum actually reported to the identification step. This is done by performing a least squares fit of the form:

$$mass\_channel\_signal = b + m \cdot \langle 1,2,3,4,5,.. \rangle + c \cdot \langle model \rangle$$

where b is an overall ignored signal bias, m is an ignored linear trend attributed to device drift within a single elution time, and c is the actual contribution of this mass channel to the elution event (the value of interest). Finally, various flags can be added to this particular mass channel's contributing peak indicating it has a small signal/noise ratio (background peak), more than two zero counts are present within the signal window (could be spurious noise spike), or the fit error using the compound model is too large (badly modeled peak may not actually be part of this elution). These flags penalize individual peaks in the last step, that of compound identification.

5. COMPOUND IDENTIFICATION

Compound identification (MS spectral matching) is performed according to the NIST AMDIS algorithm [7, 5, 6].

We now have a list of potentially dozens of elution times (to one tenth sample time accuracy) with associated of mass channel peaks and relative amplitudes, in addition to flags describing hazardous characteristics. To identify an extracted elution, each elution must be matched against entries of known compounds within a pre-existing library. The "match factor" that describes how well a candidate compound matches a library compound is formed via the summation of the dot product of mass spectra peaks in the candidate compound with the library entry to which it is being compared, weighted by the square of the mass at each channel, and penalized by any associated flags.

$$match\_factor_{library} = \sum_{m}^{mass\_channels} peak_{unknown,m} \cdot peak_{library,m} \cdot flag_{m} \cdot m^2$$

Heavier masses are more significant as they represent larger (and hence more unique) molecular fragments thus bearing greater identification potential. Unfortunately, the VCAM instrument's current sensitivity decreases with larger mass, which makes the precise weighting a matter of debate and research. For each elution time event we now have a list of match factors (approximate likelihood) for each entry within the spectral library. The library compound with the highest match factor is reported as the best possible match, and its match factor is provided so the user may have some estimate of trustworthiness.

6. COMPOUND QUANTIFICATION

Compound quantification analysis reports the concentration (in parts-per-million) of compounds that have been identified by the NIST AMDIS spectral matching algorithm. VCAM project scientists Ara Chutjian, Murray Darrah, and John MacAskill provided the compound quantification method. To quantify compound concentration, first the total ion count under an elution peak (signal) is found. Second, this ion count is compared against a concentration curve (one for each compound to be quantified), to arrive at the total compound concentration. Concentration curves are determined empirically, on the ground, with the VCAM development and protoflight units. The family of concentration curves is parameterized by two constants (α and β) as follows:

$$signal = \alpha(concentration)^{\beta}$$

Transforming the concentration curve to log-log space yields constants a and b and the total concentration:

$$\log(concentration) = a\log(signal) + b$$

7. INITIAL RESULTS

A major challenge in verifying the VCAM DASW identification and quantification algorithms has centered on the dearth of data currently available for testing. The VCAM instrument is still under active development and detailed and dedicated compound testing has only recently become the scientist and instrument team's highest priority. Table 1 presents a snapshot of all data analysis identification results on all readily available compounds circa March 2006, the last time compounds for run through the instrument for hardware detection and software identification testing. In come cases, the data analysis algorithms performed well and in others, greater accuracy is certainly desired. With the exception of acetone however, none of compounds were present in large enough numbers to provide meaningful accuracy statistics. At the time of this paper, the VCAM team has successfully uncovered and recovered from several instrument contaminations events that were hindering compound detection by the hardware and also confounding data analysis software. Compound testing has begun in earnest and we fully expect the data analysis software will meet VCAM mission accuracy requirements.

Sample	Correct	Total Samples	Accuracy (%)
acetaldehyde	1	1	100
acetone	72	80	90
dichloromethane	3	4	75
ethanol	8	19	42
methanol	2	5	40
1-butanol	5	6	83
2-propanol	3	10	30
benzene	2	3	67
heptane	2	3	67
pentane	3	3	100
toluene	5	5	100

Table 1: A snapshot of data analysis and identification rate results on all readily available compounds circa March 2006.

8. CONCLUSION

We presented our variant of the de-facto industry standard Automated Mass Spectral Deconvolution and Identification System (AMDIS) algorithm developed by the National Institute of Standards and Technology (NIST). VCAM employs AMDIS to autonomously identify and quantify chemical species from PCGC/MS data. In addition to stock AMDIS peak-finding and spectral matching, we augmented the AMDIS method with mass calibration on the front-end and compound quantification on the back-end. Analysis results on initial laboratory datasets are promising, but more testing is required.

ACKNOWLEDGEMENTS

The research described in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. We thank the VCAM project scientists Ara Chutjian (PI), Murray Darrah (Co-I), Richard Kidd, John MacAskill, Dan Fry, Erge Edu-Fry, and VCAM flight software leads Vance Haemmerle and Mike Girard for their many insights and valuable contributions to this work.

