AGNOSTIC POLYMER DETECTION USING MASS SPECTROMETRY FOR ASTROBIOLOGICAL SAMPLES.

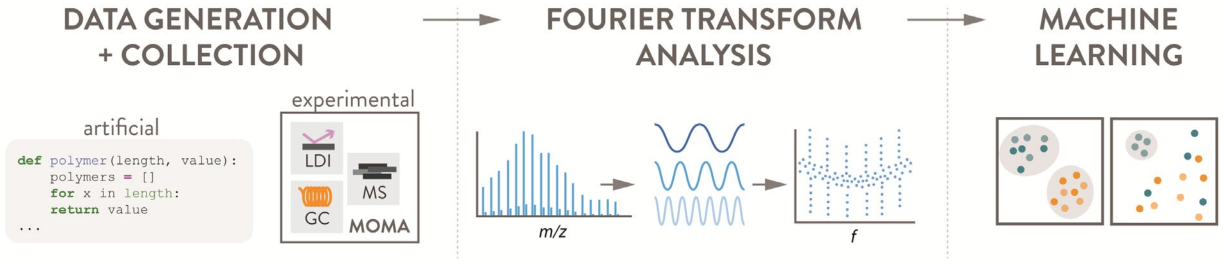
L. Chou[1,2], N. Grefenstette[3], V. Da Poian[2], C. Kempes[3], H. Graham[2], A. Roussel[1], P. Mahaffy[2], S. S. Johnson[1] [1]Georgetown University (37th and O Streets NW, Washington, DC 20057), [2]NASA Goddard Space Flight Center (8800 Greenbelt Rd, Greenbelt, MD 20771, luoth.chou@nasa.gov), [3]Santa Fe Institute

Introduction: *Agnostic biosignatures.* As more planetary missions are being planned to explore the solar system, such as ExoMars2020 and Dragonfly, a new way of approaching life detection is necessary [1]. Traditional methods for life detection have focused on biosignatures that are based on Terran biochemistry. In certain contexts, these methods may detect unmistakable biosignatures, but they also risk missing signs of life as-yet unrecognized [2]. To avoid biasing our search, we need to reconceptualize the framework within which we look for life in general, be it exotic or not. Here, we focus on developing new techniques for processing and interpreting return mass spectrometry data in order to infer the presence of agnostic biosignatures without presupposing the existence of certain Terran biochemical machineries.

Space-capable applications. In line with a more general approach to life detection, the methods used to process and interpret return data from flight instruments also need to improve in parallel with our efforts to identify agnostic biosignatures. The Mars Organic Molecule Analyzer (MOMA) on board the ExoMars2020 rover houses a dual-source linear ion trap mass spectrometer, capable of laser desorption ionization and gas chromatography, enabling considerable breadth in the investigation of organic matter across a wide range of molecular weights and volatility with an added advantage of simple sample preparation requirement [3]. MOMA is also capable of performing tandem mass spectrometry, where specific ions of certain mass-to-charge ratio (m/z) can be selected, accumulated, and further fragmented into smaller components [3]. The MOMA instrument will also serve as a prototype for the mass spectrometer, DraMS, onboard the Dragonfly octocopter that is designed to explore Titan. This work leverages MOMA-like instrument suites to determine if agnostic biosignatures can be detected in astrobiologically-relevant samples.

Polymers as a universal biosignature. One of the fundamental features of life on Earth is its use of polymers (DNA/RNA, proteins, etc.), composed of a repeated set of limited building blocks, in an organized pattern. The use of polymers allows life to access to a larger chemical space, as well as reliably store and propagate information [4]. For these reasons, we make the assumption that polymers hold a key role in universal biology. Here, we develop methods for detecting polymers in the mass spectra of unknown samples. We discuss the challenges associated with identifying polymers, recovering the constituent parts of those polymers, and, in rare cases, sequencing those polymers. We use innovative signal processing methods to investigate the data. These methods build upon traditional approaches used for analyzing known biopolymers such as sequencing, as well as other techniques for examining the data structure such as the Fourier transform.

Machine learning. While the detection of polymers or classes of polymers in a sample alludes to the potential utilization of a polymeric system by life, it requires the evidence necessary to qualify this system as biotic. This is because polymers are relatively ubiquitous in the solar system [5]. For example, on Titan, the organic chemistry that takes place in the atmosphere, induced by photolysis or radiolysis, can form complex organic compounds that polymerize into larger particles (aerosols) that are eventually deposited on the surface as "tholins" [6]. Tholins can contain a wide variety of polymers that are abiotically produced [5]. While the random natural production of polymers in extraterrestrial conditions presents a tremendous challenge to the search for agnostic biosignatures, we aim to use machine learning and data science tools to understand and classify the materials that are abiotically produced and materials that contain the polymeric signals of life. First, we will apply our data processing



```
def polymer(length, value):
    polymers = []
    for x in length:
    return value
...
```

Figure 1: Analytical and computational workflow for the work described here.

method to polymer samples of known abiotic origin (i.e., Titan tholins analogs, meteorite extracts) and known biotic origin (i.e., proteins, DNA/RNA, sugars). Then, we will perform a supervised machine learning analysis on the processed data (Fig. 1).

