A Machine Learning Journey in a World of Alien Planets

Konstantin Matchev

With: K. Matcheva, R. Forestano, E. Unlu, A. Roman, E. Panek

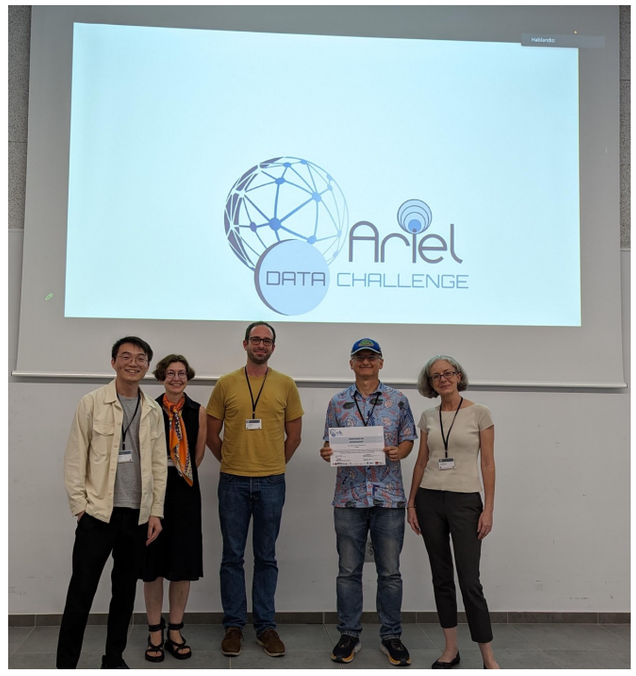
University of Rome

May 28 2025

How did I get into all this?

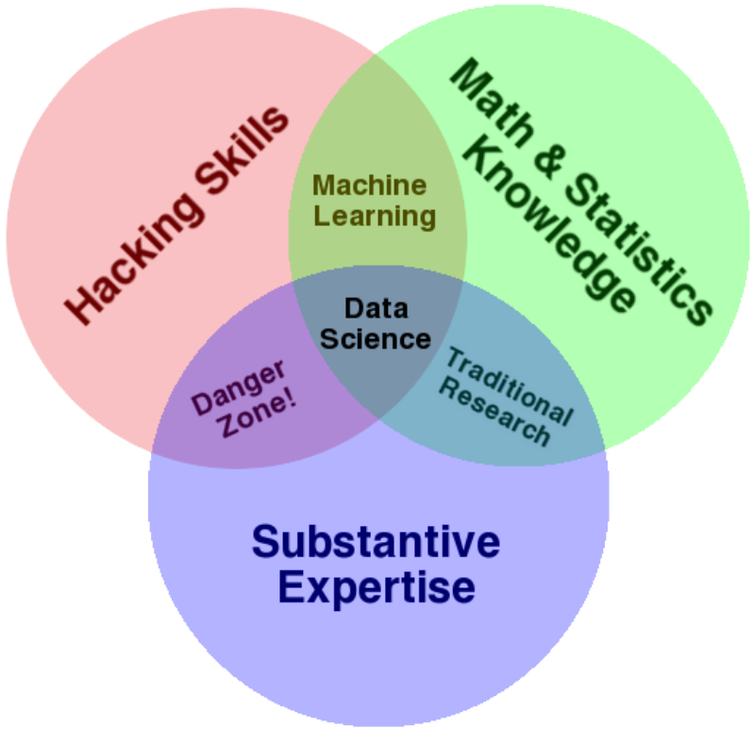
- In May 2020 the University of Florida announced a major AI Initiative
  - Many new AI courses*
  - Albert became AI
  - We became the GAItors

