Leveraging Open Science Machine Learning Challenges for Data Constrained Planetary Mission Instruments

Victoria da Poian, Eric I Lyness, Jay Y Qi, Isha Shah, Greg Lipstein, P Doug Archer, Luoth Chou, Caroline Freissinet, Charles A Malespin, Amy C Mcadam, et al.

RAS Techniques and Instruments

RASTAI **3,** 156–165 (2024) https://doi.org/10.1093/rasti/rzae009
Advance Access publication 2024 March 15

Leveraging open science machine learning challenges for data constrained planetary mission instruments

Victoria Da Poian [1,2,3★] Eric I. Lyness,[1,2] Jay Y. Qi,[4] Isha Shah,[4] Greg Lipstein,[4] P. Doug Archer Jr.,[5] Luoth Chou,[1,6,7] Caroline Freissinet,[8] Charles A. Malespin,[1] Amy C. McAdam,[1] Christine A. Knudson,[1,6,9] Bethany P. Theiling[1] and Sarah M. Hörst[3]

[1]*NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*
[2]*Microtel LLC, Greenbelt, MD 20770, USA*
[3]*Johns Hopkins University, Earth and Planetary Science Department, Baltimore, MD 21218, USA*
[4]*DrivenData, Denver, CO 80206, USA*
[5]*Jacobs JETSII Contract at the NASA Johnson Space Center, Houston, TX 77058, USA*
[6]*Center for Research and Exploration in Space Science and Technology II (CRESST II), Greenbelt, MD 20771, USA*
[7]*University of Maryland Baltimore County, Baltimore, MD 21250, USA*
[8]*Laboratoire Atmospheres, Observations Spatiales (LATMOS), Guyancourt, 78280, France*
[9]*University of Maryland College Park, MD 20742, USA*

Accepted 2024 March 6. Received 2023 September 7; in original form 2023 June 7

ABSTRACT

We set up two open-science machine learning (ML) challenges focusing on building models to automatically analyse mass spectrometry (MS) data for Mars exploration. ML challenges provide an excellent way to engage a diverse set of experts with benchmark training data, explore a wide range of ML and data science approaches, and identify promising models based on empirical results, as well as to get independent external analyses to compare with those of the internal team. These two challenges were proof-of-concept projects to analyse the feasibility of combining data collected from different instruments in a single ML application. We selected MS data from (1) commercial instruments and (2) the Sample Analysis at Mars (an instrument suite that includes a mass spectrometer subsystem onboard the Curiosity rover) testbed. These challenges, organized with DrivenData, gathered more than 1150 unique participants from all over the world, and obtained more than 600 solutions contributing powerful models to the analysis of rock and soil samples relevant to planetary science using various MS data sets. These two challenges demonstrated the suitability and value of multiple ML approaches to classifying planetary analogue data sets from both commercial and flight-like instruments. We present the processes from the problem identification, challenge set-ups, and challenge results that gathered creative and diverse solutions from worldwide participants, in some cases with no backgrounds in MS. We also present the potential and limitations of these solutions for ML application in future planetary missions. Our longer term goal is to deploy these powerful methods onboard the spacecraft to autonomously guide space operations and reduce ground-in-the-loop reliance.

Key words: Machine Learning – Open Science Challenge – Planetary Mission Instruments – Transfer Learning – Data Science – Mars Exploration.

1. INTRODUCTION

1.1 Challenges of planetary science data

Many planetary exploration missions aim at evaluating the habitability and the existence of potential life on the target bodies. Missions exploring further away in our Solar system (e.g. Titan, Europa, Enceladus, Ceres, etc.) or shorter duration missions due to extreme environmental conditions (e.g. Venus and Mercury) will face communication constraints due to limited transfer rates and short communication windows.

Current space missions operations processes are centred around ground-in-the-loop activities: data are sent back to scientists' teams on Earth for analysis and decision-making for future operations and desired analyses to run. The analysis time during which scientists collect the scientific data, analyse it, infer the information contained in it, and decide which next operations should be run on the spacecraft is often extremely limited (e.g. a few hours for the SAM instrument on Curiosity, 24–48 h for the mass spectrometer instrument onboard the ExoMars mission). Further, ground-in-the-loop operations require significant planning and coordination, restraining the missions' flexibility (Thompson et al. 2012). Our open-science challenge leverages machine learning (ML) and data science techniques to improve methods for analysing mass spectrometry (MS) data to support scientists' decision-making process during missions operations

★ E-mail: victoria.dapoian@nasa.gov

(Da Poian et al. 2022). The goal is to make science operations faster and more efficient, and ultimately maximize missions' scientific returns, especially for missions to the outer Solar system (Theiling et al. 2021) that will face more severe communication and resource limitations than missions to Mars.

One of the main challenges of applying ML techniques to planetary mission instruments is the limitation of the available training data sets. It is essential to understand that planetary mission instruments are unique due to their development and design in-house. Indeed, planetary instruments are often built for a specific mission, each with their own set of unique requirements and scientific goals. Therefore, each instrument onboard a planetary mission is precisely tailored to the specific mission's target and scientific objectives, while being constrained by mission requirements. During the development of space instruments, scientists and engineers use commercial instruments and develop testbed instruments (e.g. flight analogue instrument) to optimize and understand flight models (FMs). Testbed instruments are essential to (1) simulate space conditions, (2) test the instrument functionality (for engineering, science, and operations purposes), (3) calibrate the instrument, and to (4) start collecting data similar to the ones that will be collected during the mission. Because the development and the use of testbeds is time-consuming and resource-intensive, scientists often use commercial instruments in the early stage of the development. Commercial instruments off the shelf have lower fidelity but offer a higher accessibility and more experimental freedom (i.e. less restrictions about possible samples to analyse and methods), while FMs represent the ground truth but are used less frequently and with highly mission relevant samples to limit potential contamination and over use of hard to replace flight-like components. Our research investigates the combination of commercial instruments and flight-like instrument data sets in the development of a ML model to help analyse MS data for Mars exploration via open science challenges.

1.2 Opportunity grant for open science challenges

In 2020, the NASA's Science Mission Directorate (SMD) Strategic Data Management Working Group launched a call for challenge proposals entitled 'Using NASA Science Data and Computing for Cross-Disciplinary Science'. This call was looking for challenge proposals that could be turned into topics for significant prize-based challenges focusing on utilizing NASA's free and open science data from multiple science disciplines. The call was also looking for proposals encouraging collaboration across the various NASA science divisions. This call for proposals had up to $1M to put towards developing and executing these challenge proposals.