REFERENCES

[1] A. Chutjian, et al., "Overview of the Vehicle Cabin Atmosphere Monitor, a Miniature Gas Chromatograph/Mass Spectrometer for Trace Contamination Monitoring on the ISS and CEV," *Proceedings of the 37th International Conference on Environmental Systems*, SAE Technical Paper Series (paper number 2007-01-3150), 2007.

[2] R. G. Dromey, M. J. Stefik, T. C. Rindfleisch, and A. M. Duffield, "Extraction of Mass Spectra Free of Background and Neighboring Component Contributions from GC/MS Data," Journal of Analytical Chemistry, Vol. 48(9), 1368–1375, 1976.

[3] D. L. Jan, "Environmental Monitoring Instruments: Using ISS as a Testbed for Exploration," *Proceedings of the 2007 IEEE Aerospace Conference*, Big Sky, Montana, March 2007.

[4] R. E. March, "An Introduction to Quadrupole Ion Trap Mass Spectrometry," *Journal of the American Society of Mass Spectrometry*, Vol. 32, 351–369, 1997.

[5] S.E. Stein, "Estimating Probabilities of Correct Identification from Results of Mass Spectral Library Searches," *Journal of the American Society of Mass Spectrometry*, Vol. 5, 316–323, 1994.

[6] S. E. Stein and D. R. Scott, "Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification," *Journal of the American Society of Mass Spectrometry*, Vol. 5, 859–866, 1994.

[7] S. E. Stein, "An Integrated Method for Spectrum Extraction and Compound Identification from GC/MS Data," *Journal of the American Society of Mass Spectrometry*, Vol. 10, 770–781, 1999.

[8] H. Zhou, M. L. Homer, A. V. Shevade, and M. A. Ryan, "Nonlinear Least-Squares Based Method for Identifying and Quantifying Single and Mixed Contaminants in Air with an Electronic Nose,", *Proceedings of the 2005 IEEE Sensors Conference* (ISSN 1424-8220), 2005.

[9] NIST Automated Mass Spectral Deconvolution and Identification System (AMDIS) software, http://chemdata.nist.gov/mass-spc/amdis/.

[10] NIST/EPA/NIH Mass Spectral Library (NIST 05) ASCII Version, http://www.nist.gov/data/nist1.htm.

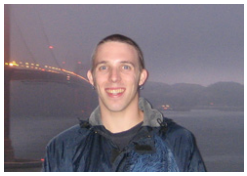
BIOGRAPHY

Ben Bornstein *is a senior member of the Machine Learning and Instrument Autonomy group at the Jet Propulsion Laboratory in Pasadena, CA. He is the lead engineer for VCAM data analysis and its flight software implementation. Ben enjoys bringing machine learning techniques and considerable hacking (programming) skills to bear to solve problems in geology, remote sensing, bioinformatics, and systems biology. He has designed and implemented software systems for several Caltech biology labs, the Institute for Genomics and Bioinformatics (IGB) at UC Irvine, USC Children's Hospital, and JPL's Mars Exploration Rover (MER) project. He is also the inventor of and lead developer for LIBSBML, an open-source library for the Systems Biology Markup Language (SBML). Ben received a B.Sc. in Computer Science from the University of Minnesota Duluth in 1999 and is pursuing a M.Sc. in Computer Science at the University of Southern California.*

Seungwon Lee *is a senior member of the High Capability Computing and Modeling Group at Jet Propulsion Laboratory. She is involved in projects developing flight instrument software for Vehicle Cabin Atmosphere Monitor and conducting research on materials modeling and simulation, nonlinear dynamics control, spectral retrieval, data reduction, global optimization, parallel computing, and advanced numerical algorithms. She received her Ph.D. in Physics from the Ohio State University and her M.S. and B.S. in Physics from the Seoul National University.*

Luke Mandrake *is a member of the Machine Learning and Instrument Autonomy group at the Jet Propulsion Laboratory in Pasadena, CA. He is involved in the application of machine learning techniques to current research endeavors with various Earth-sensing satellite systems and particularly enjoys building and studying computational models of natural systems. Luke*

received his PhD. and M.S. in computational plasma physics from UCLA and his B.S. in engineering physics from the University of Arizona.



Brian Bue *is a research programmer in the Machine Learning and Instrument Autonomy group at the Jet Propulsion Laboratory, where he participates in projects involving software and algorithm development for Earth and planetary science data analysis. In the past, he has done research in computational geomorphology, automated terrain analysis, planetary image processing and scientific visualization. He received a M.S. from Purdue University in Computer Science, and Bachelor's degrees from Augsburg College in Computer Science and Mathematics. He is currently pursing a Ph.D. in Electrical and Computer Engineering at Rice University.*