Foundation Models for Astrobiology

A white paper for the 2025 NASA Decadal Astrobiology Research and Exploration Strategy (DARES)

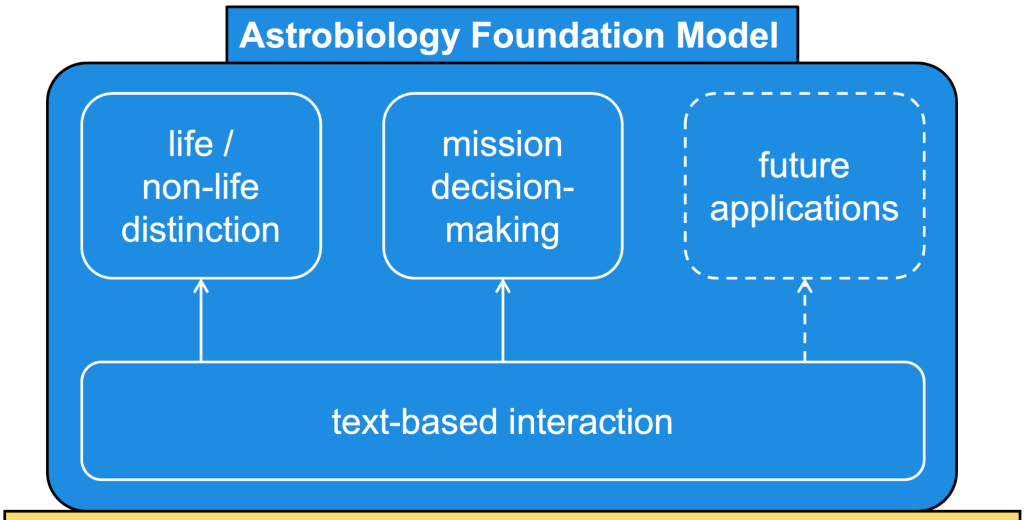
Ryan Felton, NASA ARC - ryan.c.felton@nasa.gov
Stuart Bartlett, California Institute of Technology/SETI Institute
Nathalie A. Cabrol, SETI Institute
Victoria Da Poian, NASA GSFC
Diana Gentry, NASA ARC
Jian Gong, University of Wyoming
James Hendler, Rensselaer Polytechnic Institute
Adrienne Hoarfrost, University of Georgia
Manil Maskey, NASA MSFC
Floyd Nichols, Virginia Tech
Conor A. Nixon, NASA GSFC
Tejas Panambur, University of Massachusetts, Amherst
Joseph Pasterski, NASA GSFC
Anton S. Petrov, Georgia Institute of Technology
Anirudh Prabhu, Carnegie Science
Caleb Scharf, NASA ARC
Brenda Thomson, Rensselaer Polytechnic Institute
Hamed Valizadegan, NASA ARC / KBR
Kimberley Warren-Rhodes, SETI Institute/NASA ARC
David Wettergreen, Carnegie Mellon University
Michael L. Wong, Carnegie Science
Anastasia Yanchilina, SETI Institute

1) Background and Motivation

In this RFI response we present preliminary recommendations from a workshop on Foundation Models (FMs) in Astrobiology that was hosted by the NASA Ames Research Center and the SETI Institute in Mountain View, CA, February 24-26, 2025. The goal of the workshop was to assess FMs towards applications in astrobiology and to produce a number of white papers to guide the community.

A FM [1] is a large-scale machine learning (ML) system typically trained on enormous datasets to encode fundamental, general information and relationships, enabling it to serve as a foundation for various downstream applications with limited additional training and fine-tuning. While building such large-scale FMs requires a great deal of ML expertise and effort, adapting them for specific downstream tasks is often fast and requires minimal ML expertise, making them excellent tools for rapid development in fields that call for a wide range of applications. **For astrobiology, FMs may offer unique and critical opportunities for advancing efforts in life detection and characterization.** In the last few years, FMs have emerged as a new paradigm that can dramatically accelerate the application of ML to specialized tasks and a wider range and types of data [2, 3].

