Adapting AMDIS for Autonomous Spectral Identification of Hazardous Compounds for ISS Monitoring

Lukas Mandrake, Seungwon Lee, Benjamin Bornstein, Brian Bue

Jet Propulsion Laboratory
California Institute of Technology
4800 Oak Grove Drive
Pasadena, CA 91109
<firstname.lastname>@jp.nasa.gov

Abstract—The stand-alone Vehicle Cabin Atmospheric Monitor (VCAM) instrument was designed to provide an automated method of monitoring air quality within the International Space Station (ISS) via a miniaturized mass spectrometer and gas chromatograph system.[1][2] The output of the device, a series of mass spectra as a function of time, is then processed via our implementation of the Automated Mass Spectral Deconvolution and Identification System (AMDIS) method from the National Institute for Standards and Technology (NIST) to generate potential identification with reference to a known library of hazardous chemicals. In this paper we discuss the modifications required to the AMDIS method for autonomous in-flight operation as well as additions beyond the original method. In particular, the original AMDIS method contains numerous parameters that were intended to be adjusted by an operator during the analysis to reduce false positives and adjust sensitivity. We have instead implemented solution filtration based on elution time and discuss possible arbitration algorithms for close similar matches to provide the user with a more succinct, single-valued answer.

TABLE OF CONTENTS

1. INTRODUCTION

The ISS exists as a fragile bubble of life-sustaining atmosphere in an otherwise lethal environment. However, unlike the large biologically and chemically mediated gasses of Earth, the station must constantly monitor and adjust its atmospheric components manually. Besides oxygen generation and water vapor sequestration, carbon dioxide is physically and chemically adsorbed via silica gel and molecular sieves while myriad trace contaminants are contained with activated carbon and thermal catalytic oxidation [1]. Future spacecraft and station missions already planned will require much smaller and more efficient systems than those currently employed, making life support an area of intense modern study. The importance of this research is magnified by the presence of many compounds with potential health consequences for airborne exposure that are used in functioning devices, experiments, and cargo loads aboard the ISS. These compounds can require special action to remove them from the crew atmosphere should a leak or out-gassing occur or may even trigger an evacuation. Biological activity from fungi and bacteria inadvertently brought aboard also can produce toxic components. Regular air quality checks are performed on sample bags returned to Earth, but the need for a fast-response analysis has been understood for some time [2][3]. However, many of these systems are hard-wired for detection of specific compounds during construction, narrowing their applicability should a novel event occur.

NASA's soon to be launched VCAM device represents the first fully autonomous miniaturized gas chromatograph mass spectrometer (GCMS) system for flight application [4]. While the system will be tuned to a specific group of compounds of concern, the generic GCMS system implementation permits identification and concentration estimation to be made on future contamination within a concentration range of 0.01 to 100 parts per million. Further, the device's library may be updated during operation aboard the ISS should there be unforeseen mission needs such as a novel chemical leak. To process the time series of mass spectra generated from the device, the Automated Mass Spectral Deconvolution and Identification System (AMDIS) algorithm developed by the National Institute of Standards and Technology (NIST) [5][6] was chosen based on its performance and generally accepted reputation among the mass spectrometry community.

1———————

[1] 978-1-4244-2622-5/09/$25.00 ©2009 IEEE

[2] IEEEAC paper#1296, Version 4, Updated 2008:10:29

2. Brief System Overview

Critical Device Parameters

The VCAM device is a microwave oven sized payload less than 30 kg in weight and requiring 70 to 180 W of power during operation. It will operate under nominal conditions for twelve months until its supply of carrier He and calibrants are exhausted, at which point the crew may replace the supply tanks and operation may resume. Samples will be fed to the device once per day with a calibration run once per week. Raw data will be downlinked to Earth as well as autonomously processed on-board.

Gas Chromatograph

The device's operational flow proceeds as follows. Air is pumped into a charcoal bed called the preconcentrator. Many trace chemicals in the air adsorb into this matrix. Minutes later, the preconcentrator is gently warmed to encourage nitrogen, oxygen, and water to leave the system (smaller compounds require lower temperatures to evacuate based on boiling point and chemical affinity). A sharp heat pulse is then used to drive off remaining gasses while a flow of pure He gas carries the outgas products deeper into the machine. This puff of gas passes through a 10 meter capillary tube (gas chromatographic column) that uses a special internal coating to encourage the separation of various chemical species by modifying their velocities within the tube. Full separation takes 20 minutes. This process is the gas chromatography aspect to the machine: the separation in time of various air components by mass, chemical family, and polarity.

Mass Spectrometer

As each component exits the tube, the gas enters the mass spectrometer aspect. First it is ionized by an electron beam and stored in a magnetic trap. This ionization process does not strip all atoms into free ions but instead produces a characteristic fractionation pattern of smaller molecular ions unique for each particular parent molecule. A voltage ramp is then applied to the trap in such a way as to encourage those molecular ions with the smallest mass to charge ratio to escape first, followed by successively larger mass/charge ratios. Each molecular ion is counted as it leaves the trap by a channel electron multiplier (CEM) high voltage sensor. Fifty such spectra of counts per mass/charge are produced per second and averaged to yield a characteristic fractionation pattern unique to each compound of interest. 1200 such average mass/charge spectrum are produced in sequence as various gas components escape the gas chromatograph (elute) and are processed. Half of these are taken at "high gain" (dense ionizing electron beam, more ionization) and half at "low gain" (sparse ionizing electron beam, less ionization) to encompass a greater range of input gas concentration. The resulting 2D dataset can be visualized as a grid of ion "counts" where one axis is charge/mass and the other is time. Once such a grid is produced (one for high gain, one for low), the physical device's work is complete and autonomous software takes over to determine compound presence, identity, and concentration. It should be noted that the raw output of the device is not mass calibrated, that is 4096 channels of unknown mass/charge are output. An autonomous mass calibration algorithm developed by Lee [7] performs this mapping before any autonomous identification begins.

3. Library of Interest

The VCAM instrument has a software-defined library of compounds of interest stored internally. This library is used to identify potential compounds and to determine their concentrations. Table 1 shows the compounds originally specified for VCAM's launch, though these can be modified via upload at any time. The spectra for these compounds were taken from the NIST spectral database [8], a generally accepted industrial standard, while all other library-specific information (elution time and concentration coefficients) must be produced on a VCAM-like instrument on the ground.