Bull, Slavitt & Lipstein (2016) describe the power of crowdsourcing to increase capacity for data science with open challenges where experts (and non-experts) from around the world can contribute and organizations can receive high-performing algorithms with empirically demonstrated results among different models explored during a competition. The intention of NASA SMD to organize open science ML challenges is to leverage the great benefits of this approach listed below and illustrated in Fig. 1:

(1) **Engagement of a large and global community**: Open science challenges make research more accessible and inclusive, attracting a broader range of researchers.

(2) **Fast progress**: Multiple researchers from different fields can develop different approaches and test hundreds/thousands of models quickly.

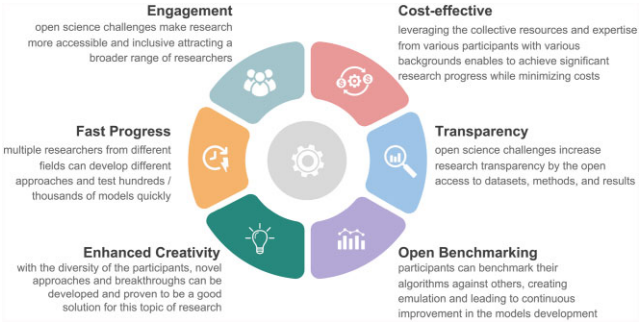

Figure 1. Some of the main benefits of organizing open science challenges for enhancing science and research topics.

(3) **Enhanced creativity**: The diversity of participants leads to diversity of perspectives and approaches to solve research problems. Novel approaches and breakthroughs can be developed and proven to be a good solution for this topic of research.

(4) **Cost-effective**: Leveraging the collective resources and expertise from various participants with various backgrounds enables to achieve significant research progress while reducing costs.

(5) **Transparency**: Open science challenges increase research transparency by the open access to data sets, methods, and results.

(6) **Open benchmarking**: Open science challenges offer a standardized evaluation process so that participants can benchmark their algorithms and models against others, creating emulation and leading to continuous improvement in the models development.

This research benefited from sharing of resources and data, and the cost-effectiveness of open science challenges to develop novel collaboration practices, and ultimately accelerated research progress on the use of commercial data sets for flight-like instrument implementation while fostering an open scientific community. While open science challenges have been used in the field of astrophysics (Dieleman, Willett, Damber 2015; Hložek 2020), this is quite a new approach in the field of planetary science. Via the two open science ML challenges described in this paper, we investigated the potential of ML techniques on MS data for planetary science analyses. The first challenge that used Evolved Gas Analysis-Mass Spectrometry (EGA-MS) data mainly investigated the potential for transfer learning between commercial and flight-like instruments and showed that transferability between these instruments data sets exist and could be leveraged to train ML algorithms for planetary science missions. These initial results on MS data could also be applied to other planetary science instruments that generate spectral data, such as Raman spectroscopy and infrared spectroscopy.

2. CHALLENGE ORGANIZATION

2.1 Proposal for planetary science data

Mass spectrometers have been deployed onboard space missions since the 1970s (Nier & McElroy 1977; Niemann et al. 1996). A mass spectrometer is an analytical instrument measuring the mass-to-charge (m/z) ratio of ionized particles in a sample. The three main components of any mass spectrometer are the ion source, mass analyzer, and detector (Fig. 2). The **ion source** generates ions from the neutral analytes. The ionization process can be done via various methods (e.g. electron impact, electrospray ionization, chemical ionization, etc.). These ions are then accelerated and focused into a beam by an electric field or a magnetic field. The

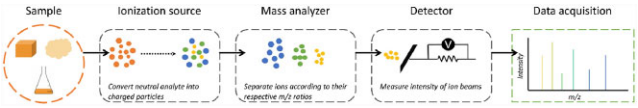

Figure 2. Representation of the three main components of mass spectrometer instruments (ionization source, mass analyzer, and mass detector) (adapted from Arevalo et al. 2020). The modularity of each subsystem allows various mass spectrometer designs for targeted applications.

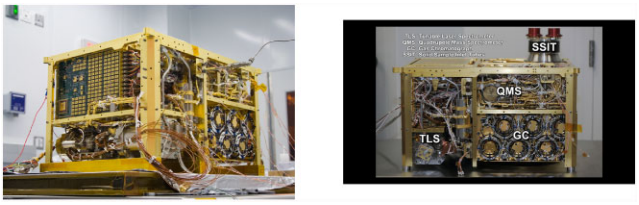

Figure 3. SAM mass spectrometer instrument onboard the Curiosity rover, operating on Mars since 2012 (credit: NASA).

mass analyzer separates the ions based on their mass-to-charge ratio (m/z). The most common tools for this separation are quadrupole, time-of-flight, magnetic sectors, linear ion trap, and Orbitrap^{TM}, etc. The **detector** counts the separated ions or measures the charges electrically, sometimes amplifying the signal to measure the ion's relative abundance.

MS and associated subsystems are used in various disciplines such as medical science, chemistry, biology, environmental science, and planetary science. For planetary science applications, the main uses of mass spectrometers and associated devices include: (1) determination of the composition of planetary surfaces and atmospheres in order to understand the origin and evolution of planetary bodies, the dynamics of atmospheric processes and climate on planetary bodies, as well as past and present habitability; (2) identification of inorganic molecules from surface samples, to determine the mineralogy of an environment and its evolution over geological times; (3) identification of organic molecules from surface samples that could be building blocks for life as we know it on Earth; and (4) measurement of isotopic ratios that can provide information about the formation and evolution of planets as well as the history of geological and atmospheric processes on a planetary body. MS is a high-heritage technique and powerful tool for planetary missions (Chou et al. 2021) as it provides valuable insights about the composition, the evolution, the processes of our Solar system, and the potential for life beyond Earth.

The Mars Science Laboratory (MSL) mission, launched in 2011 and landed on Mars in 2012 in Gale Crater, aims at investigating the potential habitability of Mars by studying Mars' geology, climate, and organic matter distribution at Mars' surface and subsurface (Grotzinger et al. 2012). The MSL mission consists of the Curiosity rover equipped with 10 powerful instruments including the Sample Analysis at Mars (SAM) instrument suite (Mahaffy et al. 2012, Fig. 3). SAM is designed to analyse the chemical and isotopic composition of samples of Martian rocks, soil, and atmosphere on Mars. It is composed of three main subsystems: (1) a gas chromatograph (GC), to separate gases prior to their identification in the MS (Gas Chromatography Mass Spectrometry, GCMS), (2) a mass spectrometer, to detect and identify the key molecules necessary for life (i.e. containing carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur, commonly known as CHNOPS), and (3) a tunable laser spectrometer, to detect light gases (such as CO_2 or CH_4 that could have been produced by life or geological processes), and investigate their isotopic composition.