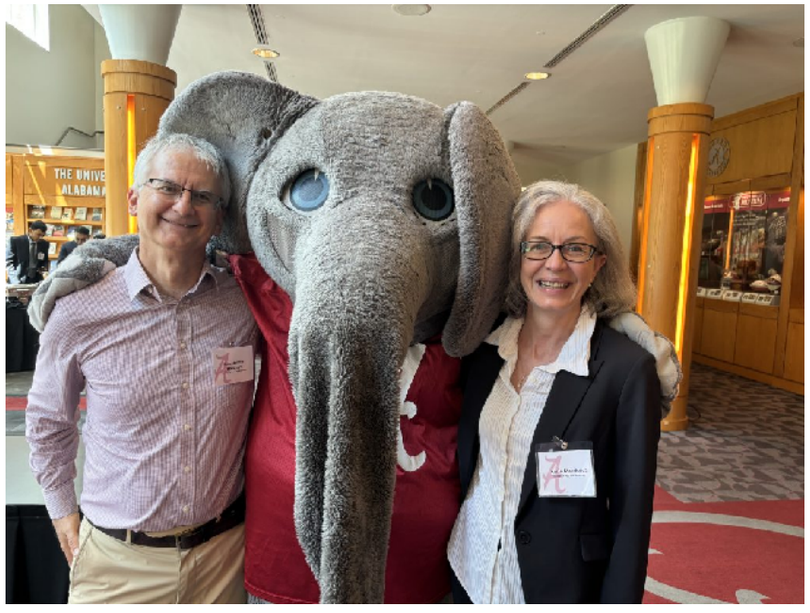
- AI and ML are bringing together the different disciplines
  - Common language; collaboration opportunities
- Ariel Machine Learning Data Challenge at NeurIPS 2022: 1st prize