Nature generates fractional values of mass/charge via multiple ionization and mass defect effects. The NIST spectral database, however, uses integer bin resolution for the mass/charge ratios of all its compound records. VCAM internally measures 4096 channels of mass/charge information, but these are folded down to 400 1-AMU channels to be compatible with the NIST database. The relative peak heights for each compound are stored as numbers between 0 and 999, normalized to the highest peak of each spectrum.

Table 1. Compounds of Interest

Compound	CAS
1,2-dichloroethane	107-06-2
1,2-propylene glycol	57-55-6
1-butanol	71-36-3
2-butanone	78-93-3
2-propanol	67-63-0
4-methyl-2-pentanone	108-10-1
acetaldehyde	75-07-0
acetone	67-64-1
benzene	71-43-2
C5 aldehyde (pentanal)	110-62-3
C5 alkane (pentane)	109-66-0
C6 aldehyde (hexanal)	66-25-1
C6 alkane (hexane)	110-54-3
carbonyl sulfide	463-58-1
chloroform	67-66-3
dichloromethane	75-09-2
ethanol	64-17-5
ethyl acetate	141-78-6
ethyl benzene	100-41-4
fluorobenzene	462-06-6
freon 11	75-69-4

freon 113	76-13-1
furan	110-00-9
hexamethylcyclotrisiloxane	541-05-9
isoprene	78-79-5
limonene	138-86-3
m-xylene	108-38-3
octamethylcyclotetrasiloxane	556-67-2
o-xylene	95-47-6
perfluoropropane	76-19-7
p-xylene	106-42-3
toluene	108-88-3
vinyl chloride	75-01-4

To understand how difficult unique identification of all entries within this library will be, a graph versus mass channel (AMU) was created where the y axis displays the ad-hoc "discrimination" D for a channel m

$$D(m) = \frac{N_{total} - N(m)}{N_{total} - 1},$$

where N_{total} is the total number of library compounds (33 in our case) and N(m) is the number of compounds which have a non-negligible spectral component at channel m ($I(m) >$ 5% of maximum peak). Thus, D(m) of zero indicates all compounds possess contributions at channel m removing most of that channel's discriminatory utility for identification, while D(m) of 1 indicates only a single compound utilizes this channel providing ideal discrimination. D(m) is treated as undefined if N(m) is zero. Figure 1 shows D in blue for our particular library as well as a running average in red. Note that higher masses are far more discriminatory than lower masses with zero or near zero discrimination possible atop masses 28 and 32 due to omnipresent air peaks leaking into the mass spectrometer chamber (vertical bars in Figure 2). Superimposed on this graph is a quadratic mass weight term $(M/M_{MAX})^2$ in preparation for Section 6.

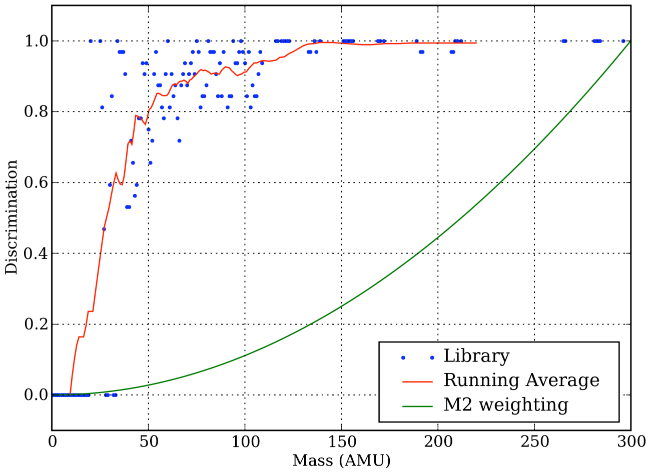


Figure 1. Channel discrimination as a function of mass. Higher masses are more uniquely identifying.

4. Peak Detection

Figure 2 shows an example ion count grid (with mass/charge ratio on the horizontal and time on the vertical. Note first the persistent peaks at 28, 32, and 44 AMU. These are ionized N2, O2, and CO2 molecules, the basic constituents of our atmosphere. They continually leak into the measurement chamber to some degree, causing a persistent signature that must be removed from analyzed spectra. We next note that at certain times structure suddenly emerges along the Mass/charge axis as a series of peaks. These are the mass spectra/fragmentation patterns for individual gas components as they elute from the gas chromatograph.

Though this grid of ion counts is the basic, raw data, a human analyst more often looks at two derived products: the total ion current (TIC) chromatogram, which sums the contribution of all mass/charge components into a single trace versus time (Figure 3), and a mass/charge spectra in the centroid time of an elution event, often as determined in the TIC. All three of these products (raw ion count grid, TIC, and mass spectra) are considered by the AMDIS method. Our implementation will be herein referred to as JPL AMDIS to distinguish it from the version available directly from NIST.