Because of the high heritage of mass spectrometers for space missions, many future missions looking for chemistry or biology of a given planetary body, orbiting it or landing on it, will be equipped with mass spectrometers, particularly as part of a suite in combination with a sample preprocessing component (e.g. the Sample Manipulation System on MSL) and separation system (e.g. GC, etc.). For example, Mars Organic Molecule Analyzer (MOMA) will be onboard the ExoMars 2028 mission (Rosalind Franklin rover), a joint mission between the European Space Agency and NASA (Goesmann et al. 2017). MOMA will notably be used to search for traces of past life on Mars by analysing samples collected up to 2 metres below the surface. In the next decade, the Dragonfly Mass Spectrometer (DraMS) will be onboard the Dragonfly mission, targeting arrival at Titan by 2034 (Grubisic et al. 2021). DraMS will address the question of the complex chemistry and potential habitability of Titan's surface. The Mass Spectrometer for Planetary Exploration is developed for the Europa Clipper mission, targeting Europa for a launch in 2024 (Brockwell et al. 2016). It will be used to analyse the composition of Europa's icy surface and subsurface in order to understand the moon's potential habitability. For small-body investigations, the Laser Ablation Mass Spectrometer will be onboard the Psyche mission, set to be launched in October 2023, to analyse the composition of the metallic asteroid Psyche to understand its origin and evolution (Hart et al. 2018). Finally, mission concepts such as the Enceladus Orbilander propose using a High-Resolution Mass Spectrometer to investigate the origin and habitability of Enceladus as well as look for signatures of life on this moon (MacKenzie et al. 2021).

Planetary missions are highly constrained in communication links and data downlinks. As instrument complexity and resolution increases, it will be increasingly challenging to downlink the full output data files of scientific instruments, especially when going further in the outer Solar system (e.g. Titan, Europa). Algorithms and methods capable of reading and interpreting the output of these scientific instruments with high confidence onboard the spacecraft will not only enable the missions to collect more data, but will also benefit the data prioritization process to send back to Earth the most interesting and promising results first. Artificial Intelligence and ML techniques could greatly benefit the development of such algorithms. ML applications require large data sets that planetary science instruments do not often have. In particular, NASA-built science instruments are highly customized and have constraints in the list of samples that can be tested, often leading to small data sets and a variety of analytical parameters that can prevent ML applications. However, during the development of space missions, laboratory equivalents of instrumental subsystems as well as commercial instruments serve a critical purpose by enabling larger data sets to be collected.

Our challenge's proposal focused on finding some alternate approaches to assisting in the interpretation of space missions' instrument data without having large amounts of data with which to train ML algorithms. We sought innovative methods (e.g. transfer learning or other novel approaches) to help analyse and interpret the output measurements of planetary mission instruments constrained by limited data sets (restraining the use of ML algorithms). Specifically, the two challenges presented in this paper aimed at (1) evaluating how ML could be applied to MS analysis, with a specific focus on Mars planetary mineralogy, geochemistry, and chemistry, and at (2) evaluating how well models trained on commercial instruments would perform on SAM-like data. Challenge 1 used EGA-MS data, while challenge 2 focused on the GCMS data to develop methods

to accurately interpret the chemical composition of material samples analysed by the SAM instrument on the Curiosity rover.

Our ultimate goal was to use 'transfer learning' to leverage the knowledge gained from commercial instruments used to develop planetary science missions (Wong & Michaels 2022). Transfer learning relies on developed algorithms trained on large data sets from similar instruments to then tune those algorithms to adapt them to planetary instruments with limited training data sets. For example, commercial mass spectrometers and flight-like models such as an Engineering Test Unit or Testbed model are instruments that collect spectra and have many tunable parameters that can produce different results with the same sample. More flight-like models are often reserved for mission-related activities and commercial instruments are used to collect large amounts of data from relevant samples (e.g. laboratory simulants or Earth-based analogues). The larger data set from commercial instruments could be used to train ML models, and then, the learned algorithms can be adjusted for very limited flight instrument data sets, ultimately benefiting missions' preparation and operations. To summarize our task in one question: Can we train ML on data sets from commercial laboratory instruments and then successfully apply those ML models to science data from FM instruments?

2.2 Process set-up

The organization of these challenges was conducted with various stakeholders. First, NASA was the project organizer and the sponsor with the principal investigator's team from the NASA Goddard Space Flight Center (GSFC). Second, the NASA Center of Excellence for Collaborative Innovation (CoECI), established in 2011 at the request of the White House Office of Science and Technology Policy, was the project coordinator. CoECI aims at collaborating with innovators across NASA and the Federal Government to generate ideas and solve important problems by working with global communities via the NASA Tournament Lab (NTL). NTL offers a variety of open innovation platforms that engage the crowdsourcing community in challenges to create the most innovative, efficient, and optimal solutions for specific, real-world challenges faced by NASA. These challenges were designed and hosted by DrivenData, a company focused on the organization of online ML challenges for projects at the intersection of data science and social impacts in various areas such as international development, health, education, research, and public services. Finally, HeroX supported the communication around the challenges and the publication of press releases.

Both challenges used MS data collected for Mars exploration missions. The first challenge focused on EGA-MS (see Section 3.2) data with data sets coming from laboratory instruments at NASA's GSFC and Johnson Space Center (JSC) that are affiliated with the SAM instrument science team. The data sets were collected from (1) commercial instruments: commercially manufactured instruments that have been configured to be used in SAM-like conditions at GSFC (Franz et al. 2020) and JSC (Archer et al. 2013; Clark et al. 2019), and (2) the SAM testbed at GSFC, a high-fidelity replica of the SAM instrument suite, operating in a Mars chamber (under Martian temperature and pressure conditions). It is worth mentioning that differences between commercial instruments and the SAM testbed lead to additional difficulties in preprocessing the non-uniform data sets. For instance, commercial instruments measure ion abundance as ion current in amperes (amps, Coulombs per second), while the SAM testbed measures abundance as counts per second. Another major difference includes a higher time resolution in commercial mass spectrometers, due to a lower scanning rate in the testbed instrument. In order to deal with these differences, some data processing steps and calibration needed to be applied to enable the comparison of commercial instrument data and testbed data. The second challenge used GCMS (see Section 3.3) data only from GSFC commercial instruments. In future challenges, we envision using actual Mars data collected by the SAM instrument thanks to NASA's Planetary Data System (PDS; https://pds.nasa.gov/) that archives and distributes publicly available digital data related to the study of surfaces and interiors of terrestrial planetary bodies.