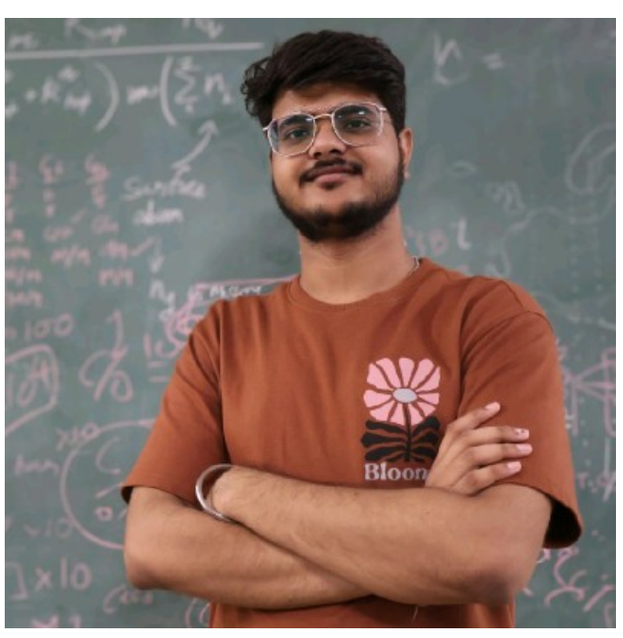


*www.phys.ufl.edu/~matchev/PHY7097_Fall2022/

Our EXO Group



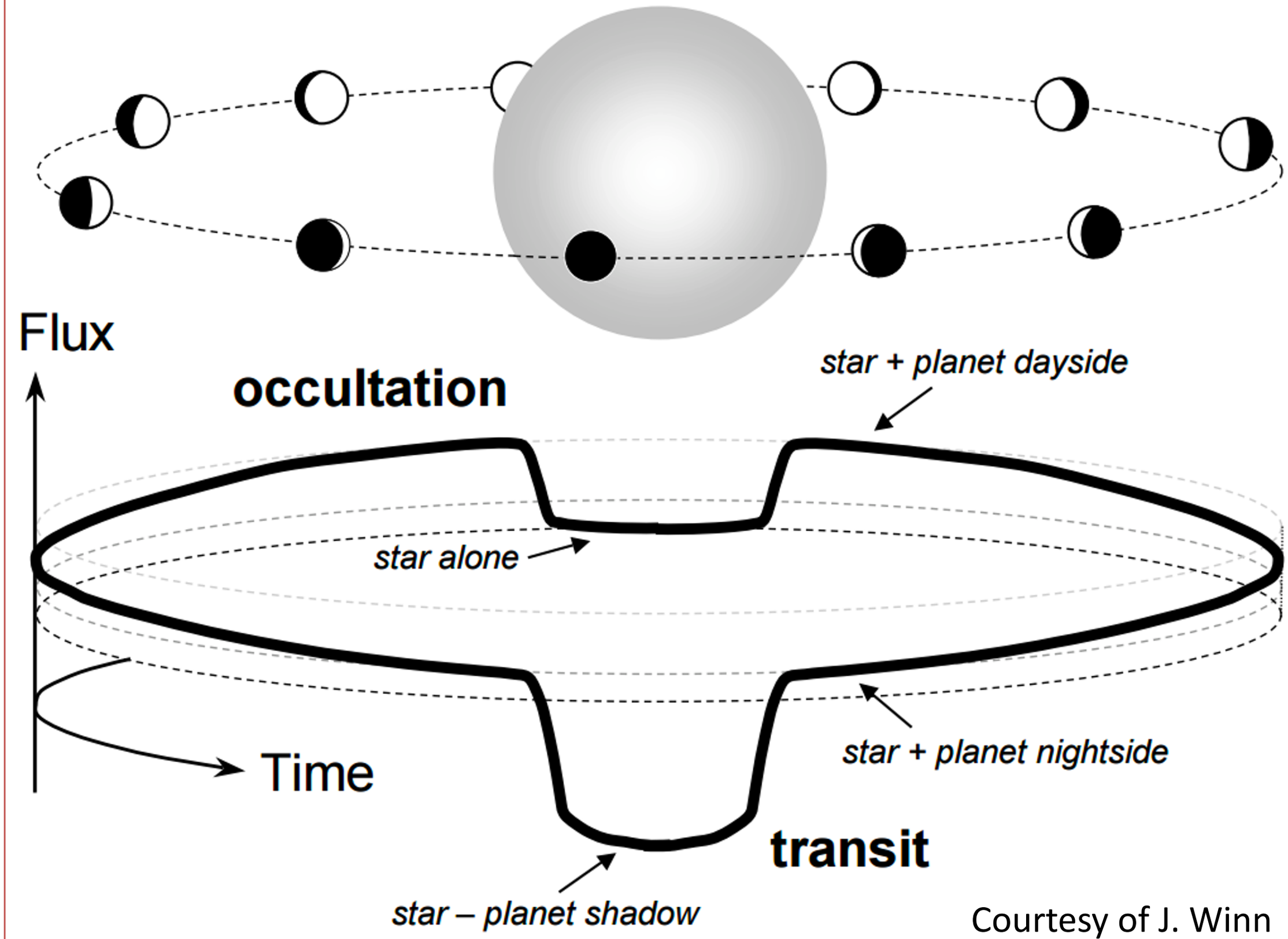




Outline/Topics

- Transmission Spectroscopy
  - Forward radiative transfer models
  - Inverse problem: analytical/Bayesian/ ML methods
- Machine Learning for Exoplanets
  - Supervised Learning
    - Ariel Data Challenge: 1st prize!
  - Symbolic Learning
    - Symbolic regression with PySR
  - Unsupervised Learning
    - dimensionality reduction
    - anomaly detection

