Zhuchang Zhan, Ph.D.

Research Scientist, MIT

• Cambridge, MA • (217) 979 3735 • zzhan@mit.edu
• LinkedIn • Github • Google Scholar

Objective: Looking for a data scientist / machine learning engineer position focused on building complex models to solve real-world problems and applying machine learning to large datasets.

- **Languages & Aptitudes:** Python, SQL, C++, Linux Shell, Latex, Tensorflow, Keras, Pytorch, Scikit-learn, Pymc3, nltk, pandas, h5py (HDF5), numpy, scipy, mongoDB, Git, Jupyter, matplotlib, seaborn, Tableau

- **Technical Skills: ETL, Regression (**Linear, Logistic, Elastic-net, Kernel, SVR, Random Forest), **Classification** (KNN, SVM, Naive Bayes), **Clustering** (K-means, Gaussian mixture, DBSCAN, Hierarchical), **Reduction** (PCA, t-SNE), **Deep Learning** (CNN, GAN, LSTM, NLP), Bayesian Inference, Monte-Carlo (Markov Chain, Hamiltonian)

EDUCATION

Massachusetts Institute of Technology Jun. 2015 - Feb. 2021
Ph.D., Physics (Subdiscipline: Planetary Sciences)

University of Illinois at Urbana-Champaign Sep. 2011 - May 2015
Bachelor of Science., Physics and Astronomy

DOCTORAL THESIS

Expanding Plausible Biosignature Gas Candidates

Advisor: Prof. Sara Seager

Searching for gases produced by alien life (biosignature gases) is a vital challenge of the century but only ~20 biosignature gases are identified in the last two decades.

- Developed a triage framework that enables the assessment of 10k+ gases (100x existing work) to identify plausible biosignature gas candidates and discovered 20+ new biosignature gases, 2x previous works combined.
- Pioneered infrared spectra prediction for millions of gases with no spectral data using a combination of maximum common substructure (**MCS**) and **hierarchical-clusters** of molecules with spectral data.

RESEARCH PROJECTS

NASA TESS Variable Star Detection Pipeline (Nature Publication)

Jun. 2017 - Present

The Transiting Exoplanet Survey Satellite (TESS) is a space telescope dedicated to find planets beyond our solar system.

- Constructed an **ETL** pipeline to reduce **terabytes** of **time series data** and discover new categories of variable stars (**periodic signal**) using **Fourier transform** and stellar flares (**anomalies detection**) using **PCA.** The pipeline is used in the discovery of planet remnant core and published in **Nature**.
- Led the discovery of ten “Complex Rapidly Rotating Mdwarf Stars" and applied **convolutional-neural network (CNN)** to discover 40x more in low temporal resolution data for prioritized follow-up observation.
- Developed an auto vetter to aggregate key stellar parameters, which reduced the human vetting process time by 80%. The auto vetter is regularly used by the TESS research team to validate the discovery of new planets and variable stars (over 100+ members and used in 10+ journal articles).
- Developed a flare (intensity spike) prediction model using **LSTM** with 90% + accuracy. The model enables multi-wavelength follow-up observations to be conducted 2 hr earlier than existing methods.

Discovery of Phosphine, A Possible Sign of Alien Life, on Venus (Nature Astro.)

Jun. 2018 - Present

Led the building and maintenance of the Seager Group Natural Product Database (1M+ molecules) and led the big-data and machine-learning driven analyses to study complex biological reaction networks. Communication with renowned scientists in a US-UK joint-collaboration to discover phosphine in the upper atmosphere of planet Venus.

- Developed analysis & validation tools for the discovery of phosphine on Venus and as a potential biosignature gas.
- Built multiple custom web scrapers using **selenium** and **rdkit** to extract over 1M+ molecule data and 100k+ biological pathways from 10+ sources and aggregate the data in a Cloud **SQL** database, which form the basis for assessing phosphine’s biological production in the Venusian atmosphere.
- Predicted molecular properties by testing **20+ regression models** and predicted molecules produced by life (natural products) using an **ensemble majority voting model** from the best predictors.