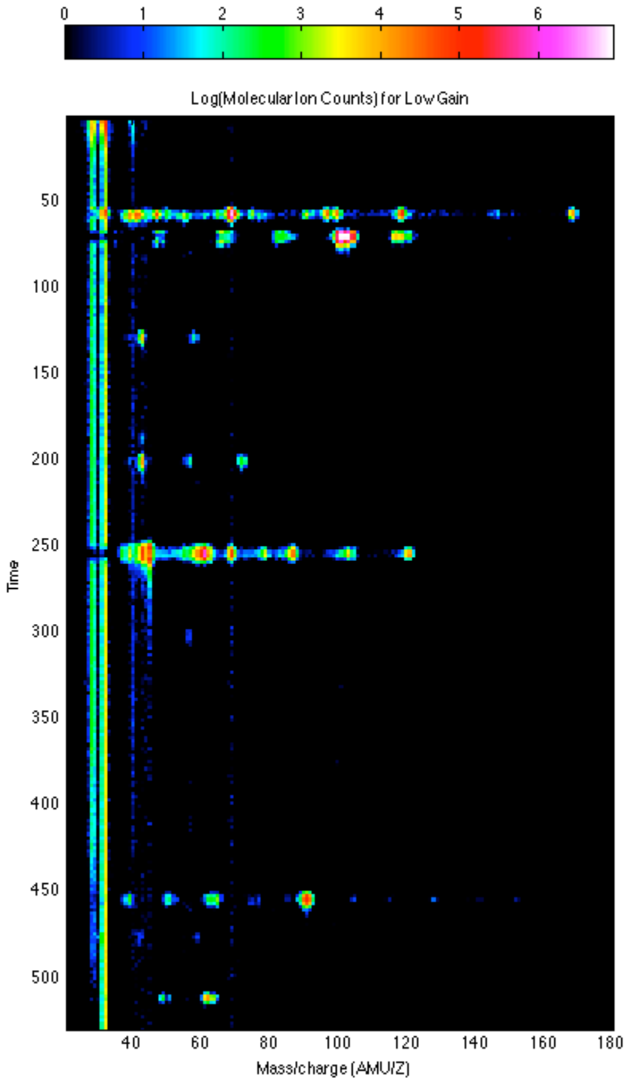


Figure 2. Molecular Ion log(Count) grid. Time = 0 is at the top of the y axis, while smaller mass/charge rations are to the left. A log scale has been used to enhance contrast of small peaks for display only.

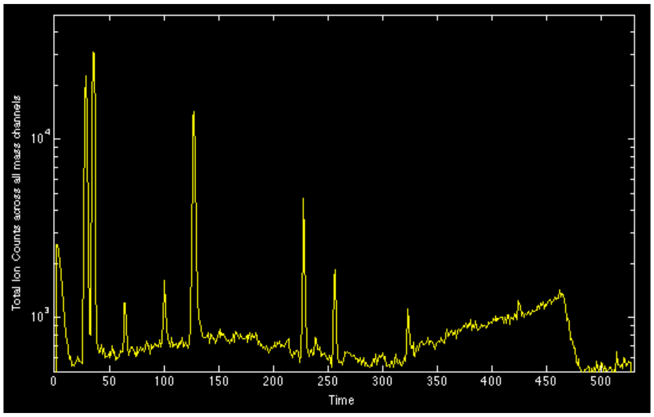


Figure 3. Total Ion Current (TIC) summed across all mass channels. Note the log y axis.

Estimating Noise

Our first task in autonomously simulating a human analyst is to examine the 2D count grid and identify the elution events (times when gas components exit from the chromatograph) for further study. The AMDIS method refers to this as peak extraction. Peak identification is achieved by first computing the "noise factor" N_f which is a unitless measure of the ambient noise well away from all peak events [6]. This factor is the ratio of the observed signal fluctuation to the square root of the mean signal strength (the expected noise a CEM generates). The observed signal fluctuation is obtained by counting the number of times a signal oscillates about its mean in a given window (12 time samples) and then taking the median absolute value of the observed deviation. Each window thus measured produces a sample N_f. This process is repeated once for the TIC and once for each mass channel using only windows that have no zero ion counts at any time and also cross their own mean value more than 1/3 of their length (at least 4 times). This ensures that neither peak events nor data sparse channels corrupt the N_f value. Finally, the median of all candidate N_f values is taken as the representative N_f for the entire dataset. It typically ranges from 0.8 to 5.0 in our experiments with 0.83 for the case shown in Figures 2 and 3.

Cataloging Peaks

Once a noise estimate is calculated, we may go about determining the location of all "relatively significant" peaks, that is, those that are high enough above the ambient noise to be considered elution events. The AMDIS method does this once using the TIC to capture low-concentration components that might not have strong individual channel traces, and once for each mass channel independently for single peak compounds that might have weak TIC signatures. These results are appended to form a master peak list.

To determine if a peak is suitable, it first must satisfy a variety of conditions. The full procedure is complex, but the rules may be summarized:

1) It must be an absolute peak (larger than immediate neighbors)

2) The total width of the peak must be greater than 3 time samples. Calculating this width is itself a complex process.

3) The relative peak height must be ~230% higher than the (surrounding background + anticipated background noise at peak height) if the peak is the narrowest permitted but only 30% higher if the peak is the broadest permissible.

4) The peak must have a sufficient maximum rate (slope) compared to the expected noise fluctuation

based on the maximum peak signal, i.e. there must be at least one sharp rise somewhere on the peak that cannot be explained by noise.

At the end of the peak extraction procedure we obtain a list of hundreds of peaks that exist in the TIC or an individual mass channel. Each individual elution event will usually generate several peaks: one for each major mass/charge line and one in the TIC.

5. Peak Clustering / Compound Generation

Armed with a sea of potentially significant peaks, we must now decide which peaks likely belong to the same compound and produce a mass spectra for each potential compound found. In simple cases this is trivial: simply gather "nearby" peaks in time together across all mass channels and call their appropriately normalized relative amplitudes the mass spectra. In more complex cases this procedure has two significant flaws. First, it assumes that each elution event is well separated in time, but coelution events can frequently occur (note the coelution event in Figure 3 around time 25-30), causing observed contribution peaks from one peak to be merged with the another. Second, the mass spectra can be affected by background bias with large but meaningless peaks added into the true spectrum such as from persistent chemicals leaking into the measurement chamber. We desire a more robust procedure than naïve single-time mass spectra extraction.

A more reliable method begins with the physical observation that every unique compound that elutes has a characteristic elution curve in time (Figure 4). Even when two compounds are nearby each other, these rise and set times help to identify signals on different mass channels originating from the same compound. AMDIS harnesses this observation in two ways: by clustering peaks which maximize at precisely the same time, and by constructing a mean model for each potential compound which is then used to extract a more precise mass spectra.

