The Need for Speed (and Accuracy): Surrogate Models in Exoplanet Atmospheric Characterization

Anshuraj Sedai, Dr. Prajwal Niraula, Dr. Sai Ravela

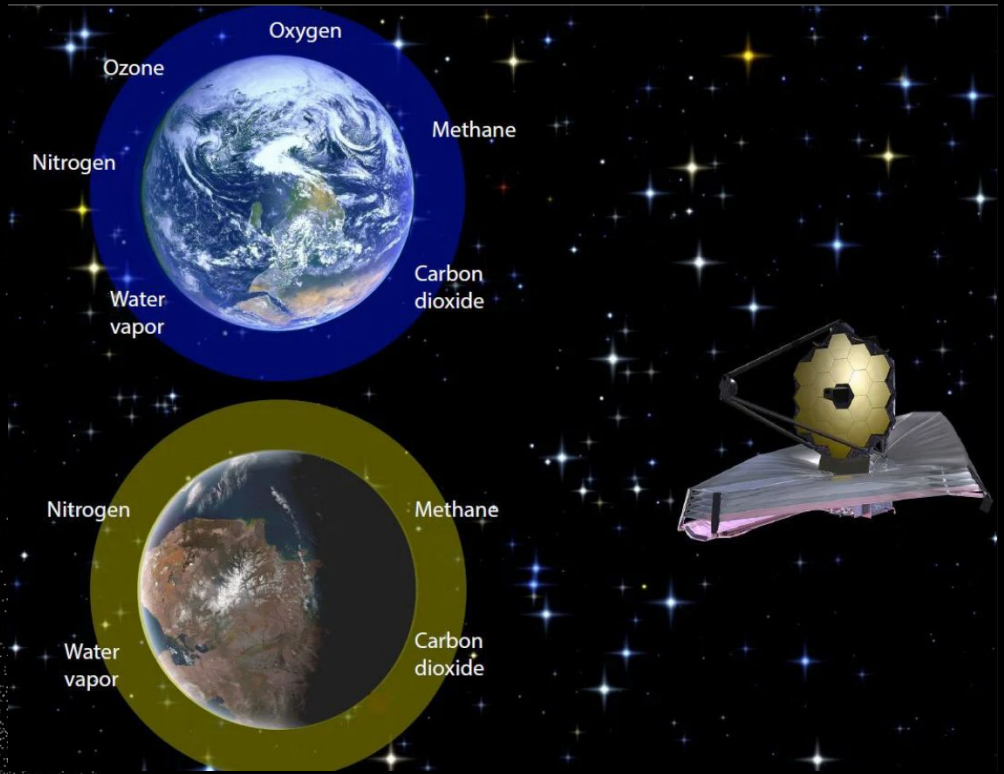
CALDWELL UNIVERSITY

Massachusetts Institute of Technology

MSRP MIT Summer Research Program

What are exoplanets and why do we study their atmospheres?

- Exoplanets: Planets that orbit stars outside our solar system
- Understanding the atmospheric conditions is crucial in determining the planet's habitability [1]
  - Find and quantify the constituents of the atmosphere and planet's surface
- But there is only so much we can do with observations and the vast amount of data they produce

