The opportunities from machine learning applications in astrobiology

Caleb Scharf, NASA Ames Research Center, Moffet Field, California.
caleb.a.scharf@nasa.gov

The search for life represents a unique data challenge within modern science. Machine learning, as it is now and may be in the future, offers many new opportunities for addressing this challenge. Astrobiology should systematically investigate and evaluate those opportunities or risk diminished scientific returns.

At its core the search for life is a search for life's imprints in the world, reflected in the *features* in data; meaning the measurable properties, variables or attributes in data associated with those imprints. To try to determine whether life exists, or has existed, elsewhere in the universe, evidence of the nature of those features is accumulated through measurement and experimentation. But the necessary measurements or experiments needed to verify and characterize life vary according to circumstances. One extreme example would be if extant extraterrestrial organisms were acquired in sufficient abundance so that a definitive claim could be made for the presence of life that is directly and uniquely assignable to its native surroundings. At the other extreme is the accumulation of many correlated pieces of evidence - both contextual and direct – that individually do not fully confirm or reject biogenicity but together lead to a probabilistic estimate of the existence of life (e.g., [1-3]). Such evidence might require the mapping and monitoring an entire planetary object (or even an entire planetary system) across heterogeneous spatial conditions and varied temporal scales.

In any given case, including these extremes, proof of discovery will be less significant without the construction of a plausible history and characterization of the lifestyle of any living systems. For example, the discovery of an abundant, high complexity organic molecular species in an environment (with complexity gauged by an approach such as assembly theory [4]) would be very strong evidence of biogenicity. But if there were no robust hypothesis for the origin of such molecules (e.g., if there was no match to any known biochemical function or byproduct, or if these molecules existed in surprising conditions), it would be hard to see this as full confirmation of biogenicity – although it would clearly prompt much further investigation.

To date however the search for life has tended to prioritize the imprints of life that are thought to offer the greatest chance of successful *feature measurement*, but not necessarily to maximize certainty and informational leverage on the nature of any biogenicity. This has sometimes resulted in controversial or inconclusive outcomes. A classic example is the C^{14} Labeled-Release nutrient experiments onboard the Viking 1 and 2 landers, where a test that in principle yields guaranteed measurements in many terrestrial circumstances (that, by definition, are extreme in terms of their fecundity), proved enormously challenging and contradictory under Mars surface chemical conditions and in the context of results from other concurrent experiments.

A more modern example would be the proposed use of the presence of molecular oxygen in remotely sensed rocky planet atmospheres. In certain conditions the detection of molecular oxygen would seem to be an excellent signature of a possible biosphere and is accessible with foreseeable technological detection capabilities. But in other, abiotic, conditions it is a potential false positive, and the absence of oxygen does not eliminate the possibility of the existence of a robust anoxic biosphere, as was the case for the first 2 billion years of life on the Earth [5]. Theoretical studies do yield additional candidates for atmospheric compounds that could be more informative and definitive as biosignatures [6], but these are increasingly challenging to measure. Consequently, a consensus has formed that the likeliest successful search for life will inevitably involve a multitude of independent measurements and an accumulation (or "ladder" [2]) of evidence that, in effect, blends low and high informational value in features.

Precisely how to rank or weight informational value or best handle data from heterogeneous conditions remains a topic of considerable debate [2, 7]. For instance, while heterogeneity might be dealt with by examining statistical distributions of measurements, that approach could dilute underlying signals or fail to account for, or deconvolve, the inherent limitations of the instruments being used to make the measurements. Or, if similar measurements are made independently and capture data features relating to different facets of a single underlying, highly complex phenomenon, it is critical to understand how the features relate to each other.

Beyond business as usual: machine-in-the-loop

Machine learning presents an opportunity for astrobiology to seek new ways to address these challenges and is already seeing rapid growth in use, driven in part by a cascade of algorithmic breakthroughs following 2012's revolution in deep learning [8]. Recent works have explored machine learning for biosignature detection and assessing environments in astrobiology [e.g., 9-12] and paleobiology [e.g.,13–16] as well as mineral discrimination in planetary science [e.g. 17].