Peak Maximization Clustering

For model-building process, a parabola is fit to each peak and its two nearest neighbors to calculate a more precise estimate of the maximization time to 0.1 timestep resolution. Then, by the method of Colby [9], we calculate a maximum sharpness S for the slope of each peak over the entire peak extent window relative to the expected peak noise given by

$$S = \max\left|\frac{I_0 - I_n}{nN_f\sqrt{I_0}}\right| \quad \forall \quad n \in window_{peak}.$$

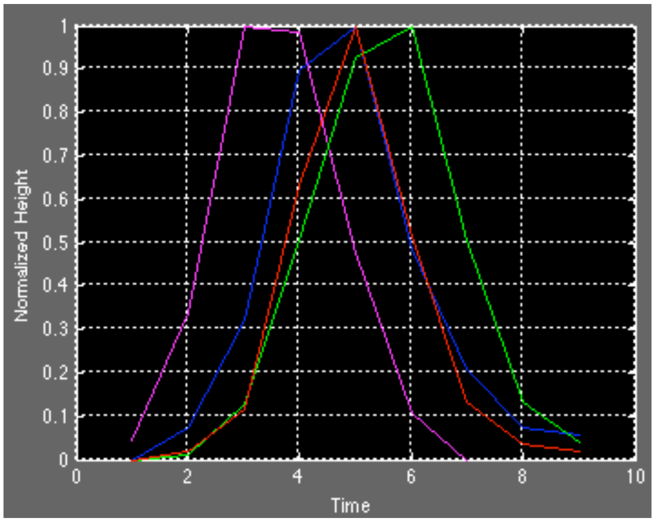


Figure 4. Peak shapes for the 3rd through 7th peaks in the TIC of the example in Figure 3 with the beginning of the peaks aligned. Note the different shapes and rise/set times.

where I_n is the intensity at a given index n (n=0 is the peak value) and N_f is the noise factor. These max sharpness values (one per peak) are added to an array of bins each representing 0.1 timesteps. Local maxima within this sharpness array are then used to detect candidate compounds, with compounds suppressed within an empirically derived range about each major compound candidate. The sharpest peak within a locally maximum bin is called the "dominant peak" of the compound, and any peaks with sharpnesses found to be within 75% of the sharpest peak that also maximize within the suppression window are added as contributing peaks.

Note that peaks detected in the TIC are a special case: as they have no relation to mass channels, they are added directly as potential compounds without further entries in their contributing peak list.

Compound Model / Mass Spectra

An average run generates a handful to dozens of potential compounds, each with a list of contributing peaks. The next step is to determine a mass spectrum for each which is achieved by constructing a model shape. The contributions of each peak, after suitable normalization and background subtraction, are averaged together to generate a model in time M(n). TIC contributions simply use the normalized shape of the TIC itself in the window of interest. These models are then fit to **every** mass channel individually using the least-squares method described by Dromey [5] by the equation

$$I(n) = a + b \cdot n + c \cdot M(n),$$

where a and b represent an uninteresting linear background and c is the projection of the mean model on the current mass channel. Taking all positive c and normalizing the largest peak to the standard NIST intensity range of 0 to 999, the mass spectrum for this compound is ready to be compared to the NIST library. It should be noted that the AMDIS method available from NIST includes the capability of fitting multiple models M(n), N(n), O(n) and so on to a single peak in the hopes of assisting deconvolution of coelutions. Due to resource constraints and desire for simplicity, we chose not to implement this feature and instead maintain a single-model resolver.

At this point quality control flags may also be recorded for each of the associated peaks within the spectrum. AMDIS notes several potential problems:

1) If the peak modeled by I(n) above represents less than ~25% of the total signal (i.e. a bad match to M(n)) the mass peak is rejected as not part of the spectra

2) If the peak has signal less than 3 times the expected noise at the peak maximum, it is flagged as a possible noise peak and downweighted in the identification step. This penalty ($W_{penalty}$) is expressed as a value from 0 (unpenalized) to 99 (maximally penalized) depending on peak height.

3) If the peak modeled by I(n) represents less than ~50% of the total signal the peak is downweighted as likely background similar to (2).

This procedure ends with a list of perceived compounds with associated mass spectra and quality control flags for each peak if necessary. The identification step compares these compounds with those within our library. An optimal example is shown in Figure 5 for a 1-butanol observation. The violet spectrum is the raw data directly from the VCAM instrument centered at the peak maximum. The dark superimposed lines are those peaks AMDIS found that were closely associated via peak maximization clustering. The green spectrum is 1-butanol from the NIST spectral library.

6. Compound Identification

Identification is the procedure of matching each perceived candidate compound with a library entry (or none if badly matched with all). Each compound is matched individually, independently of any other perceived compounds in the run. While this does not utilize all of the intuition available to a human operator (acetone always elutes after pentane but before 1-butanol, for instance), our later use of elution time filtration recaptures much of such time ordered information.

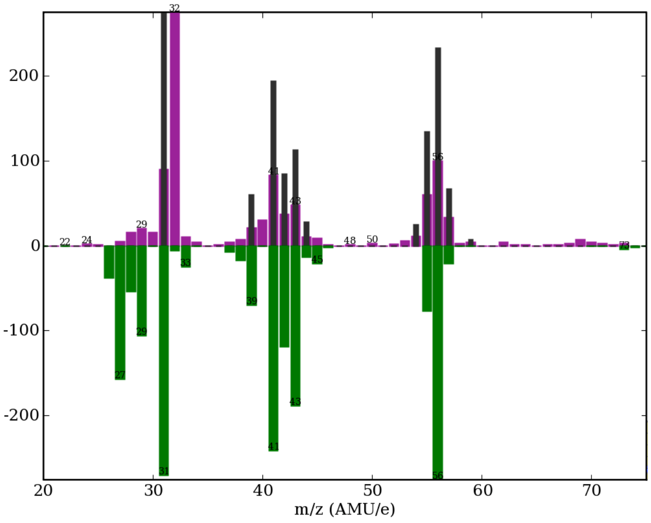


Figure 5. Sample 1-butanol spectrum. Violet = raw data. Grey superimposed = AMDIS clustered peaks. Green = NIST library.

In both NIST and JPL AMDIS, a compound is first subjected to a preliminary screening based on a simple dot product to establish general similarity between spectra:

$$S = \frac{\left(\sum_{i=1}^{N=\min(4,\mathrm{len}(L))} L_i \cdot U(M_i)\right)^2}{\left(\sum_{i=1}^{N} L_i^2\right)\left(\sum_{i=1}^{N} U^2(M_i)\right)},$$

where the sums over i range over the (at most) four largest peaks of the current library entry to compare (N<5). L_i is the height of the ith library peak, M_i is the mass, and U[M] is the height of the unknown compound as a function of mass (grey spectrum in Figure 5). In both user records and library records, peaks are ordered by their height such that the most significant peaks are compared first. Note that both U and L have been normalized such that their largest respective peaks are equal (in our case, 999). S represents the cosine squared of the "angular distance" between vectors L and U in the space of the four largest peaks of L. AMDIS requires S to be less than 0.4; otherwise, the compounds are seen to be too different for consideration and the library entry is rejected. This initial screen can introduce false negatives (missed identification) should the mass calibration be off by more than 0.5 AMU for any dominant library spectra peaks.