3. CHALLENGES SET-UP

These two ML open science challenges are multilabel classification tasks. A multilabel classification problem is a type of supervised learning problem where each input data (in our case, mass spectra) is labelled with multiple classes (also called labels). Each mass spectrum can belong to zero or more classes (in our case, chemical families) rather than just a single class. The model outputs a probability distribution for each class, with probability scores between 0 and 1 that indicate the likelihood of that label being present for the given input data. Several performance metrics can be used to evaluate the performance of the ML models. For these two challenges, we used the logistic loss and the average precision as they encompass other performance indicators and combine them to assess the overall model's performance.

3.1 Performance metrics

Open science challenges rely on the use of various performance metrics to evaluate and compare the effectiveness of different research approaches and methodologies. For these two challenges, we use the log loss metric and the average precision metric.

3.1.1 Log loss metric

The logistic loss (also called cross-entropy loss) is a commonly used loss function in ML and statistics models, especially for classification tasks. The log loss calculates the difference between the predicted probabilities and the true labels. For a single observation, the log loss is expressed as follows:

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$$

with y, a binary variable indicating if the label is correct (0 or 1), and p, the predicted probability that the label is present.

The logistic loss aims at penalizing the model when it is confident (e.g. predicts high probability) for incorrect predictions. The log loss is a reliable and widely used metric for the evaluation of classification models as it provides a better measure of performance to incorrect predictions and is more sensitive to differences in predicted probabilities between classes. This is even more important in application to planetary science, as the misclassification of one class might be more important than another one. Lower log loss scores indicate better performance of the model. For these challenges, the metric is the average across label classes of the binary log losses for each class.

3.1.2 Average precision metric

The average precision is calculated as the weighted mean of precisions at each threshold. The precision measures how well the algorithm finds true positives (TPs) out of all the positive predictions

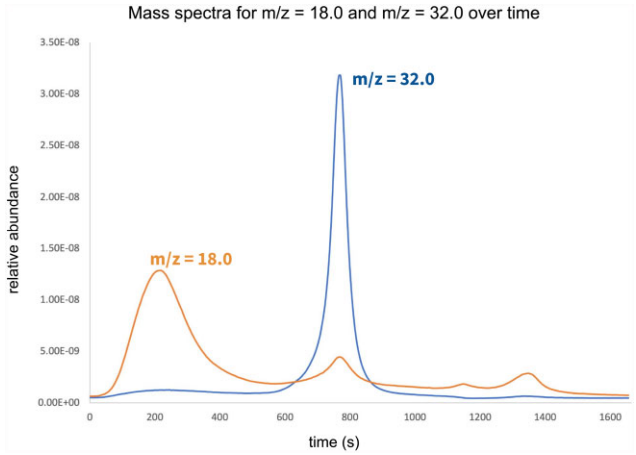

Figure 4. Example of an EGA mass spectra focusing on masses $m/z = 18.0$ and $m/z = 32.0$.

(TPs and false positives). This performance metric aims at rewarding the algorithm that assigns positive samples with higher scores than negative samples. It ranges from 0.0 (completely wrong predictions) to 1.0 (perfect predictions). The average precision metrics reported for these challenges are micro-averaged across label classes-each label's prediction for each observation is treated as an observation in a global precision-recall calculation.

It is worth noting that many other performance metrics exist for classification tasks. Log loss and average precision provide concise and comprehensive summaries of overall model performance across operating thresholds. Log loss considers predictions with a probabilistic interpretation and rewards models that are statistically well-calibrated. Average precision measures the quality of prediction scores' rank ordering. Models with strong performance will generally score well on both of these metrics, but they are not inherently correlated with one another.

3.2 Evolved gas analysis-mass spectrometry

The Evolved Gas Analysis (EGA) mode in SAM (Mahaffy et al. 2012; Sutter et al. 2017; McAdam et al. 2022) involves heating a solid sample at a rate of 35°C min^{-1} from ambient to 850°C under a He flow and measuring in real-time the quantity of released gases using a mass spectrometer. The temperature at which specific gases are released provide information about the sample's mineralogy and geochemistry. The EGA-MS's measurements are time series that scientists study to identify the gases produced by the sample over time while being heated. Scientists' expertise and domain knowledge enable the determination of the chemical and mineralogical composition of the studied sample. Fig. 4 illustrates an example of (1) a sample ion abundance plotted over time and (2) the temperature profile the sample was heated at. A specific volatile compound will produce a series of fragments that are recorded as m/z by the MS. As a simplified example, sulfate minerals will decompose at temperatures above 600°C and release SO_2, a gas that is characterized with m/z 64, 48, and 32 (among others). The detection of those concurrent m/z at the same high temperature thus determines the presence of sulfate in the sample. The specific type of sulfate in a sample can then be constrained by using the temperature of evolution of the SO_2 peak.

The challenge data set was shared as .csv file format. Each input data in the challenge data set represented the study of a physical sample. The features for each sample are the EGA-MS measurements containing four dimensions:

(i) **Time**: The time in seconds since the start of the reference time (e.g. the start of sample heating),
(ii) **Temp**: The temperature of the sample in °C at the time of the measurement,
(iii) ***m/z***: The mass-to-charge ratio of the measured ion, and
(iv) **Abundance**: The count or current of ions being detected per scan (note: abundance values are compared in a relative way within each sample's analysis).

For the EGA-MS challenge, competitors were asked to predict the probability that each of the classes described in Table 1 was present in the sample. These classes represent certain families of mineralogies that are of scientific interest in analysing conditions for the history of Mars and its past habitability. Details about this EGA-MS challenge are available on DrivenData website: https://www.drivendata.org/competitions/93/nasa-mars-spectrometry/page/437/). Each sample can have multiple class assignments or can have none. In the labels file (as shown in Table 2), a '1' indicates that the studied sample

Table 1. Summary table of the 10 labels used in the EGA-MS challenge to represent main families of minerals of interest for the study of Mars.