Methods: *Artificial data generation:* In order to enhance our ability to classify samples based on their mass spectrum, machine learning processes require large amounts of data. Thereby, we seek to generate artificial mass spectral data that could significantly improve our machine learning algorithms. This artificial polymer mass spectrum is generated based on a defined monomer set (i.e., amino acids). We assign a value for the length and define a sequence for the polymer and determine the potential mass fragments based on the assumption that the polymer will break at every monomer unit in the mass spectrometer. The mass fragments are then concatenated together to form a mass spectrum with artificial values for intensity. All computational work described in this study is written in the Python 3.0 programming language.

Noise generation: Analytical noise is an important component of mass spectral data obtained in the laboratory or from space. For this work, we analyze the frequency-rank (intensity vs. rank of *m/z*) distribution of the mass spectrum of known polymer samples, such as peptides (angiotensin), Titan tholin analogs, meteorite extracts, and complex mixtures (i.e., microbial extract and sediment samples) to determine any underlying power law (i.e., Zipf's distribution) that potentially describes the analytical noise intrinsic to our instrument. Zipf's law states that the frequency distribution of observations is inversely proportional to its rank [7]. Thus, we define a function that potentially describes this distribution and artificially produce noise peaks that follow that distribution. We then add these peaks to the artificially generated polymer mass spectra.

Sequencing. In some cases, we use a sequencing algorithm to determine the order in which the monomers are linearly arranged in a polymer. To do so, we create a program that compares the mass differences between peaks and matches them to a known monomer set (i.e., amino acids). We analyze all the possible sequences, or combination of sequences, and determine the limitations and error rates of this program.

Fourier Transform. The mass spectrum of polymers can contain information about the monomer size(s), polymer length, and potentially the sequence of polymers within the sample. We hypothesize that these properties of the polymers can be extracted using Fourier analysis methods and/or sequencing methods on the mass spectrum. Fourier transform can help identify harmonics in the mass spectra signals, which would aid in the identification of fragmentation patterns of repeating subunit losses, and thus, the presence of a polymeric compound. Sequencing, on the other hand, can be used to calculate potential combinations of subunits in a monomer that appears as a peak in a mass spectrum. Successful sequencing would reveal the existence of a polymer that contain a subset of repeated building blocks which could point to a potential utilization of an information storage system.

Experimental verification. We will test our methods on data that we artificially generated as well as data derived from MOMA-like experiments. Our samples include known abiotic polymers such as tholins analogs and meteorite extracts, and biotic polymers, such as peptides and DNA, as well as combinations of those samples, and complex environmental samples.

Preliminary Results: We successfully created an algorithm to generate artificial mass spectra for polymers using the 20 canonical amino acids as monomer subunits, though this program can use any starting monomer set. Initial results indicate that the Fourier transform method is able to detect the presence of polymers within a sample and correctly identifies the one-monomer building block with errors between 0.14 to 3.29 %. We also observed noticeable differences between the Zipf's distribution of pure samples (containing one molecule), to that of complex samples, and environmental samples. These differences will inform on our noise generation algorithm and further enhance our ability to produce realistic artificial spectra for machine learning purposes.

Future work: A major goal of this work is to be able to apply the Fourier analysis method on artificially produced and experimentally derived mass spectra for agnostic biosignatures detection. To do so, we will create a signal processing (noise removal) algorithm that can be universally applied to MOMA-like mass spectral data. We will then perform Fourier analyses on both artificial and experimental data. Ultimately, the transformed power spectrum will be used to classify samples of known biotic polymers and known abiotic polymers, providing a potential supervised machine learning method that can be used to detect agnostic polymer biosignatures in unknown samples.

References: [1] Johnson S.S. et al. (2018) White paper for *NAS Strategy for the Search for Life in the Universe.* [2] Johnson S.S. et al. (2018) *Astrobiology* 18(7) 915 - 922. [3] Li X. et al. (2017) *International Journal of Mass Spectrometry*, 422: 177 – 187. [4] Benner S. (2017) *Astrobiology*, 17:9, 840-851. [5] Pernot, P (2010). *Analytical chemistry*, 82(4), 1371-1380. [6] Sagan C. and Khare B. (1979) *Nature*, 277: 102-107. [7] Benz R. W. (2008) *Journal of Chemical Information and Modeling*, 48, 1138–1151.