For those library entries remaining after the initial screening, we compute a match factor (mf) that measures more thoroughly any spectral similarities. AMDIS includes three separate match factors and one amalgam that we utilize in JPL AMDIS. All match factors range from 0 to 100.

The "simple" match factor follows a library spectrum and computes the dot product against the unknown spectrum. Note that there could be sizeable peaks in the unknown spectrum that are not sampled by this factor as they have no counterpart in the library spectrum.

$$mf_{simple} = 100 \cdot \frac{\left(\sum_{i=1}^{N}\sqrt{L_i \cdot U(M_i)}\right)}{\left(\sum_{i=1}^{N} L_i\right)\left(\sum_{i=1}^{N} U(M_i)\right)}$$

where N is now the total number of peaks in the library spectrum, L_i is the height of the ith library peak, M_i is the mass of the ith library peak, and U is the height of the unknown compound as a function of mass.

The "weighted" match factor includes a weighting factor for peaks at higher masses as being more unique or relevant. In our library, this consideration is certainly true (Section 3). It also reduces the penalty on library peaks that are not present in the unknown spectrum.

$$mf_{weighted} = 100 \cdot \frac{\left(\sum_{i=1}^{N} M_i^2 \beta_i \sqrt{L_i \cdot U(M_i)}\right)}{\left(\sum_{i=1}^{N} \alpha_i M_i^2 L_i\right)\left(\sum_{i=1}^{N} M_i^2 U(M_i)\right)}.$$

Here α_i is a penalty reduction term equal to 0.5 for peaks not present (nonzero) in the user spectrum and 1 for those that are. β_i is a penalty term which is zero for flagged peaks (in the unknown spectrum) and 1 for unflagged peaks.

The "reverse" match factor is precisely the same as the "weighted" factor except that β_i is now replaced with a more sophisticated penalty based on the $W_{penalty}$ quality flag determined during the compound model building step:

$$\beta_i = \frac{1}{1 + 0.2 \cdot W_{penalty}}.$$

These three match factors are finally combined in a complex way (beyond the scope of this paper) taking into additional account compound purity (total signal fit by the compound model), the number of common peaks between the unknown and library compound, the estimated noise at the peak of the compound, and whether the peak was taken from the TIC. The final number is referred to as the "Net" match representing a goodness of fit to a particular library entry.

7. Elution Time Filtration

Until now, we have only utilized our knowledge of mass spectra to assist in the identification of candidate compounds. However, as shown in Section 3, there are many compounds whose fractionation patterns are extremely similar: shifted by a single mass, subsets of another more complex pattern, or different only by relative abundance ratios. To assist our identification, we take advantage of the fact that various compounds elute at different, predictable times. This is the gas chromatography aspect of VCAM, and in principle it is possible to identify compounds solely upon their elution time. In our case, we will utilize this additional information to narrow the possible candidates before identification is complete. This logic extends the NIST AMDIS method beyond its typical configuration.

Our implementation is simple and efficient: windows of possibility are defined for each library entry. For each compound identification event, only the subset of the library known to be available at the time of the unknown compound is considered. This top-hat probability distribution is appropriately draconian in its removal of distant compounds, but care must be taken not to accidentally eliminate nearby candidates. In fact, these windows must be made to overlap due both to co-eluting compounds (compounds that are emitted at roughly the same time) and to account for run-to-run variability in elution time (a noise process). Addition of elution time filtration reduces false positive detection by up to 95% and greatly facilitates deciding between structurally similar compounds with separated elution times.

A common chromatographic measure of elution time is the Kovats Retention Index (RI) [10]. This is a unitless measure relative to the n-alkanes with logarithmic interpolation. However, our attempts to extract reliable values for RI from the NIST library were fraught with disagreement between sources. We instead elected to implement the above algorithm based on relative time (minutes) from a fixed point during the VCAM measurement cycle summarized in Figure 6. To show areas where many compounds share similar elution times, Figure 7 was produced as a histogram of Figure 6 using 20 second bins. Note the large number of compounds at the beginning of the run; this initial rush proves to be very challenging to an elution time filtration method since, in our case, up to seven compounds may have overlapping potential identification regions. The size of acceptance window for a compound is not precisely defined mathematically, being a function of peak width, run-to-run elution time variability which varies for each compound, and the precise point in a peak that AMDIS selects as the true maximizing event based on peak shape models and least squares fitting. The elution time windows used in this report were the "worst case" windows made by selecting the left-most part of any peak ever observed for a particular compound and the similar right-most part of any peak ever observed. Thus, these windows are very wide, permissive, and not as discriminatory as could be imagined. Hand-trimming of elution time acceptance windows on a given test dataset will permit significant improvement in the elution time filtration method's ability to resolve especially early eluters and will be performed as a final tuning step in VCAM's development.