A defining characteristic of all machine learning approaches is a capacity to "self-program" to model features in data and find those that are most informative. In the simplest, almost trivial, cases this might involve modeling least-squares curve-fits. The most complex cases involve machine learning approaches that – in effect – build a model for mapping the features in data as multi-dimensional vectors projected into a so-called "latent space"; an embedding within a high-dimensional manifold that nonetheless has lower dimensionality than the feature space of the data used to train on. This is a type of data compression that attempts to capture all features in data in a way that preserves their most information-rich relationships, enabling tasks like classification or prediction.

In principle that makes machine learning approaches capable of discovering far more complex (i.e., non-linear, high-dimensional) patterns in data, with greater sensitivity to informative correlations between variables that would evade classical techniques. In practice this also makes the preparation of data a much more critical process, since unexpected biases and unappreciated imprints of method (e.g., how measurements are binned, or variations in instrument sensitivity) can confuse machine learning discoveries. The complex, non-intuitive abstraction of data within machine learning models also presents a very significant barrier to interpreting how and why any machine learning approach arrives at its results or outputs. Furthermore, the continued rapid pace of evolution in machine learning approaches presents an additional challenge for scientific applications where consistent interpretability and reproducibility of results are paramount.

Astrobiology can mitigate these issues and continue to overcome some of its data challenges by pursuing intensive work on state-of-the-art machine learning to evaluate opportunities and areas for scientific advancement. This could take the form of a systematic program to: ***(1) Rapidly develop and deploy a variety of astrobiology-orientated machine learning applications and assess their efficacy***, and ***(2) Assess the totality of astrobiological data to date and its future-facing suitability for machine learning use (e.g., bias, breadth, sensitivity, heterogeneity)***. Such steps would enable the field to narrow some of the options for machine learning development and proceed along the most optimal path.

Machine learning is also very likely to play a larger future role in instrument and mission design and operation [18]. Especially in autonomous science, and support of science during human exploration. Research that explores opportunities in machine-learning-enabled autonomous science for astrobiology (including hypothesis generating systems, data collection decision making, and real time system control) for astrobiology would help ensure that the field maintains relevant capabilities and is positioned to play key roles in mission concept design and eventual flight. That work could both leverage and inform the elements **(1)** and **(2)** above.

Deep learning and the nature of astrobiological data:

As described above, some machine learning approaches, such as various forms of deep learning, build a model during training that provides a mechanism for mapping new input data into a multi-dimensional latent space. For example, a convolutional autoencoder working with images will learn how to convert input to an efficient latent representation (within the model's latent space) of the features in those images (in the form of a multi-dimensional array).

Part of that latent representation of an image will describe fundamental properties like edges, gradients, or curves. In deeper layers of the encoder more specialized features can be captured, such as the patterns of materials or subtle relationships between light and shadow under differing conditions. With sufficiently large datasets and large models (e.g., many layers in the encoder/decoder) a system like this can be used (with additional learning structures) to accurately classify images from many classes or perform tasks like image generation or denoising. In this type of machine learning the convolutional autoencoder model (that is learnt) is represented by a fixed structure of interconnected layers and the many parameters that describe different spatial filters and, often, activation functions that introduce non-linear responses. That model is the result of feedthrough of many training images and a process of internal adjustment. Once a model is stopped from further learning it then serves as a mechanism to take an input and create a latent representation of that data which can then be used for the tasks of classification or reconstruction.

In other words, if successful, such a model will map an input to a new, multidimensional representation that contains all the essential features for analysis in the original input, but which itself is not an image in any traditional sense. This raises a few key questions: (***a) if the scientific value of data lies in the results of its use with a trained machine learning model, then is it the compressed latent representation of that data that is relevant, not the raw data?*** *And* ***(b) can the needs of the machine learning model direct the needs for measurements, instruments, and experimental design (e.g., mission architecture)?***

Both questions have relevance across many scientific areas, but are amplified for astrobiology, where solar system exploration under strong constraints of communication bandwidth, instrument complexity, necessary data breadth, and the risks of false positive life detections are all factors of concern.

It is therefore suggested that work be supported to examine the efficacy of the current paradigm of “raw” data utilization in astrobiology and whether existing protocols and data formats are optimal for machine learning applications. This would apply not only to planetary exploration imaging data, but also to in situ or remote compositional analyses and contextual information on physical conditions. Furthermore, should an instrument return unprocessed data (that is seldom truly unprocessed) if the ultimate use is by a predetermined machine learning model that only requires the latent representation of a dataset? And if the latent representation is all that matters for answering scientific questions, then could the instrument itself be designed differently and more optimally?