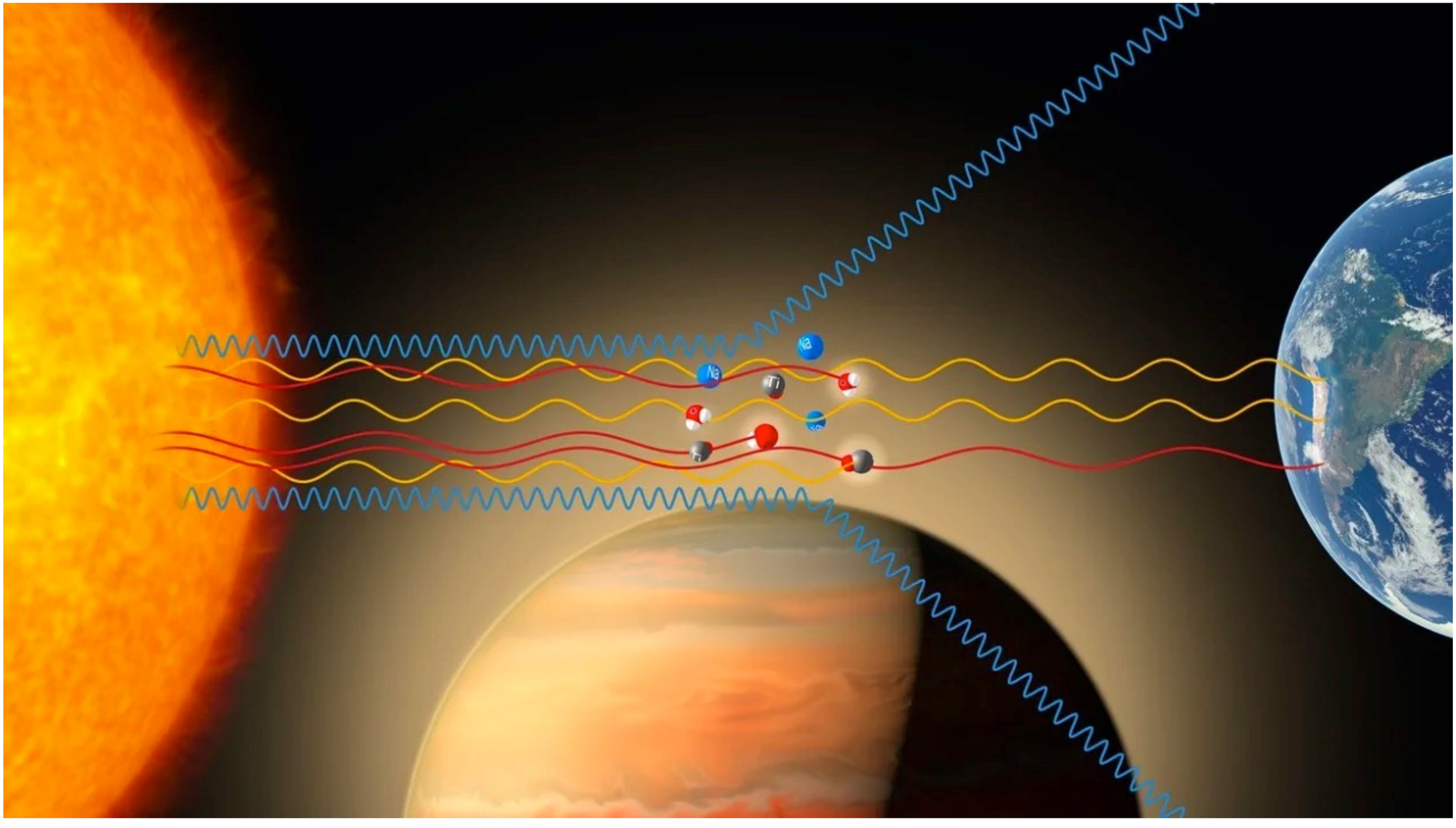
Planetary transits

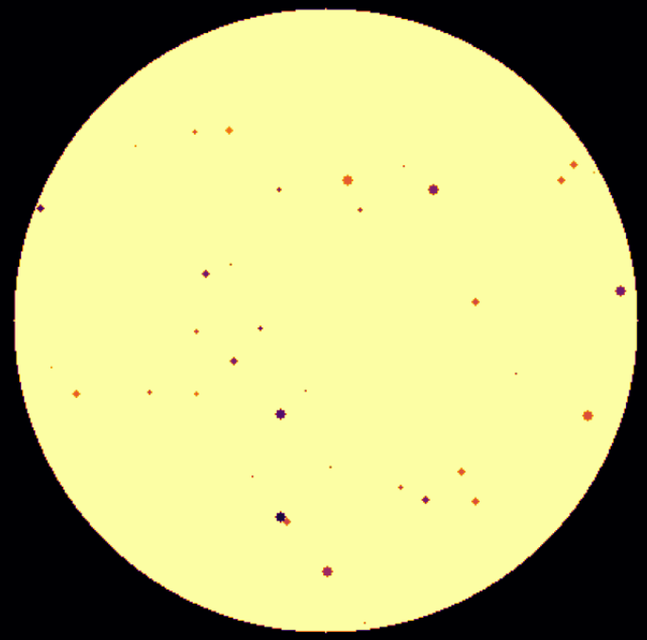


Courtesy of J. Winn

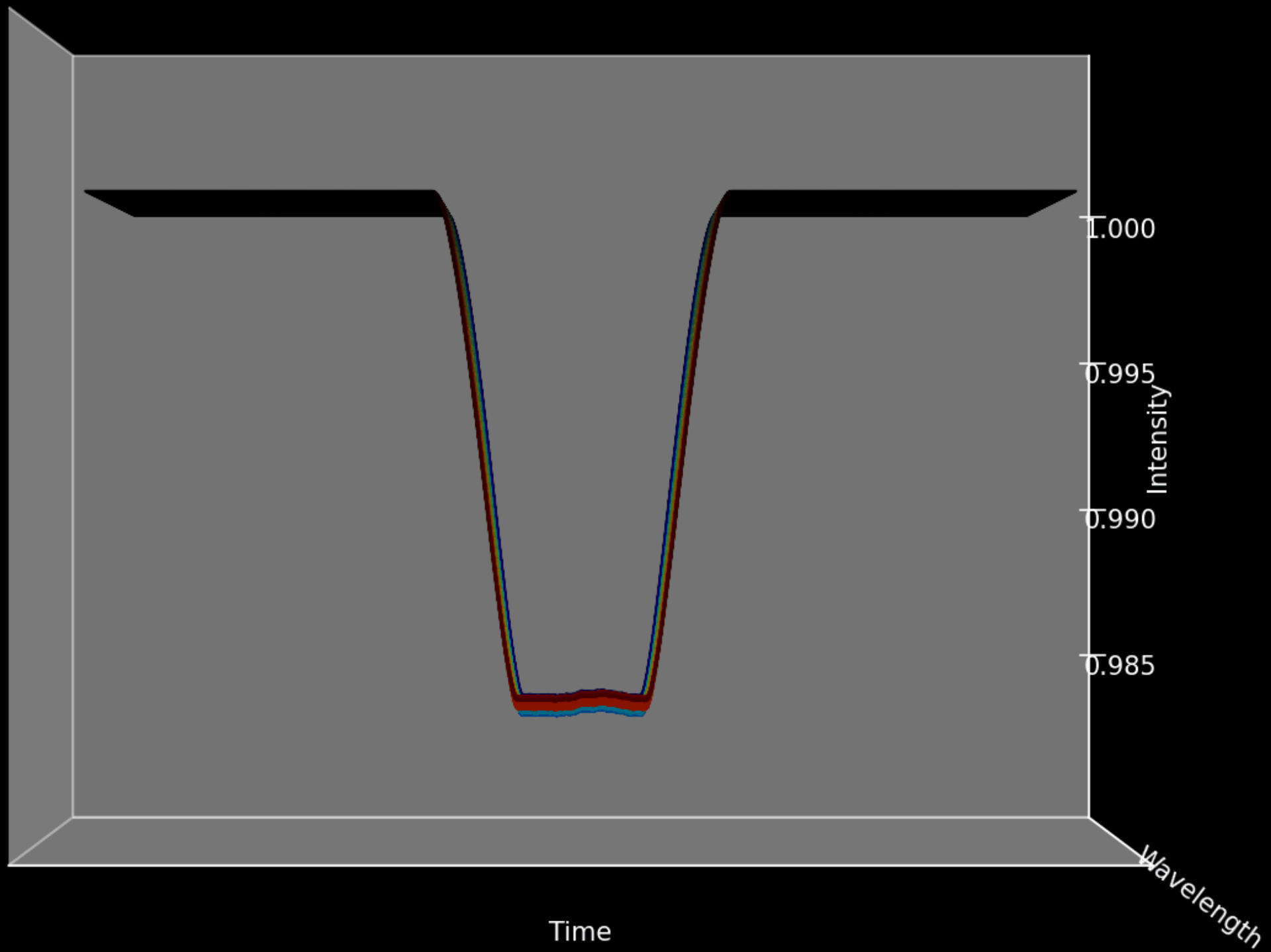
Transmission spectroscopy

- The observed transit depth depends on the wavelength
  - The spectrum “knows” about the chemical composition of the atmosphere