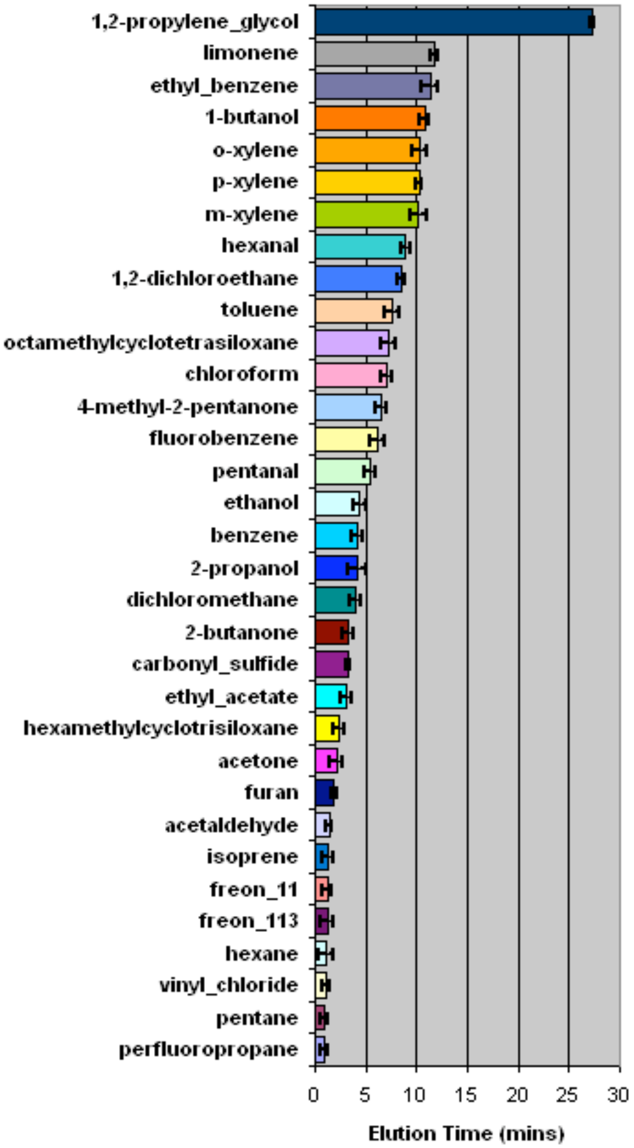


Figure 6. Elution times for each library entry in minutes with uncertainty estimate (acceptance width).

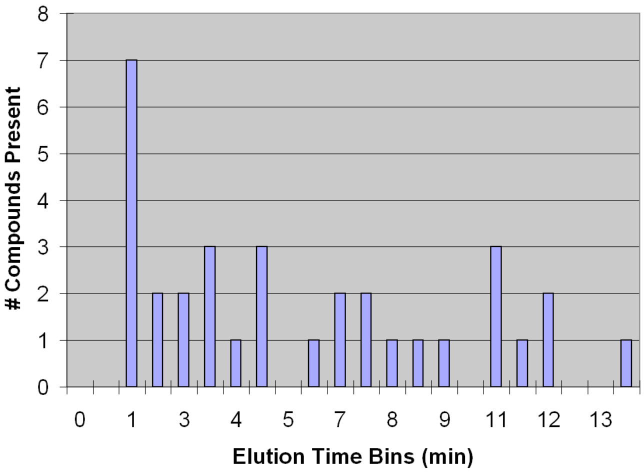


Figure 7. Number of compounds with near elution times.

8. CONSOLIDATION OF DETECTIONS

A typical run such as is shown in Figure 3 can generate several identifications for each visible peak in the TIC. Table 2 shows a complete list of detections for the example run including multiple detections, while Figure 8 shows the same results graphically. One of these peaks originates from TIC analysis and the other from mass channel components. One simple algorithm to filter these results for human consumption might examine overlapping windows and take only the highest Net M_f for identical compounds, which in this case would clean the list completely of redundant detections. Similarly, reporting only those detections with the largest extracted counts (area beneath peak) would tend to dismiss smaller fragmented detections due to noise. More troublesome situations may arise, however, in which misidentifications are mingled with correct identification at various match factors. While it is tempting to produce algorithms to try and detangle such situations, it was decided for the sake of completeness to downlink the raw output of JPL AMDIS to the ground for human inspection. Considering that we will likely never pursue more than a handful of compounds in any one run, it is not onerous to require a human to consider multiple returns.

As the VCAM instrument has varying degrees of sensitivity to the contents of our library, the current VCAM hardware implementation operates at two gain settings (ionization current). Each run includes a low and high gain simultaneous measurement. Some compounds may be seen in both high and low gain equally well, while others may be saturated in high gain or have negligible signal in low gain. Such a system offers another consolidation challenge, for false detections in the high gain are numerous due to saturation events. One proposed method would restrict certain compound identifications and quantifications to only be valid in one gain setting, low or high, or to require successful detection be found in both. However, these consolidation methods are not currently scheduled to be implemented for launch.

The final step of JPL AMDIS feeds results such as those in Table 2 into a post-processing algorithm [11] that computes concentration based on integrated TIC counts given a correct identification of the compound and the window in time over which it elutes. This logic is an extension beyond the NIST AMDIS process and is discussed thoroughly in Lee's accompanying paper.

Table 2. Example Run Output

Compound Detected	Elution Time	Width	Domnt Peak (AMU)	M_f Net
perfluoropropane	28	6	TIC	58.7
perfluoropropane	28	5	102	26.1
freon 11	34	5	TIC	67.4
freon 11	34	5	102	67.2
acetone	63	9	TIC	79.1
2-butanone	100	8	TIC	78.4
2-propanol	125	5	45	55.3
2-propanol	128	7	45	51.8
toluene	227	6	TIC	83.9
1,2-dichloroethane	256	5	TIC	74.7
1-butanol	322	6	56	82.3

9. Results

As of this paper, the results for the final validation and verification of the VCAM instrument are not available. However, results from a preliminary validation set have been analyzed for the purposes of documenting the JPL AMDIS code performance. Results listed here are not intended to represent the eventual capability of the VCAM instrument or its performance aboard the ISS, as substantial machine tuning will have occurred between these results and the final launch system.

Five physical bags containing mixtures of the 33 compounds specified in Section 2 were prepared and analyzed at four concentrations with three repetitions each for statistical sampling. A calibration run using fluorobenzene, acetone, perfluoropropane, and air [7] was performed before each set of 12 runs for an individual bag to ensure proper mass calibration. Figure 9 shows our preliminary identification results. For the experimental run reported here, 1,2-propylene glycol, carbonyl sulfide, and hexamethylcyclotrisiloxane were not available for testing and are shown in grey. The xylenes (m, p, and o) were collapsed to a single xylene entry due to identical mass spectral fragmentation patterns differing only by slight relative heights.

On the whole, our identification averaged 90% correct. Of the errors, 30% were due to isoprene, 12% respectively to octamethylcyclotetrasiloxane (OMCTS) and hexane, and 9% to pentanal with the remainder scattered across the other compounds. We will briefly document the challenging compounds now as examples of failure modes.

For isoprene, it was later shown a coelution with hexane at 1/100th the signal intensity of hexane made it invisible on the TIC trace and extremely difficult to detect in the mass channel regime. This lead not only to misidentification as hexane but missed events, subsumed by the larger peak. Later adjustments were made to enhance VCAM's sensitivity to isoprene.