Label	Brief geochemical description
Basalt	Extrusive igneous (volcanic) rock, low in silica (Si) content, dark in colour, comparatively rich in iron (Fe) and magnesium (Mg)
Carbonate	A salt that contains CO_3^{2-} and a cation, generally Fe^{2+}, Ca^{2+}, or Mg^{2+}
Chloride	A salt that contains the Cl^- anion
Iron_oxide	Chemical compounds composed of iron (Fe) and oxygen (O)
Oxalate	Minerals containing the $C_2O_4^{2-}$ anion
Oxychlorine	Oxidizing chlorine-containing salts of general composition ClOx, that includes the widespread Martian perchlorates (ClO^{4-})
Phyllosilicate	Compounds with structures containing tetrahedral s sheets (silica tetrahedrons consisting of a central silicon atom surrounded by four oxygen atoms) and octahedral sheets (arrangements of OH^- and cations), commonly called clay minerals
Silicate	Minerals containing polyatomic anions consisting of silicon and oxygen (e.g. SiO_4^{2-})
Sulfate	A salt containing SO_4^{2-} and cations such as Fe^{2+}, Ca^{2+}, or Mg^{2+}
Sulfide	A compound containing one or more S^{2-} ions

Table 2. Example of the label file for each sample. '1' indicates the presence of the mineral phase in the studied sample, while '0' indicates otherwise.

Sample_id	S000	S001	S002	S003	S004	...
Basalt	0	0	0	0	0	...
Carbonate	0	1	0	1	0	...
Chloride	0	0	0	0	0	...
Iron_oxide	0	0	0	1	1	...
Oxalate	0	0	0	0	0	...
Oxychlorine	0	0	1	0	1	...
Phyllosilicate	0	0	0	0	1	...
Silicate	0	0	0	0	0	...
Sulfate	1	0	0	1	0	...
Sulfide	0	0	0	0	0	...

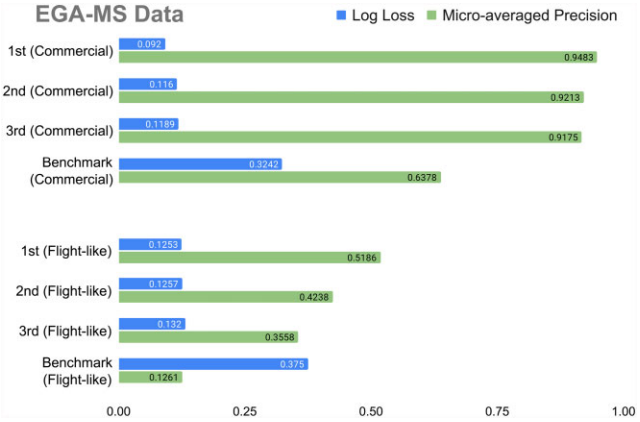

Figure 5. EGA-MS results for the top three winners and the benchmark. Comparison between the models results on the commercial data and on the SAM testbed data. Log loss in blue (the lower the better), and micro-averaged precision in green (the higher the better). We can note that the top three winners' solutions outperform the benchmark model for commercial and testbed data, and that the task on testbed data is more difficult but proposed solutions also outperformed the benchmark model. The top three winners' solutions are further detailed in Section 4.

contains a mineral phase from that specific family, while a '0' indicates otherwise.

In order to better understand the possibility of transferability to flight-like instruments, we organized a bonus prize specifically dedicated to the performance on the SAM testbed data. Because SAM testbed samples were limited to only 76 inputs, 12 samples were used in the training set and 64 samples in the test set as we were looking to emphasize this evaluation. The task of making correct predictions for the SAM testbed data were clearly more difficult than for the overall data set. The two main performance metrics of log-loss score and overall micro-averaged average precision (described in Section 3.3) were respectively higher and lower than for the overall test set as shown in Fig. 5. It is essential to note that log-loss scores cannot be directly compared across different data sets (between commercial and testbed predictions for instance, or even between different label classes). Nevertheless, the difference from the top three solutions with the basic benchmark provided to the participants proves that some transferability can be applied to data from commercial instruments to flight-like instruments. For both challenges, the original benchmarks we provided the participants contain basic exploration data analysis steps, preprocessing, and model development. The EDA step aims at better understanding the proportion of samples in each training, validation, and testing set for each instrument type (commercial versus flight-like), and some main features of MS data. The preprocessing step standardizes the mass values, removes background noise, and converts abundances to relative abundances. The benchmark model uses a simple modelling approach as 'one versus all' for this multilabel classification: binary classifiers using logistic regression are developed for each label class independently. The winners' solutions are described in more detail in Section 4.1.

3.3 Gas chromatography mass spectrometry

GCMS is an analytical method used to determine the molecular composition of samples. The SAM GCMS experiment involves heating a solid sample up to 850°C in a pyrolysis oven in order to vaporize the samples, directing compounds that volatilized over a

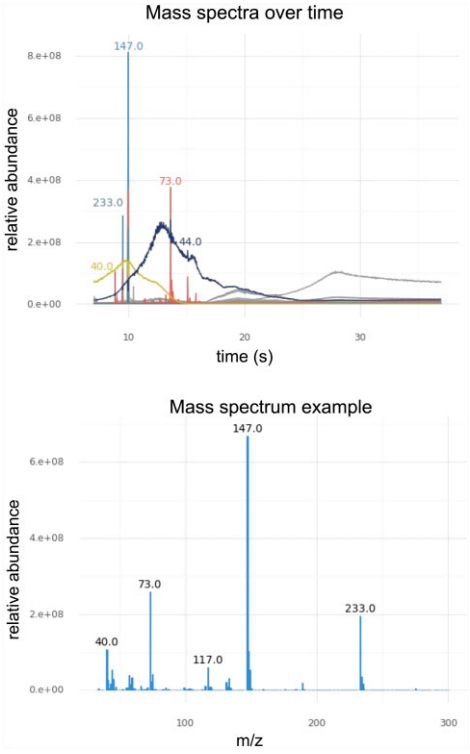

Figure 6. Example of a GCMS experiment output. The top plot represents the 'ion chromatograms' which show the intensities over time for ions by their individual mass. Ion chromatograms of *m/z* 40.0, 44.0, 73.0, 147.0, and 233.0 are highlighted. The mass spectrum example (bottom panel) represents the fragmentation peaks for a compound of this sample.

chosen temperature range during the heating ramp into GC capillary columns for their separation, and at the outlet of the columns, analysing the discrete compounds with the MS. A derivatization agent (a chemical reagent) can be added to the sample to aid in the vaporization of compounds with low volatility such as amino acids. The role of the GC column is to separate the chemical species released from the sample into their individual components. Components are released from the column at different times based on their chemical and physical properties. The time at which the compound is released from the GC column is the compound's retention time (Fig. 6). Different GC columns (composed of a stationary phase), carrier gas (a mobile phase such as helium), and the GC oven temperature programs will result in different retention times for the same analyte. Thus, the retention time of a given compound on a given column will be the same under the same analytical conditions. Once the component leaves the GC column, it is guided to and through the mass spectrometer in order to be identified. The outputs of GCMS experiments contain a chromatogram (representing the abundance, through one mass ion (*m/z*) or the sum of all selected mass ions, over time) and for each recorded time a mass spectrum is generated (Fig. 6).