Habitable Exoplanet Atmosphere Spectra Model [Github] Jun. 2017 - Feb. 2020

Developed the first model that facilitates studying the 10,000+ gases life can produce (~1000x more molecules than modern methods) and model used in multiple recent peer-reviewed journal publications.

- Implemented atmosphere information retrieval with **Hamiltonian Monte Carlo (pymc3)**, which enables the retrieval with ~10x speed and > 10x free parameters compared to exisiting MCMC infomation retrieval methods.
- Led a team of 5 to construct a comprehensive gas spectra database (100,000+ gases) by scraping and merging multiple online resources using **beautifulsoup** and **selenium** and stored in **SQL** databases.
- Refactored and modularized legacy **C++** photochemistry code by swapping grid calculation with optimized matrix operations and added custom rapid data-block loader (**h5py)** to improved code efficiency by 260%**.** Added python wrapper to improve user readability and collaboration. Added auto-save and event logging to increase reliability.

Chinese Food Safety Evaluation Using NLP on Court Cases Jun. 2019 - Present

Collaboration with MIT Sloan FSAS Initiative to analyze the Chinese court case database using NLP.

- Identified 14,000+ unique Chinese agencies from millions of Chinese court documents using a combination of **empirical rules** and **NLP** (nltk, jieba). The documents contain unstructured, non-standardized, and often abbreviated **Chinese text** and characters.
- Applied **NLP** on the court documents to identify the lead function of each agency and classify the agencies as government or private. The new approach reduced human checking by 80 - 90%.

NASA Martian Greenhouse Sustainability Research Jun. 2018 - Sep. 2020

Advancing humans as a multi-planetary species on Mars is my lifelong aspiration. Combined my astrobiology research with establishing a self-sustainable greenhouse on Mars during the 2019 "NASA Big Idea Challenge".

- Led the biology simulations of the MIT (BEAVER) team and won **2nd** place (out of 100+ teams nationwide)
- Designed the *BEAVER* greenhouse's iconic spiral cone architecture and constructed a data-driven **simulation model** to study the crop harvest cycle and crop rotation. Display results using **matplotlib** and **seaborn**.
- Assessed the accumulation of 100+ trace gases produced by eight varieties of crops grown in an enclosed environment and study how some gas disequilibrium can be toxic and may have a long-term effect on inhabitants.

SELECTED PUBLICATIONS (20+ coauthor papers)

- **TESS Discovers a Remnant Planetary Core in the Hot Neptunian Desert** (2020) D Armstrong, *et al*., Didier Queloz (*2019 Nobel Physics Prize*), *et al*., ***Z. Zhan***. ***Nature*** (583 (7814), 39-42)
- **Complex Rotational Modulation of Rapidly Rotating M Stars Observed with TESS** (2019) ***Z. Zhan***, MN Günther, S Rappaport, K Olah, A Mann, AM Levine, et al., ***Astrophysical Journal***. (876 (2), 127)
- **Stellar Flares from the First TESS Data Release: Exploring a New Sample of M-Dwarfs** (2020) MN Günther, ***Z. Zhan***, S Seager, PB Rimmer, S Ranjan, et al., ***Astrophysical Journal,*** (159 (2), 60)
- **Assessment of Isoprene as a Possible Biosignature Gas in Exoplanets with Anoxic Atmospheres** (2021) ***Z. Zhan***, S Seager, J Pektowski, et al., ***Astrobiology*** (Accepted for Publication)
- **Detection of Phosphine Gas in the Cloud Decks of Venus** (2020) J Greaves, et. al., ***Z. Zhan***. ***Nature Astronomy*** 1-10
- **Persistence of Flare-Driven Atmospheric Chemistry on Rocky Habitable Zone Worlds (2021)** H Chen, **Z. Zhan**, A Youngblood, ET Wolf, AD Feinstein, DE Horton. ***Nature Astronomy***, 1-13

ACADEMIC HONORS AND AWARDS

- Ph.D. Fellowship awarded by NASA TESS Research Grants 2018-20
- Ph.D. Fellowship awarded by Amar G. Bose Research Grants 2015-18
- NASA Big Idea Challenge: Martian Greenhouse Design: 2nd Place out of 100+ university teams Apr. 2019