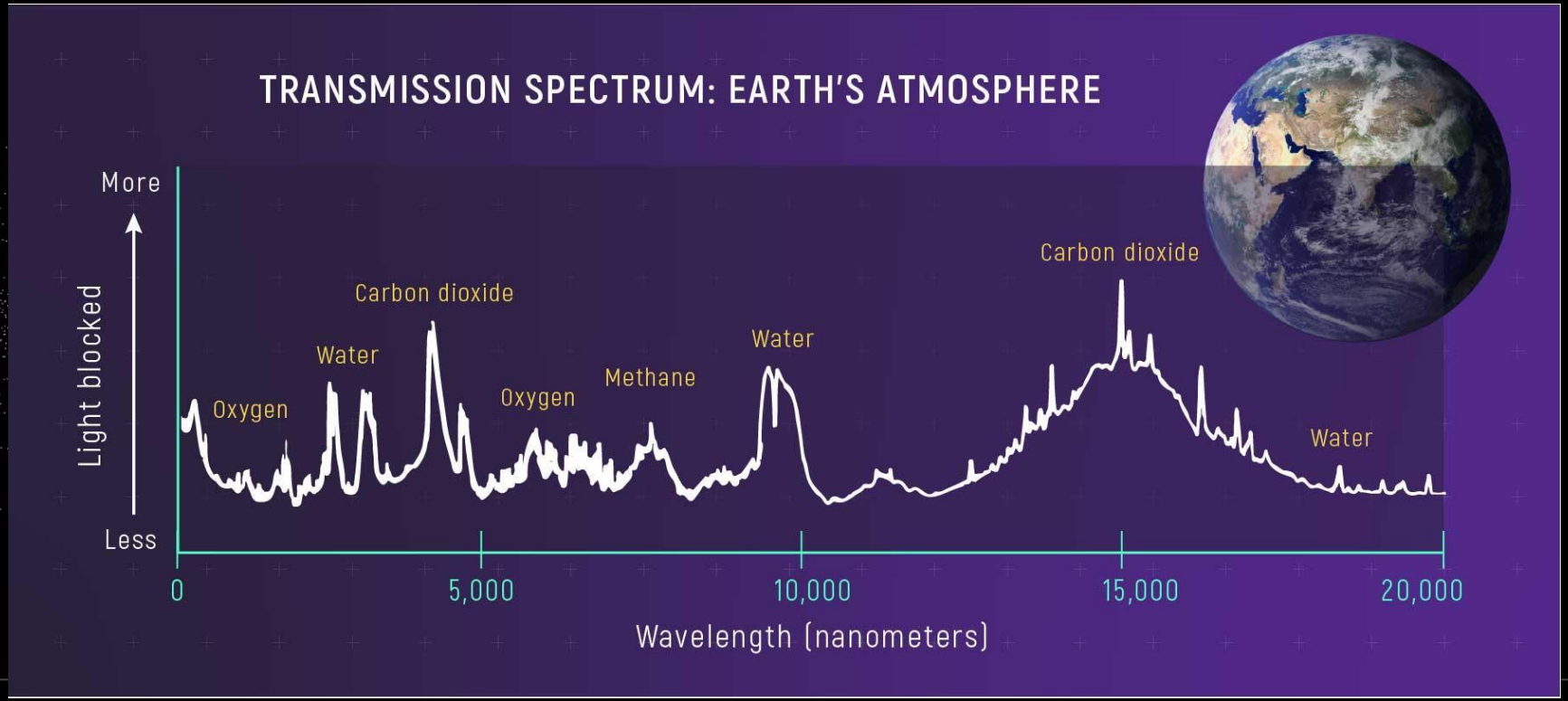


Credit: NASA/Wikimedia Commons/Joshua Krissansen-Totton

[1] Madhusudhan & Seager 2009, ApJ 707, 24.

How do we study their atmospheres?

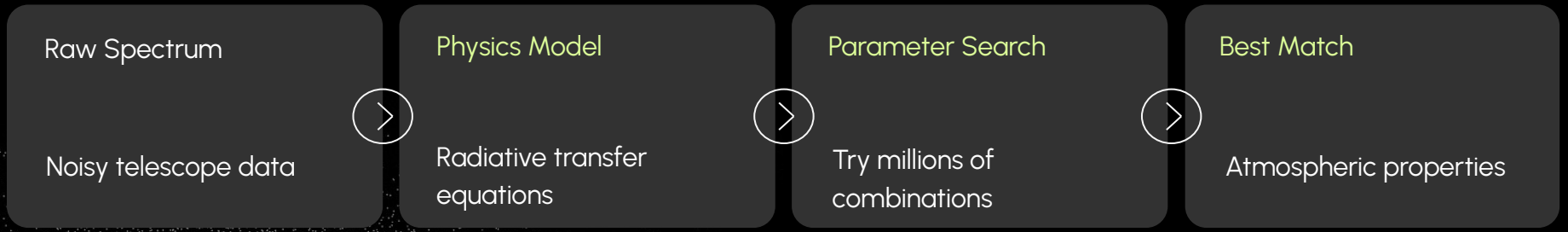
We analyse the light that passes through or comes from their atmospheres.



Credit: NASA, ESA, and L. Hustak (STScI).

How do we gain insights from it and the challenge?

Getting from raw spectrum to atmospheric composition is incredibly complex.



- **Complex Physics:** Must solve radiative transfer through entire atmosphere
- **Many Parameters:** Temperature, pressure, molecular abundances at many layers
- **Statistical Sampling:** Need millions of model runs to find best fit
- **High Precision Required:** Tiny signals need careful analysis

Result: Very accurate results based on solid physics! [2]

Hours to days per planet!
Computationally expensive
Limited for large surveys

[2] Himes et al. 2022, PSJ 3, 91.