OMCTS was sometimes misidentified as toluene due to the very high mass of its parent peak (281 AMU). Any initial mass calibration error is magnified via extrapolation from the highest calibration mass of 169 AMU (perfluoropropane) resulting in a poor fit. As only toluene, chloroform, or perhaps4-methyl-2-pentanone have elution times overlapping OMCTS, toluene is the closest next match and can compete with OMCTS should its parent peak be very shifted. The AMDIS algorithm is currently being modified to include a mass folding capability to desensitize results from mass shifts at such very high masses.

Hexane was misidentified as pentane twice and isoprene twice. The isoprene misidentification is due to the same observation as for isoprene's difficulties: a real isoprene component was present at very low signal, superimposed on hexane nearly perfectly. Given that their elution windows overlap substantially, it becomes a competition between their respective match factors as to which will be reported. The pentane misidentification results from the nearly identical spectra of pentane and hexane as well as their overlapping elution windows. In some high-concentration samples, pentane can produce additional peaks in its fragmentation pattern above its parent peak and resemble hexane through this saturation mechanism. The difficulty separating pentane and hexane may be aided by hand-trimming of the elution windows of acceptance for the final device tuning.

10. Conclusions and Future Work

The JPL AMDIS method is a robust algorithm now packaged as part of the VCAM instrument to fly aboard the ISS to determine atmospheric constituents. The ability to self-calibrate mass scale, filter results by absolute elution time, and calculate atmospheric concentration in ppm was added to the NIST AMDIS algorithm to support flight requirements as well as tuning the algorithm parameters to VCAM data requirements. All 33 required compounds were successfully identified across a range of concentrations and mixtures.

There are many extensions of the current algorithm that should be noted for future application. Some are easy to implement while others will require substantial statistical research.

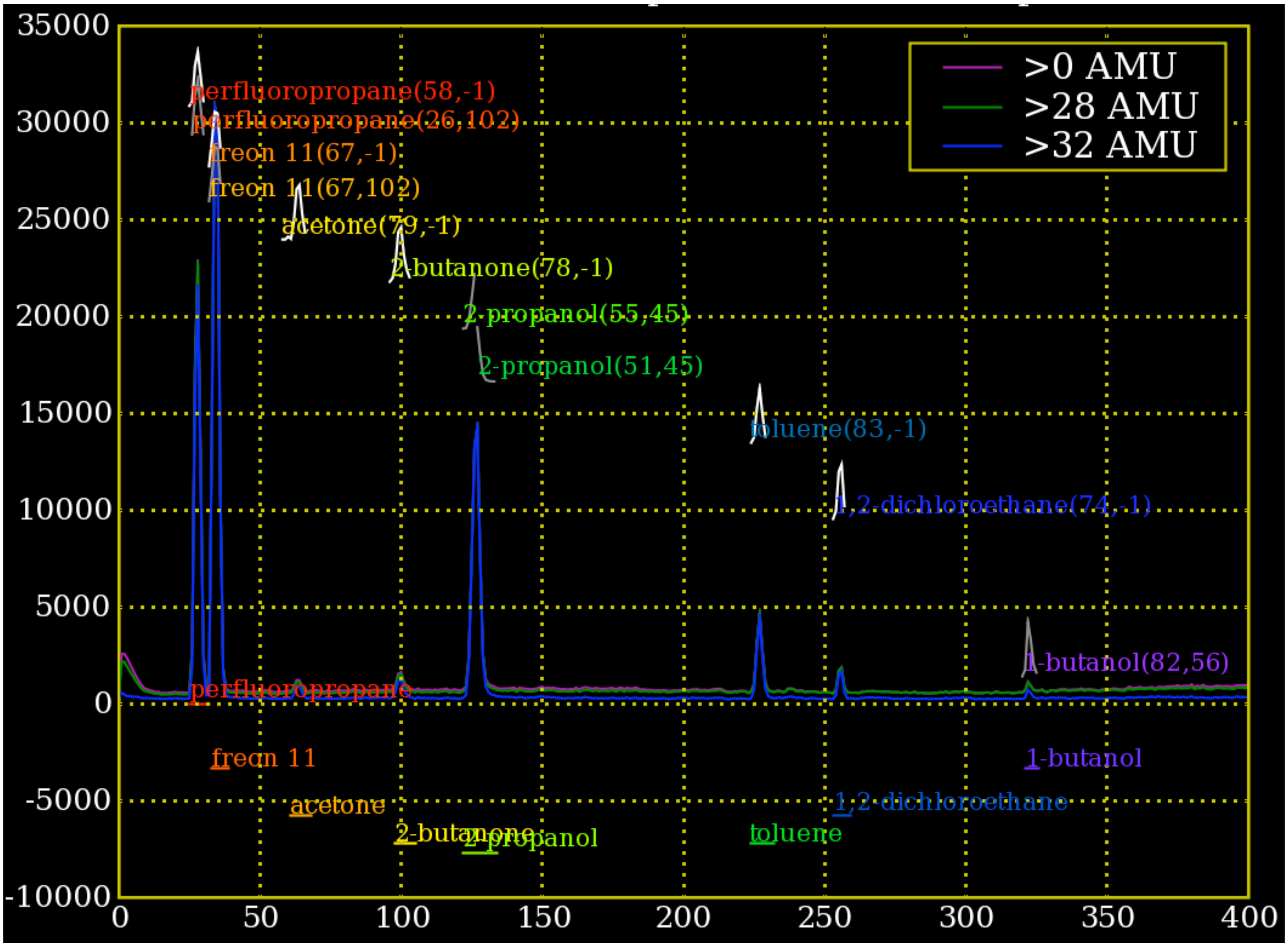


Figure 8. Results from a JPL AMDIS run. Detections are listed above Y=0, while known, injected compounds are shown below Y=0. Note the multiple hits on each peak.

Library Preprocessing

Currently, a mass squared weighting term assists the weighting of higher mass observations as more uniquely identifying; however, this is a gross approximation for any general library of interest that may or may not represent any given library. A pre-processing step could examine the current target library extracting useful information on peaks that are unique to certain compounds or uselessly common to all (see Figure 1). This information could then be fed as a series of weights to the compound identification code to assist its selection in a natural analogy to how human operators often make their determination e.g. "I see mass 169, it must be perfluoropropane." In our case, any substantial signal in mass channels above 150 could be taken as irrefutable proof of small sets or even uniquely identified compounds without additional computation, while a n-order mass term (empirically fit and likely much larger than second order) would be more appropriate for the range 20 to 150. Further, this process could be powerfully used to tease out coelutions between compounds in which only one has very high mass components, as these could be used exclusively in the model-building process improving the peak clustering / compound spectral building process.