The challenge data set was shared as .csv file format. Each input data in the challenge data set represented the study of a physical sample. The features for each sample are the GCMS measurements containing three dimensions:

(i) **Time**: The time in seconds since the start of the reference time,
(ii) ***m/z***: The mass-to-charge ratio of the measured ion at a defined retention time, and

Table 3. Summary table of the nine labels used in the GCMS challenge to represent main families of organic compounds and minerals of interest for the study of Mars.

Label	Brief chemical description
Aromatic	Any of a large class of unsaturated chemical compounds characterized by one or more planar rings of atoms joined by covalent bonds of two different kinds
Hydrocarbon	Any of a class of aliphatic (e.g. non-aromatic) linear or branched organic chemical compounds composed only of the elements carbon (C) and hydrogen (H)
Carboxylic_acid	Any of a class of organic compounds in which a carbon (C) atom is bonded to an oxygen (O) atom by a double bond and to a hydroxyl group (−OH) by a single bond. Examples are fatty acids or amino acids
Nitrogen_bearing_compound	Samples with nitrogen (N)-containing compounds such as amines [organic compound derived from ammonia (NH_3)] or nitriles [any of a class of organic compounds having molecular structures in which a cyano group (?C≡N) is attached to a carbon (C) atom]
Chlorine_bearing_compound	Sample containing chlorine (Cl). Typically the type of compounds detected in presence of perchlorates or other oxychlorines in the sample
Sulfur_bearing_compound	Sample containing sulfur (S). Typically the type of compounds detected in presence of sulfate minerals
Alcohol	Any of a class of organic compounds characterized by one or more hydroxyl (−OH) groups attached to a carbon atom of a hydrocarbon chain
Other_oxygen_bearing_compound	Samples contain oxygen atoms but are not carboxylic acids or alcohols. Examples are esters (R-COOR') and ethers (R-OR')
Mineral	Naturally occurring homogeneous solid with a definite chemical composition and a highly ordered atomic arrangement, usually formed by inorganic processes

(iii) **Abundance**: The rate of ions being detected per second (note: abundance values are compared in a relative way within each sample's analysis).

For the GCMS challenge, competitors were asked to predict the probability that each of the molecular classes or minerals described in Table 3 was detected in the chromatogram. Because of the nature of pyrolysis-GCMS, the compounds detected in the chromatogram cannot be directly extrapolated to the ones that were originally present in the sample. These classes represent a range of chemical families that are of scientific interest in analysing conditions for past habitability, or that have been found on Mars. Details about this GCMS challenge are available on DrivenData website: https://www.drivendata.org/competitions/97/nasa-mars-gcms/page/519/. Similar to the EGA-MS challenge, each sample can have any number of class assignments. For the GCMS challenge, only commercial instrument data were used. As shown in Fig. 7, the top three winners performed similarly overall.

Similarly to the EGA-MS challenge, we provided a benchmark to the participants containing basic exploration data analysis steps, preprocessing, and model development. The EDA step aims at better understanding the proportion of samples in each training, validation, and testing set for each instrument type (commercial versus flight-like), and some main features of MS data. The preprocessing step standardizes the mass values, removes background noise, and converts abundances to relative abundances. The benchmark model uses a simple modelling approach as 'one versus all' for this multilabel classification: binary classifiers using logistic regression are developed for each label class independently. The winners' solutions are described in more detail in Section 4.1.

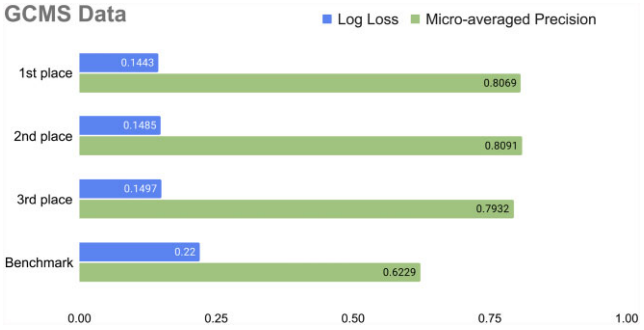

Figure 7. GCMS results for the top three winners and the benchmark only on commercial data. Logistic loss in blue (the lower the better), and micro-averaged precision in green (the higher the better). We can note that the top three winners' solutions outperform the benchmark model especially for the micro-averaged precision metric. The top three winners' solutions are further detailed in Section 4.

4. CHALLENGE RESULTS

These two challenges organized by NASA and hosted by DrivenData focusing on MS data for Mars exploration raised a lot of interest from the community. Indeed, we experienced an extensive engagement from all over the world, with 9962 site visitors from 142 different countries (data obtained with Google Analytics).

Regions	%	Countries	%
Asia	41	USA	24
North America	27	India	21
Europe	21	Turkey	6
South America	5	Russia	3
Oceania	5	UK	3

4.1 EGA-MS challenge (Feb–Apr 2022)

The first challenge organized in the first quarter of 2022 that focused on EGA-MS data received an extensive engagement, with 713 unique participants and 656 submissions. Out of these submissions, 93 participants beat the benchmark model's score (0.3242 aggregated log loss) with a log loss value of 0.092 and 0.95 average precision for the first place solution. The first place participant also won the SAM testbed modelling methodology bonus prize for technical merits and potential to be applied to future data. The participants used a variety of approaches: two-dimensional (2D) deep learning model, ensembles of different tree-based and deep learning trained on 1D representations of the data as shown in Table 4.

The winners (dmytro, _NQ_, and devnikhilmishra) brought a wide variety of creative strategies to perform this task such as feature engineering, data augmentation, and ensembling. Feature engineering is the process of selecting and transforming or creating relevant features from input data to improve the ML models performance.

Table 4. Summary table of the GCMS challenge top three winners, bonus winner, and benchmark models, representing the main modelling approaches used per model and the two performance metrics used for evaluation.