2) An Astrobiology Foundation Model Overview and Applications



Suggested structure of a multimodal astrobiology foundation model, with two quantitative models for life/non-life discrimination and mission decision-making and one language-based model. The structure could be expanded to include other models for future applications.

A) Exploring the Life/Non-Life Continuum

The search for life in the universe is astrobiology's central goal. Defining the essential properties of life, and defining criteria for evidence of life, are two views of the same problem. Challenges to a solution stem from at least two sources. One is our lack of an example of non-Earth-based life, where biochemistries will emerge via distinct coevolutions between living systems and their hosting environment, and may demand detection strategies beyond those based on terrestrial assumptions. The other is that the distinction between biotic and abiotic processes may not be a clear-cut divide but a context-dependent continuum, particularly when considering prebiotic conditions, viruses, prions, and hypothetical alternative biochemistries [4].

Emerging AI/ML tools offer a potentially new approach not only towards identifying complex combinations of data signatures that could be diagnostic of life, but also towards revealing *which* combinations are diagnostic – in essence, an empirical definition of life. **The time is right to build, as part of an initial development stage and application of an astrobiology FM, a tool for biosignature detection.** This will require an iterative approach, starting with proof-of-concept validation using existing data, refining predictive power by addressing key knowledge gaps, and ultimately integrating diverse - multimodal - datasets (*e.g.*, missions and existing data on Earth biosphere) to create a scalable, adaptive framework for life detection. The steps towards this would be:

1) Developing a Proof of Concept by Revisiting Mission Data and Terrestrial Analogs. Building a FM for biosignature detection begins by leveraging existing datasets to construct an initial proof-of-concept, which may include comparing similar

datasets between terrestrial analogs, Mars, the Moon, and various asteroids as “abiotic” endmembers. We identify several data types where we believe there is sufficient breadth in the literature now: visible imagery, VNIR reflectance, elemental and isotopic abundance, gas chromatography mass spectrometry, Raman, XRF/XRD, and material morphology. This subset represents low hanging fruit, and would allow the development of a proof of concept FM for application to biosignatures. The model would then be extended by:

2) Integrating Datasets and Closing Knowledge Gaps. To create a scalable, flexible framework for discovery, data breadth should be grown. An objective would be to standardize all data collection across planetary missions and terrestrial datasets; another objective is to address key data gaps, particularly those that are most urgent in the near term and directly relevant to upcoming astrobiology missions (*e.g.*, progress in distinguishing biotic from abiotic signatures, and mimics). Filling these gaps would strengthen the FM’s predictive power and enhance its ability to recognize coevolutions beyond Earth.

Ultimately, the proposed architecture of a multimodal FM would be robust for downstream uses and could conceivably help formulate hypotheses on the basic principles of life; characterize the boundaries between biotic and abiotic systems and processes; and develop targeted exploration strategies (*e.g.*, science questions, hypotheses, experiments, instruments, technology) for astrobiologically relevant environments.

To ensure the validity of such an FM, multiple layers of analysis must be combined, including: (a) chemical and morphological correlations, (b) biological and ecological relationships, (c) spatiotemporal patterns, (d) thorough contextual analysis (required to evaluate potential biosignatures within their planetary environment); and (e) cross-validation through multiple independent detection methods, that may lead to new insights, correlations and patterns. By unifying these elements into a multimodal detection framework, this model would represent the most comprehensive and adaptable strategy for identifying potential biosignatures, offering the most effective path forward in the search for life beyond Earth, and the understanding of life on Earth.

B) Mission-focused Decision Making

In situ analysis on other planetary bodies provides a direct assessment of an environment. Because of this, flight missions are extremely valuable for the detection of non-Earth-based life forms and their processes. However, flight missions are extremely resource and time constrained compared to laboratory-based analyses on Earth due to size, weight, and power limitations, as well as communication delays and data transfer rate limitations. We therefore **suggest the second downstream application of an astrobiology FM would be a specialized Astrobiology Mission Model (AMM) to be utilized during mission design and mission operations**. The AMM could also be

used to evaluate previous mission datasets to help identify patterns in the data that may indicate life, biotic processes, or habitable environments that were not detected during initial analyses.

