Machine Learning on Mars Mass Spectrometry Data: Unlocking Insights for Autonomous Planetary Exploration.  S. P. Do¹, V. Da Poian²,3,4, E. Lyness⁵, R. Danell⁶, X. Li⁷, and W. Brinckerhoff⁸. ¹Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, ²NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD, 20771 ³Microtel LLC, 7703 Belle Point Drive, Greenbelt, MD, 20770, ⁴Johns Hopkins University, Earth and Planetary Science Department, Baltimore, MD, 21210, ⁵Danell Consulting, Winterville, NC, 28590.

Introduction: The search for potential life on other planets continues to drive exploration missions across the solar system, with Mars remaining a prime target. The Rosalind Franklin (ExoMars) mission, led by the European Space Agency (ESA) and supported by collaboration with NASA, seeks possible evidence of past and present life on Mars through its eponymous rover [1,2,3]. The success of such an ambitious mission critically depends on advanced scientific investigations, with the Mars Organic Molecule Analyzer (MOMA) instrument playing a core role in this endeavor [4] to detect and analyze organic matter on the Martian surface and subsurface [1,5]. MOMA data may provide crucial insights into the planet's past habitability and potential evidence of life.

Space missions face significant constraints, including limited communication bandwidth that results in long communication delays and limited data downlink capacities. Traditional ground-in-the-loop workflows, relying on manual data review and decision-making by scientists on Earth, increasingly hinder mission efficiency. To overcome these challenges and optimize performance, the concept of science autonomy has emerged as a compelling capability [6,7]. The long term goal of science autonomy is to enable instruments like MOMA to independently optimize their operations and support data analysis, reducing reliance on Earth interactions and prioritizing data telemetry as appropriate. In this context, machine learning (ML) has gained considerable attention, providing intelligent data analysis techniques through filtering and sample identity prediction, allowing real-time decision-making by the science instrument. An intermediate goal of science autonomy for Rosalind Franklin, launching in 2028, aims to support scientists during their analysis of Mars data for tactical and strategic decision-making during surface operations. While previous research has explored the application of ML models on mass spectrometry (MS) data from MOMA's flight-like engineering test unit (ETU), opportunities for accuracy improvement remain [4]. By refining ML models and augmenting data volumes, we aim to optimize data selection and analysis and enable the rover science team to make well-informed decisions rapidly based on real-time observational evidence. Building on studies demonstrating ML’s potential in space missions (e.g., the Autonomous Exploration for Gathering Increased Science (AEGIS) system on the Curiosity rover [8,9]), our research seeks to pioneer novel methods to improve the accuracy and robustness of ML models for analysis of MS data through data augmentation techniques. Data augmentation is a regulation technique that consists of relevant transformation operations on the original data to create additional training data. Its primary application is to increase the volume of a model’s training set to prevent overfitting and increase the robustness of the model, ultimately advancing mission science objectives [10,11].

Methods: In this work, we leverage laser desorption MS (LDMS) data libraries obtained from the MOMA ETU for analysis using ML models, and we transform the input data from native 1D mass-to-charge (m/z) arrays into 2D image representations to leverage the capabilities of convolutional neural networks (CNNs). The resulting 2D images serve as input data for our ML models. Two separate models are developed using about 35,000 LDMS spectra as input – one to predict chemical family identity (9 labels) and one to predict individual sample identity (68 labels). Given the data-hungry nature of CNNs, we apply data augmentation techniques to increase the effective size of the training data. We use seven augmentation methods, focusing on intensity augmentation (randomly selected MS peaks and all peaks), shifting (randomly selected peaks and all peaks), stretching, and noise augmentation (Gaussian and random). The augmented data was combined with the original images to create an expanded training dataset.

Results: The application of the proposed ML model architecture and data augmentation techniques yield promising outcomes in both chemical family and sample classification tasks. Prior to data augmentation, the model demonstrated a commendable accuracy of 93.966% in predicting the correct chemical family and 90.184% in identifying the correct sample identity. With the integration of data augmentation, the model's performance increased significantly, achieving exceptional accuracies of well over 95% in classifying both chemical families and sample identities (Figure 1). This improvement can be attributed to the efficacy of CNNs on the 2D image format instead of the 1D arrays, the selected model architecture, and the judicious use of specific augmentation methods, specifically the intensity augmentation and random intensity augmentation. Notably, all data augmentation methods, excluding Gaussian noise, contributed to the
overall enhancement of results, and the most accurate models had the most data to train on – 3 times the amount of data in the original dataset.

Figure 1. The training and validation accuracy plot over epochs showcases the supervised model's performance in predicting the sample identity label. The model was trained on a combination of the MOMA ETU data and augmented data with randomly increased peak intensities up to 5% and 10% from the original dataset. The precision accuracy, macro average, and weighted average over this training set were 99.046%, 98.188%, and 99.117%, respectively, over more than 15 epochs.

**Discussion & Conclusion:** This research demonstrates the potential of ML for planetary science data analysis. The model achieved over 95% accuracy in predicting chemical samples using MOMA ETU MS data. The improvement in classification accuracy, enabled by data augmentation, helps address the challenge of data scarcity for ML on space missions. Integrating ML models with appropriate data preprocessing and augmentation enhanced the analysis of MOMA data. Future research will explore the application of additional avenues of data augmentation and its impact on model performance, such as leveraging generative adversarial networks. While this research primarily focused on the MOMA ETU LDMS dataset, future work can expand to other datasets (i.e. non-LDMS data).

**Acknowledgements:** This research was supported in the MOMA-MS Project via NASA's Mars Exploration Program.