Simultaneous Compound Identification

The current system identifies each compound individually without concern for the presence of any other compounds found in the run so long as their elution times are within a window of acceptance. This process could be instead placed into a simultaneous matrix that could benefit from known ordering information on a much finer resolution than simple elution time windows. For examine, hexane and pentane have well defined, overlapping windows of acceptance, but empirical runs have shown that should both be present in a run they always occur in the order pentane then hexane. Human operators commonly use such ordering intuition during difficult identifications. Such a matrix could be formulated in a least-squares manner analogous to the model fitting step, replacing the current highly empirical algorithm of peak clustering, peak flagging, initial screening, and match factor calculation. Substantial improvement in reduction of redundant output and lowscore misidentification could be thus obtained without the need for further ad-hoc filtration based on match factors.

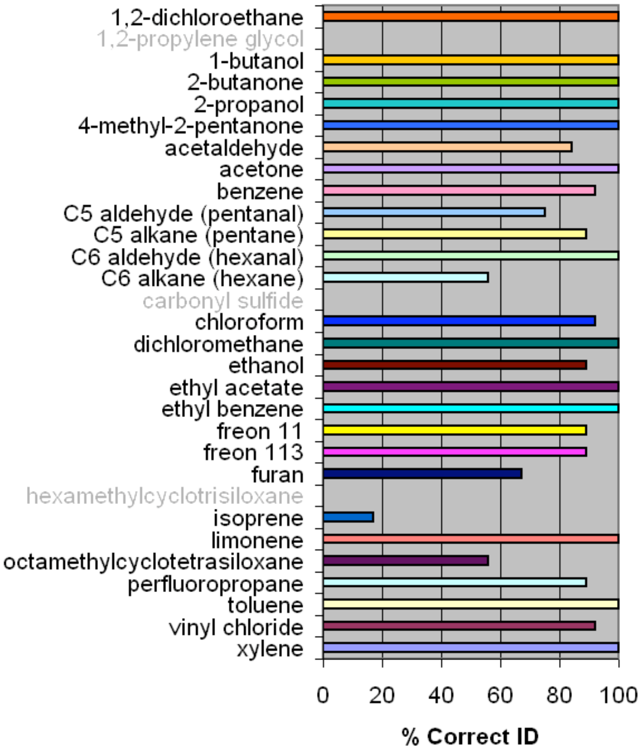


Figure 9. Percent correct ID for library species. Xylenes have been collapsed to single entry. Greyed out compounds were not present.

Bayesian Approach

Perhaps most powerfully, the addition of Bayesian theory could determine an actual probability distribution of identity over the target library (instead of an ad-hoc match factor of unknown physical interpretation) to help determine more useful output in truly ambiguous cases in which two solutions are nearly equally likely. The current NIST AMDIS implementation will simply randomly choose between two equally valid solutions. Large signals which do not well correspond to any library compound could also be thus recognized and treated specially. Such behavior would be very advantageous to a truly autonomous system that must prepare for unknown future events and still gracefully report useful observations.

11. 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. Copyright 2009. All rights reserved.

References

[1] J. Perry and M. LeVan, "Air Purification in Closed Environments: Overview of Spacecraft Systems," NBC Defense Collective Protection Conference, Orlando FL, Oct 29-31, 2002.

[2] K. Persaud et al, "A Smart Gas Sensor for Monitoring Environmental Changes in Closed Systems: Results from the MIR Space Station," Sensors and Actuators B, vol 55, pp. 118-126, 1999.

[3] D. L. Jan, "Environmental Monitoring Instruments: Using ISS as a Testbed for Exploration," IEEE Aerospace Conference, pp. 1-7, March, 2007.

[4] 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.

[5] R. G. Dromey et al., "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.

[6] S. E. Stein, "An Integrated Method for Spectrum Extraction and Compound Identification from GC/MS Data," Journal of the American Society for Mass Spectrometry, Vol. 10, Issue 8, pp. 770-781, August 1999.

[7]S. Lee et al, "Autonomous Calibration of Vehicle Cabin Atmosphere Monitor," 2008 IEEE Aerospace Conference, pp. 1-8, doi: 10.1109/AERO.2008.4526524, March 2008.

[8] S. R. Heller, "The History of the NIST/EPA/NIH Mass Spectral Database," Today's Chemist at Work, Volume 8(2), pp. 45-46:49-50, February 1999.

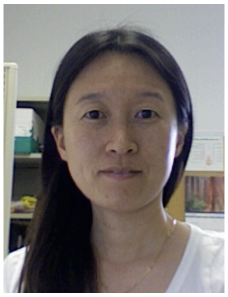
[9] B. N. Colby, "Spectral Deconvolution for Overlapping GC/MS Components," Journal of the American Society for Mass Spectrometry, Vol. 3, pp. 558-562, 1992.

[10] E. Kovats, "Characterization of Organic Compounds by Gas Chromatography. Part 1. Retention, Indices of Aliphatic Halides, Alcohols, Aldehydes, and Ketones," Helvetica Chimica Acta, vol 41, pp. 1915-32, 1958.

[11] S. Lee et al, "Quantification of Trace Chemicals using Vehicle Cabin Atmosphere Monitor," 2009 IEEE Aerospace Conference, March 2009

BIOGRAPHY

Dr. Lukas Mandrake *is a member of the technical staff in 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 Earth-sensing satellite missions, autonomous spectroscopy, and autonomous sensor networks and is a key member of the VCAM development / data analysis team. Lukas particularly enjoys studying computational models of natural systems. Lukas received his Ph.D. and M.S. in computational plasma physics from UCLA in 2004 and his B.A. in engineering physics from the University of Arizona in 1995.*



Dr. 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 the Vehicle Cabin Atmosphere Monitor and conducting research on materials modeling and simulation, non-linear dynamics control, spectral retrieval, data reduction, global optimization, parallel computing, and advanced numerical algorithms. She received her Ph.D. and B.S. in Physics from the Seoul National University.*

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 (IGE) 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.*

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 a Bachelor's degree from Ausburg College in Computer Science and Mathematics. He is currently pursuing a Ph.D. in Electrical and Computer Engineering at Rice University..*