		Modeling Approaches					Performance Measures	
	Rank	Model Description	1D Deep Learning	2D Deep Learning	Peak Features	Statistical Features	Log Loss (lower is better)	Micro-averaged Precision (higher is better)
EGA Challenge **713 Participants** **656 Submissions** **(93 beat the benchmark)**	1st place +Bonus	Representation of mass spectra as 2D images (temperature vs m/z values) then used as an input to CNN, RNN and transformer-based models Many different preprocessing methods	X	X			0.092	0.9483
	2nd place	Feature engineering includes scaling m/z channels and area under the curve, peak value, peak width, and others A LightGBM (light gradient boosting) model trained with these features ensembled with a NN with 2 Conv1d modules, operating over temperature, followed by a linear layer across m/z channels and then a multi-target classifier	X		X		0.116	0.9213
	3rd place	Conversion of the multilabel problem into a binary classification problem LightGBM k-fold ensemble model to get the initial predictions, then fed these predictions along with top 5k features to a 31 fold ensemble, catboost model		X			0.1189	0.9175
	Benchmark	Logistic regression model for each label class using the max relative abundance of each m/z over a 100° C temperature bin as regressors					0.3242	0.6378

In this EGA-MS challenge, feature engineering was used to capture the ion abundance curves. Data augmentation techniques are used to artificially increase the size of the input data set, but applying different transformations to the original data such as adding noise, shifting some peaks, etc. Ensembling techniques involve combining predictions from several ML models to produce a single and more accurate prediction.

The first-place winner converted the input mass spectrum in 2D image representations (temperature versus *m*/*z* values) before using these as inputs to convolutional neural networks (CNNs) and recurrent neural networks. This participant did not use feature engineering but used ensembling techniques and made extensive use of augmentation techniques by representing a single sample 16 times. The second-place and third-place winners both used feature engineering and ensembling techniques. The second-place winner used a light gradient boosting (LightGBM) model trained with the engineered features and 1D deep learning neural networks. Finally, the third-place winner who converted the multilabel problem into a set of binary classification problems, used LightGBM model along with ensembling techniques (Poplavskiy, Lander & Mishra 2022).

4.2 GCMS challenge (Oct–Dec 2022)

The second challenge was organized in the last quarter of 2022. It focused on GCMS data and again received an extensive engagement, with 537 unique participants and 491 submissions. Out of these submissions, 43 participants beat the benchmark model's score (0.2200 aggregated log loss) with a log loss value of 0.14 and 0.81 average precision for the first place solution. The top of the leaderboard (nvnn, dmitryakonovalov, and ouranos,) was very close with the top five participants separated by less than 0.01 aggregated log loss. The bonus prize for the best write-up of methods also expanded understanding of modelling approaches and increased visibility beyond the top three winners. The winners of this second challenge used deep learning models similar to the first place EGA challenge winner with the main differences being in how (and whether) they combined the predictions of these deep learning models with those from other models as shown in Table 5.

For this second challenge, the winners also brought a wide variety of creative strategies to perform this task such as feature engineering, statistical features generation, and ensembling. They also were able to leverage the successful techniques of the first challenge (e.g. converting the mass spectra from 1D representation to 2D image representations). The best solutions included deep learning models using image or sequence representations of the input mass spectra. The first- and second-place winners both used deep learning approaches while the third-place and bonus prize winners first generated features to describe the input mass spectra. The third-place winner engineered statistical features across the entire sample (e.g. means and standard deviations of ion intensity per time interval), while the bonus prize technique engineered features commonly used in signal processing such as peak height and peak width. The top solutions of this GCMS challenge mainly used 2D deep learning models CNNs. The third-place winner used CNN models as well but also combined two tree-based models (logistic regression and ridge classification). All these awarded solutions used some form of ensembling by either training multiple models with different types of preprocessing and of model architectures or by training models on different subsets (also called 'folds') of the input data NVN et al. (2022).

5. DISCUSSIONS

5.1 Challenges successes

The organization of these two challenges was successful in many aspects. The main one being the close collaboration between the different partners: the sample science team, the data science team, and the management team. The sample scientists worked closely with the challenge organizers in framing the proper problem of each challenge and in preparing the data sets. This close collaboration proved essential in developing thorough documentation of specific domain knowledge to help participants understand the problem driving these challenges. Secondly, the challenge host and organizer DrivenData developed a well-written and understandable benchmark code and tutorials to provide the participants a mature starting point. Thirdly, setting up these challenges required the preparation of research data sets for ML applications, with meaningful labels for analysis. Developing and featuring these unique data sets for the broader community to engage with was an important product of the

Table 5. Summary table of the EGA-MS challenge top three winners and benchmark models, representing the main modelling approaches used per model and the two performance metrics used for evaluation.

		Modeling Approaches					Performance Measures	
	Rank	Model Description	1D Deep Learning	2D Deep Learning	Peak Features	Statistical Features	Log Loss (lower is better)	Micro-averaged Precision (higher is better)
GCMS Challenge 537 Participants 491 Submissions (43 beat the benchmark)	1st place	Ensemble of a 1D CNN-transformer then fed into a 1D event detection network and a 2D CNN (each channel has different preprocessing method), used pretrained CNN backbones	X	X			0.1443	0.8069
	2nd place	Ensemble of 13 2D CNNs with different preprocessing methods (variable time bins), used pretrained backbones		X			0.1485	0.8091
	3rd place	Ensemble of 4 models: logistic regression, ridge classification (with feature selection), simple CNN, efficientnet CNN, used pretrained backbones		X		X	0.1497	0.7932
	Bonus	Used novel model averaging for weights of 10 deep learning models, engineered features to describe peaks			X		0.1508	
	Benchmark 1	Logistic regression model for each label class using the max relative abundance of each m/z over 30-sec time bins as regressors					0.22	
	Benchmark 2	resnet18 CNN trained on 2D image representation, which m/z value and time as spatial dimensions and relative abundance as intensity		X			0.3076	0.6229

challenge. These data sets and labels are publicly available on the registry of Open Data on AWS (https://registry.opendata.aws/) and details about the two challenges can be found on the DrivenData website (https://drivendata.org/). Finally, the bonus prizes required written documentation that provided great visibility into the various models' approaches and potential for future applications, as well as recommendations for future challenges.