Enter Machine Learning Surrogate Models

Simplified approximations of more complex, higher-order models

The main idea:

- Train neural network on thousands of synthetic atmospheric spectra
- Learn patterns between atmospheric properties and light signatures
- Make rapid predictions on new observations
- From hours to minutes! [3]

But can we maintain accuracy while gaining speed?

[3] Benneke & Seager 2012, ApJ 753, 100.

Our approach

- Step 1: Generate Training Data
  - Generated synthetic atmospheric spectra using Markov Chain Monte Carlo (MCMC) coupled with radiative transfer code, petitRADTRANS [3]

Credit: petitRADTRANS

- Step 2: Train and Optimize Neural Network
  - Trained separate transmission- and emission only neural surrogate models using MARGE library on the synthetic spectra dataset [2] (Took us ~9.5 minutes)
  - **Fine-tuned the model to handle noisy data and avoid generalization**
- Step 3: Test on Real Exoplanets
  - Compared synthetic transmission and emission spectra predictions with observational data of exoplanets HD 189733 b and GJ 1214 b [4, 5]

[2] Himes et al. 2022, PSJ 3, 91. [3] Benneke & Seager 2012, ApJ 753, 100. [4] Kempton et al. 2023, Nature 620, 67. [5] Zhang et al. 2025, AJ 169, 38

What did we find? (1 of 3)

Transit depth → planet blocked starlight

Model-data mismatch

Model does not fit well

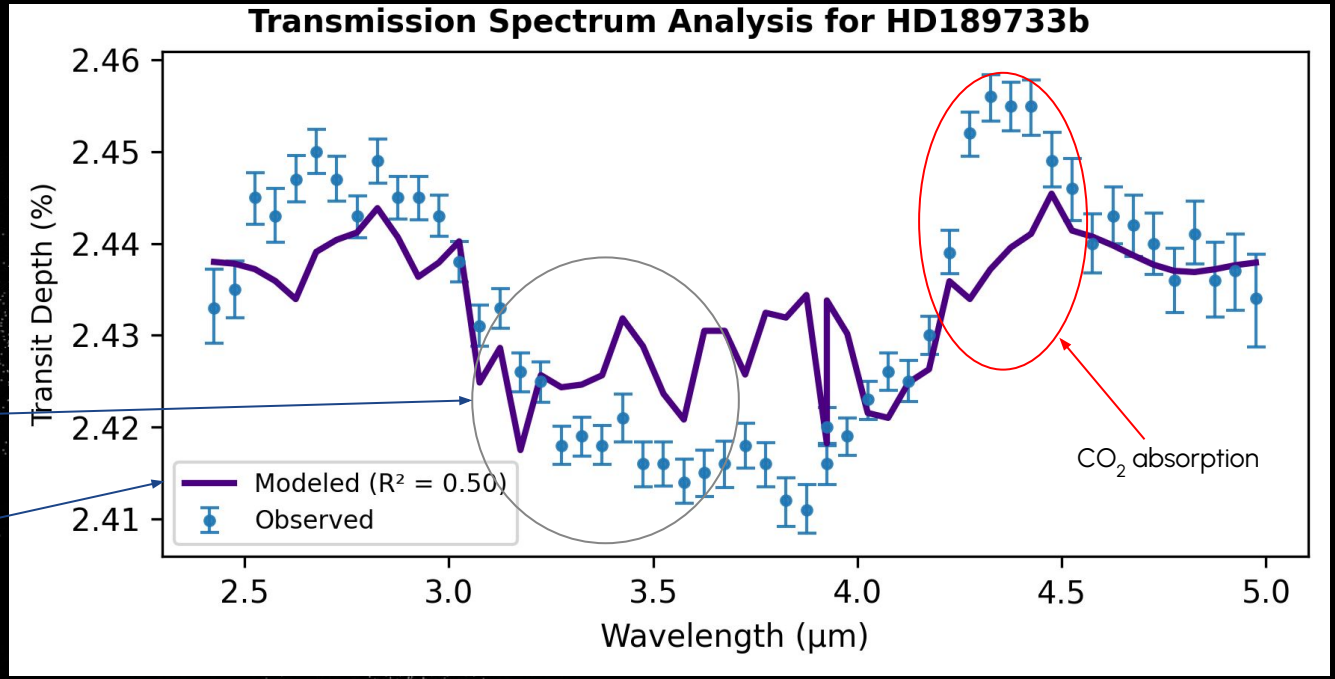


Fig. 1: Transmission spectra for HD 189733 b (2.5-5.0 μm). Model (R^2=0.50) vs. observed atmospheric absorption data.