For mission design, infrastructures already exist to begin developing an AMM. The astrobiology FM would be fine-tuned with data from previous missions and analog studies (database knowledge), previous mission measurements (mass spectrometry data, XRF data, etc., including metadata), information about planetary bodies and what has been studied, both commercial and prototype instrument capabilities, previously developed proposals and concept studies, as well as laws and regulations that may impact analog field campaigns. The main barriers to begin development would be in the form of funding, paywall limitations from specific journals, the need to "tokenize" measurement data to make it transferable to the FM, the ML readiness of available data, and synchronizing the various data sets together. For mission operations, a developed AMM could be augmented with specific details on the instruments for a specific mission, including exact instrument capabilities and similar commercial analogs, data transmission limitations, and mass and power constraints. Additionally, details about specific biosignatures potentially relevant to the target body that have not yet been incorporated into the AMM could be supplied and analog studies could be conducted during mission preparation in order to increase the performance and achieve the defined mission goals. Taken together, the incorporation of ML into mission design and operations is inevitable. An AMM could be used to enhance mission operations during flight as well as science data interpretation, enabling a more successful mission from conception to operation that would maximize the chance of achieving objectives.

C) AB-Chat (AstroBiology-Chat)

Astrobiological research requires extensive knowledge of both specialized and broad literature, as well as data availability across many areas. This interdisciplinarity presents unique challenges for parsing research records, assembling descriptive text, identifying knowledge gaps, and communicating across disciplinary boundaries.

We suggest the development and application of a specialized text-centric large language model (LLM) incorporating interdisciplinary astrobiology knowledge to assist in identifying critical research gaps, facilitating hypothesis generation and testing, and supporting interactions with the other FM modes. NASA has already invested in science-focused LLMs, for example the INDUS model (based on RoBERTa [5]), that could serve as the basis for a fine-tuned astrobiology text-FM.

The potential applications of an AB-Chat model are extensive. There is opportunity for systematic mining of astrobiology literature; including abstracts, peer-reviewed journals, and preprints. Generation of broadly comprehensible content from technical and specialized literature can facilitate broader scientific communication

and outreach. And mission and research development can be enhanced by synthesis of research or mission proposals and technical reports.

AB-Chat development could proceed rapidly building on top of the INDUS LLM. **Phase 1: Data Collection & Curation:** Further aggregation of astrobiologically relevant literature/data from key journals and additional resources not already captured by the INDUS training corpus (*e.g.,* biosciences, origins of life), and mission-related documents. Develop structured ontologies and taxonomies for critical astrobiology subdomains such as prebiotic chemistry, biosignatures, and planetary environments. **Phase 2: Model Development & Training:** Fine-tuning with curated astrobiology-specific datasets. Implement retrieval-augmented generation techniques to enhance knowledge recall and accuracy. Continuously validate and calibrate model outputs against expert-reviewed literature and validated scientific databases. **Phase 3: Validation & Benchmarking:** Performance evaluations using real-world astrobiological research queries and scenarios. Benchmark AB-Chat outputs against established scientific review and validation processes. Systematically address biases, refine model interpretability, and optimize user interaction for research efficiency. **Phase 4: Deployment & Expansion:** Integrate AB-Chat into existing research infrastructures and databases (*e.g.*, NASA's Planetary Data System). Develop user interfaces for hypothesis querying, literature synthesis, and real-time research support. Continue to expand training datasets with ongoing and upcoming astrobiology mission data to ensure the model remains cutting-edge and relevant.

We recommend that NASA invests resources into (defining/developing/supporting/funding) **foundational models** specifically **trained on large astrobiology datasets** that can then be used as community resources to further NASA goals in astrobiology.

References

1) Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

2) Szwarcman, Daniela, et al. "Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications." *arXiv preprint arXiv:2412.02732* (2024).

3) Shinde, Rajat, et al. "WxC-Bench: A Novel Dataset for Weather and Climate Downstream Tasks." *arXiv preprint arXiv:2412.02780* (2024).

4) Jheeta, S. et al., The way forward for the origin of life: Prions and prion-like molecules first hypothesis, *Life* 11(9), 872, (2021): https://www.mdpi.com/2075-1729/11/9/872

5) Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).