5.2 Challenges and potential improvements

The great collaboration between the experts (mass spectrometer scientists) and the challenges' organizers was highly time-consuming while being critical for the set-up. For future challenges, we recommend to involve the science team early in the process and provide funding support for these problem framing and data set preparation tasks. We also recommend scientists to be aware and open to any future data science potential tasks on their data and project early in the development phase, in order to optimize the data strategy and metadata collection in a thorough and well-documented manner. ML applications usually require a large volume of data and consistent data sets. In many current applications, ML techniques are applied using an opportunistic and existing data set to investigate methods to extract meaningful insights from it. For future challenges, we would recommend collecting data with potential future ML applications in mind. Although we acknowledge it may be difficult to set up, the data collection considering ML applications will need (1) to better keep track of experimental parameters, metadata, and label annotations (preferably in a virtual manner, instead of laboratory notebooks); (2) to keep experimental procedures consistent (same metadata, same experimental profiles); and if possible (3) to collect more representative data set for specific cases of interest. Finally, it would be beneficial to better understand contamination during experiments (from sample to sample, from chemical noise of the instruments) and incorporate that variable into the problem framing and modelling in a useful way.

5.3 Main takeaways

We demonstrated that multiple ML approaches can be leveraged effectively with MS data for planetary science samples. Data science techniques and ML models can be used to better analyse, investigate, and understand the chemical composition of materials from other planets, and could greatly benefit future space exploration missions. This work adds up to previous work proving the capabilities of ML-based and data science-based methods using MS data in the field of planetary science (Da Poian et al. 2022; Theiling et al. 2022).

These two open science challenges also provide some evidence that models trained on commercial data have some transferability to rover science instruments onboard planetary exploration missions. Scientists and engineers could then leverage existing models and develop models to inform their research for planetary missions. Further tests and research will be needed to determine the extent and the limitations of this transferability for SAM data on Martian applications and for other planetary targets such as Ocean Worlds moons (e.g. Titan, Europa, and Enceladus).

6. CONCLUSION

These two open science ML challenges' results demonstrated that multiple ML approaches can be leveraged effectively with MS data for planetary science analysis. The results of these challenges provide evidence that models trained on commercial instrument data set have some transferability to flight-like science instruments. This is a substantial step forward in the development of ML algorithms for planetary science discoveries.

Open challenges are a marvellous resource and valuable platform for advancing research in various fields. The formulation and implementation of open science ML challenges require a well-organized framework to tackle data preparation, benchmarking steps, evaluation metrics choice, and long-term sustainability challenges. The most time-consuming step after defining the challenges tasks was the data preparation that included the data collection and labelling, as well as the anonymization and data protection measures to prevent misuse or unauthorized access. When defining the challenges' tasks, our team highly focused on resource development to attract a diverse and engaged participant community. With the help of NASA team members and science experts, the DrivenData team developed concise and clear resource documentation about planetary science missions' limits and MS data. Our team also brainstormed on the most suitable evaluation metrics for the tasks to solve and the available data sets.

These initial results from these two open science ML challenges will serve as bases for future work using SAM data collected on Mars and archived on NASA's PDS system. We will also investigate collaborative science using EGA-MS and GCMS data together in a single task, and collaborative science using other instruments onboard the Curiosity rover (e.g. CheMin, Curiosity cameras).

The main takeaways are the engagement and enhanced creativity from worldwide participants in various fields, the reproducibility of the developed models (participants are required to provide detailed descriptions of their method and code), and the benchmarking of the challenges allowing participants to compare the performance of various models.

ACKNOWLEDGEMENTS

The research described in this paper was hosted by NASA Goddard Space Flight Center (GSFC) and by DrivenData, under a contract with the National Aeronautics and Space Administration (NASA). The authors would like to acknowledge the reviewers and the editors for providing feedback which greatly helped in improving the clarity of this manuscript. We thank the NASA Headquarters teams, Steven Rader, Katie Baynes, Steven Crawford, and Megan Ansdell for the funding support, the constant management support, and for their guidance along the project. We also thank the SAM scientists' team for sharing their data and their expertise. Finally, we thank all the worldwide participants for these two incredible challenges, the complete list of the participants can be found on DrivenData webpages: https://www.drivendata.org/competitions/93/nasa-mars-spectrometry/leaderboard/ and https://www.drivendata.org/competitions/97/nasa-mars-gcms/leaderboard/. A portion of this work was supported by NASA under award number 80GSFC21M0002.

REFERENCES

Archer Jr. P. D. et al., 2013, J. Geophys. Res. Planets, 119, 237
Arevalo R., Ni Z., Danell R. M., 2020, J. Mass. Spectrom., 55, 1
Brockwell T. G., Meech K., Pickens K., Waite J., Miller G., Roberts J., Lunine J., Wilson P., 2016, Proc. IEEE Aerospace Conference. IEEE
Bull P., Slavitt I., Lipstein G., 2016. preprint (arXiv:1606.07781)
Chou L. et al., 2021, Frontiers Astron. Space Sci., 8, 173
Clark J. V. et al., 2019, J. Geophys. Res. Planets, 125, e2019JE006173
Da Poian V., Lyness E., Danell R., Li X., Theiling B., Trainer M., Kaplan D., Brinckerhoff W., 2022, Frontiers Astron. Space Sci., 9, 848669
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Franz H. B. et al., 2020, Nat. Astron., 4, 526
Goesmann F. et al., 2017, Astrobiology, 17, 655
Grotzinger J. et al., 2012, Space Sci. Rev., 170, 5
Grubisic A. et al., 2021, Internat. J. Mass Spectrometry, 470, 116707
Hart W. et al., 2018, IEEE Aerospace Conference. Big Sky, MT, p. 1
Hložek R. et al., 2020, preprint (arXiv:2012.12392)
MacKenzie S. et al., 2021, Planet. Sci. J., 2, 77
Mahaffy P. et al., 2012, Space Sci. Rev., 170, 401
McAdam A., et al., 2022, J. Geophys. Res. Planets, 127, e2022JE007179
Niemann H. et al., 1996, Science, 272, 846
Nier A. O., McElroy M. B., 1977, J. Geophys. Res., 82, 4341
NVN N., Konovalov D., Nasios I., Ninalga D., 2022, Winning code from the Mars Spectrometry 2: Gas Chromatography challenge, Zenodo, available at: https://zenodo.org/records/8284743
Poplavskiy D., Lander A., Mishra N., 2022, Winning code from the Mars Spectrometry: Detect Evidence for Past Habitability challenge. Zenodo, available at: https://zenodo.org/records/8284806
Sutter B. et al., 2017, J. Geophys. Res. Planets, 122, 2574
Theiling B. et al., 2021, BAAS, 53, 048
Theiling B. et al., 2022, Astrobiol., 22, 0062
Thompson D., Castillo-Rogez J., Chien S., Doyle R., Estlin T., McLaren D., 2012, AIAA Meeting Paper - SpaceOps 2012 Conference. ARC
Wong L., Michaels A., 2022, Sensors, 22, 1416

This paper has been typeset from a TEX/LATEX file prepared by the author.