What did we find? (2 of 3)

Higher R^2, stronger model fit

Secondary eclipse → planet's thermal emission

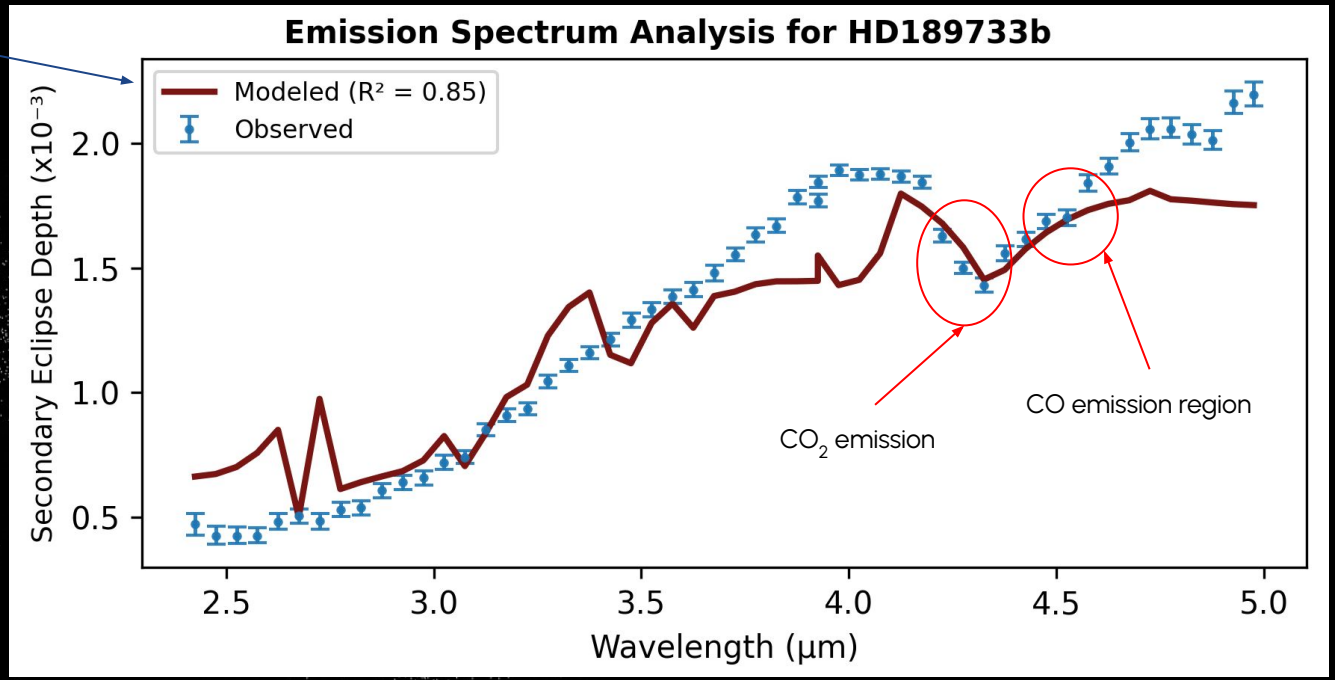


Fig. 2: Emission spectra for HD 189733 b (2.5-5.0 μm). Model (R^2=0.50) vs. observed thermal emission data.

What did we find? (3 of 3)