References:

[1] Catling DC et al. (2017) Exoplanet Biosignatures: A Framework for Their Assessment. Astrobiology Journal 18: https://doi.org/10.1089/ast.2017.1737
[2] Neveu M et al. (2017) The Ladder of Life Detection. Astrobiology Journal 18: https://doi.org/10.1089/ast.2017.1773
[3] Walker SI et al. (2018) Exoplanet Biosignatures: Future Directions. Astrobiology Journal 18: https://doi.org/10.1089/ast.2017.1738
[4] Marshall SM et al. (2021) Identifying molecules as biosignatures with assembly theory and mass spectrometry. Nature Communications 12: https://doi.org/10.1038/s41467-021-23258-x
[5] Meadows VS et al. (2018) Exoplanet Biosignatures: Understanding Oxygen as a Biosignature in the Context of Its Environment. Astrobiology Journal 18: https://doi.org/10.1089/ast.2017.1727
[6] Seager S, Bains W, Petkowski JJ (2016) Toward a List of Molecules as Potential Biosignature Gases for the Search for Life on Exoplanets and Applications to Terrestrial Biochemistry. Astrobiology Journal 16: https://doi.org/10.1089/ast.2015.1404

[7] Cockell CS, Samuels T, Stevens AH (2022) Habitability Is Binary, But It Is Used by Astrobiologists to Encompass Continuous Ecological Questions. Astrobiology Journal 22: https://doi.org/10.1089/ast.2021.0038
[8] Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet Classification with Deep Convolutional Neural Networks. Advances in neural information processing systems 25.
[9] Guttenberg N, Chen H, Mochizuki T, Cleaves H (2021) Classification of the Biogenicity of Complex Organic Mixtures for the Detection of Extraterrestrial Life. Life 11:234. https://doi.org/10.3390/life11030234
[10] Corenblit D, Decaux O, Delmotte S, et al (2023) Signatures of Life Detected in Images of Rocks Using Neural Network Analysis Demonstrate New Potential for Searching for Biosignatures on the Surface of Mars. Astrobiology 23:308–326. https://doi.org/10.1089/ast.2022.0034
[11] Warren-Rhodes K, Cabrol NA, Phillips M, et al (2023) Orbit-to-ground framework to decode and predict biosignature patterns in terrestrial analogues. Nat Astron 7:406–422. https://doi.org/10.1038/s41550-022-01882-x
[12] Figueroa MC, Gregory DD, Williford KH, et al (2024) A Machine-Learning Approach to Biosignature Exploration on Early Earth and Mars Using Sulfur Isotope and Trace Element Data in Pyrite. Astrobiology 24:1110–1127. https://doi.org/10.1089/ast.2024.0019
[13] Wiemann J, Heck PR (2024) Quantifying the impact of sample, instrument, and data processing on biological signatures in modern and fossil tissues detected with Raman spectroscopy. Journal of Raman Spectroscopy 55:761–773. https://doi.org/10.1002/jrs.6669
[14] Wiemann J, Crawford JM, Briggs DEG (2020) Phylogenetic and physiological signals in metazoan fossil biomolecules. Sci Adv 6:. https://doi.org/10.1126/sciadv.aba6883
[15] Loron CC, Rodriguez Dzul E, Orr PJ, et al (2023) Molecular fingerprints resolve affinities of Rhynie chert organic fossils. Nat Commun 14:1387. https://doi.org/10.1038/s41467-023-37047-1
[16] Cleaves HJ, Hystad G, Prabhu A, et al (2023) A robust, agnostic molecular biosignature based on machine learning. Proceedings of the National Academy of Sciences 120:. https://doi.org/10.1073/pnas.2307149120
[17] Julve-Gonzalez S, Manrique JA, Veneranda M, et al (2023) Machine learning methods applied to combined Raman and LIBS spectra: Implications for mineral discrimination in planetary missions. Journal of Raman Spectroscopy 54:1353–1366. https://doi.org/10.1002/jrs.6611
[18] Theiling BP et al. (2022) Science Autonomy for Ocean Worlds Astrobiology: A Perspective. Astrobiology Journal 22. https://doi.org/10.1089/ast.2021.0062