Higher R^2, stronger model fit

Secondary eclipse → planet's thermal emission

Four modeling approaches

Model spread

All models agree

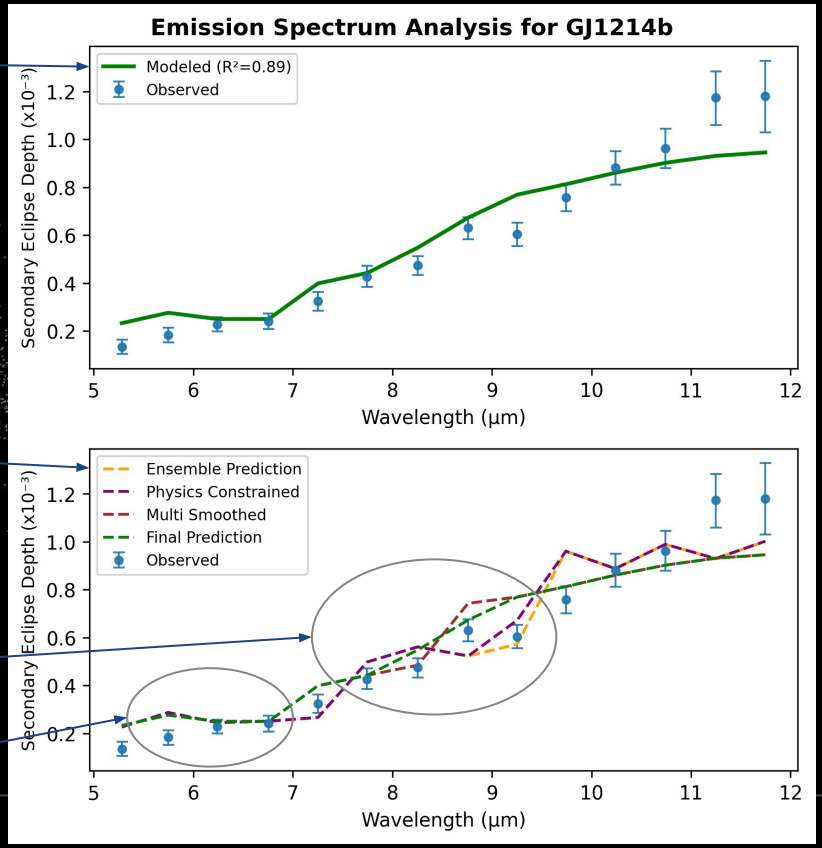


Fig. 3: Emission spectrum for GJ 1214 b (5-12μm). Top: Best-fit model vs. observations. Bottom: Comparison of four modeling approaches (ensemble, physics-constrained, multi-smoothed, final prediction) with observational data

What does this imply?

- Surrogate models accurately recover emission spectra (R^2 = 0.85-0.89) but show moderate performance for transmission spectra (R^2 = 0.50).
- Physics-informed constraints (radiative transfer principles, molecular opacities) and ensemble methods improve model reliability.
- Models successfully identify atmospheric molecules (CO_2, CO) in both emission and transmission observations.
- Results demonstrate surrogate models as fast, reliable tools for emission retrieval (~9.5 min vs. hours), with transmission modeling requiring further development.

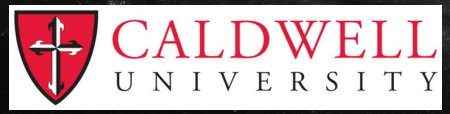
[2] Himes et al. 2022, PSJ 3, 91. [3] Benneke & Seager 2012, ApJ 753, 100. [4] Kempton et al. 2023, Nature 620, 67. [5] Zhang et al. 2025, AJ 169, 38

What's next?

- Incorporate physics-based constraints to improve transmission modeling and generalize across diverse exoplanet types.
- Work on model interpretability
- Develop user-friendly tools for astronomers

Thank you for your attention!

Any questions?



Bonus Slides!

Model Architecture

Neural Network Design:

- 2 Convolutional layers (pattern recognition)
- 4 Dense layers (decision making)
- ReLU activation functions
- Spectral-weighted loss function

Training Details:

- 14,000 total spectra (10k train, 2k validation, 2k test)
- 250 training epochs
- Channel-wise normalization
- Targeted regularization to prevent overfitting

Why is transmission spectroscopy harder?

Emission (Easier)

- Strong signal from hot planets
- Direct thermal glow
- Less atmospheric complexity
- More predictable patterns

Transmission (Harder)

- Weak signal - tiny atmospheric layer
- Complex scattering effects
- Cloud/haze contamination
- Limb temperature variations

Atmospheric characterization is the key to finding habitable worlds.

Biosignature Gases We Look For:

- Water (H_2O): Essential for life as we know it
- Oxygen (O_2): Produced by photosynthesis
- Ozone (O_3): Protects surface from UV radiation
- Methane (CH_4): Potential biological origin

Why Speed Matters:

- Prioritize targets for detailed follow-up
- Rapid assessment of potentially habitable worlds
- Enable systematic surveys of planetary populations