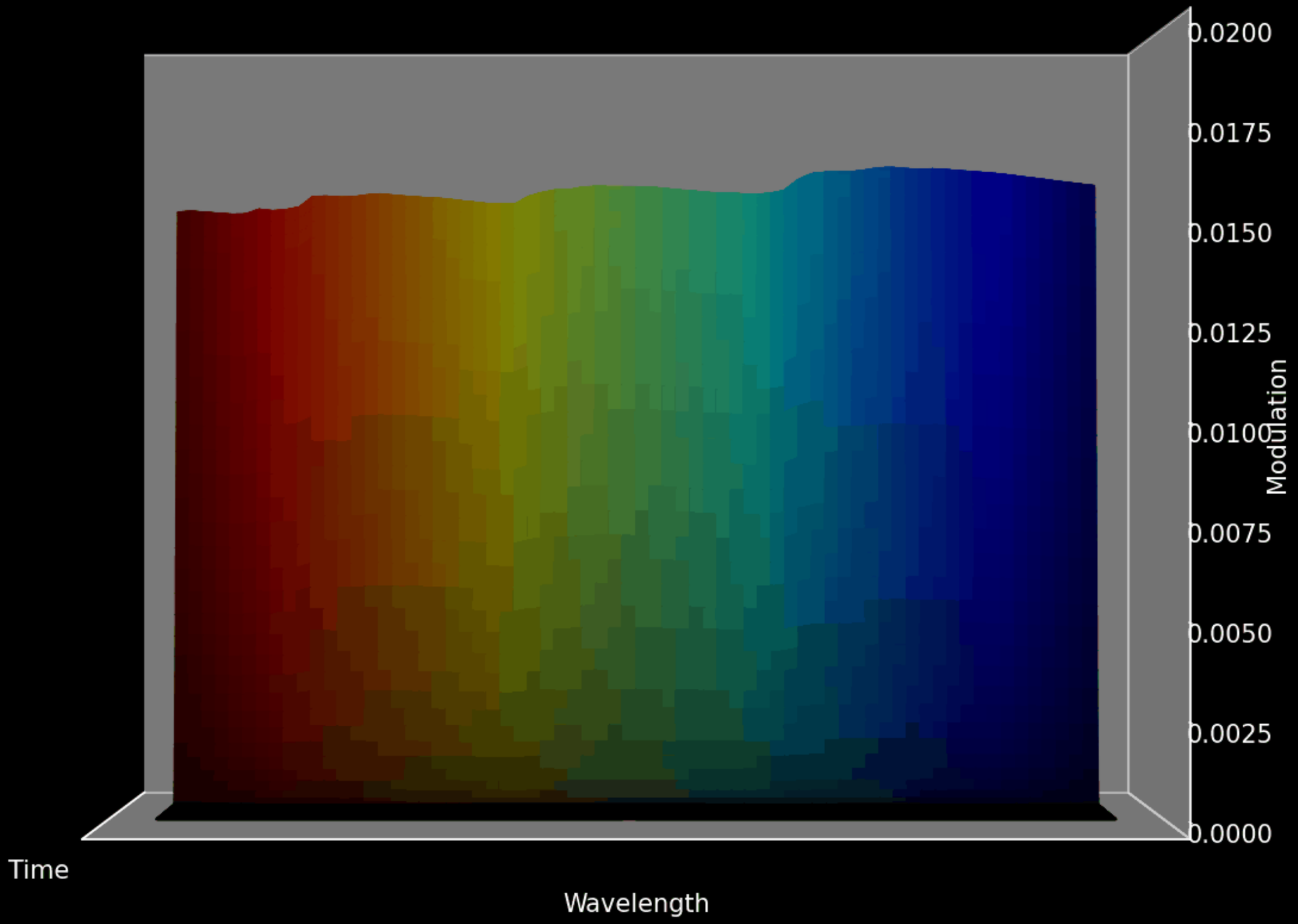




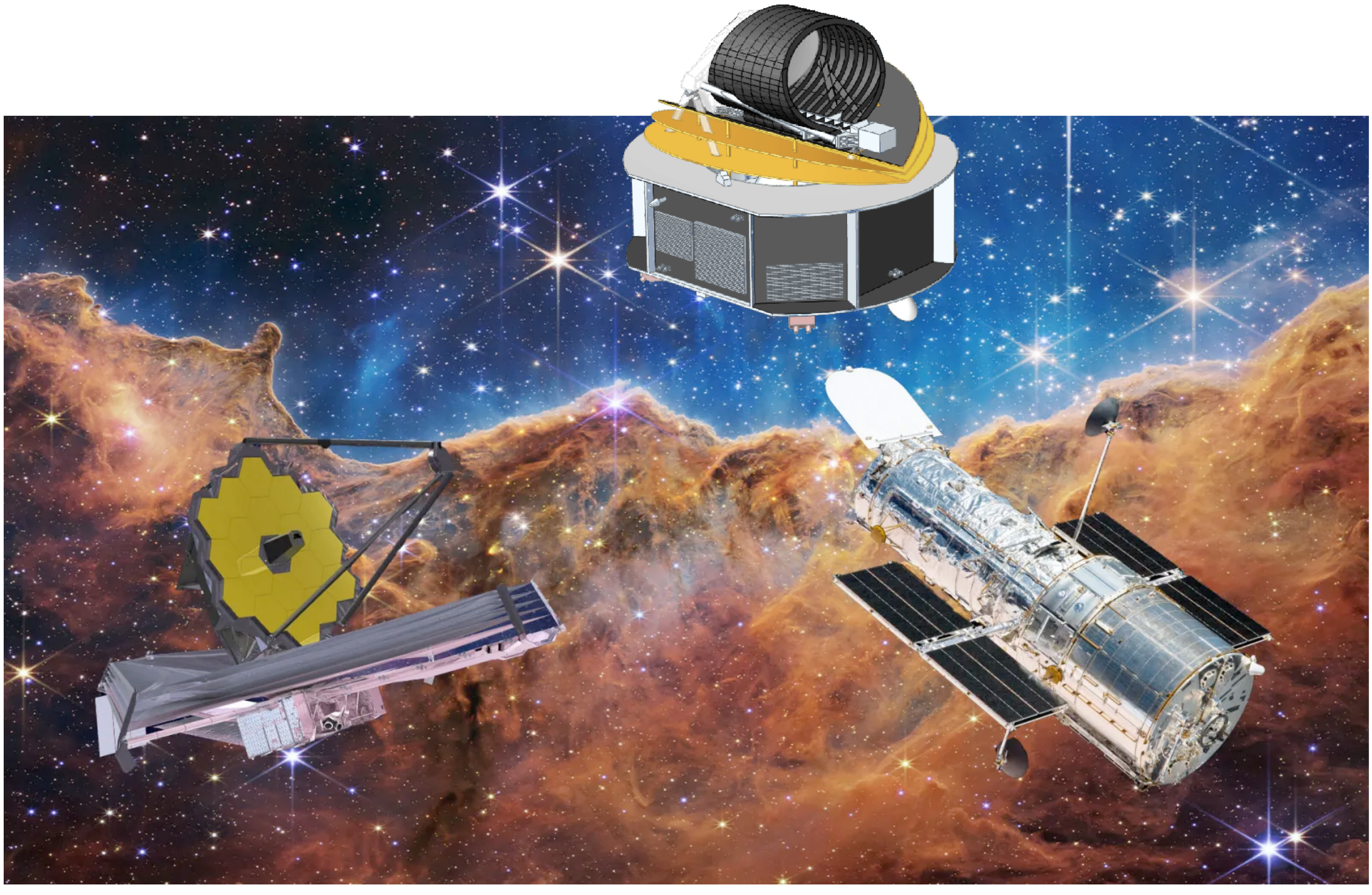
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
Intensity
Time


1.000
0.995
0.990
0.985
Intensity
Time
Wavelength

$$M(\lambda) = (R_T/R_S)^2$$

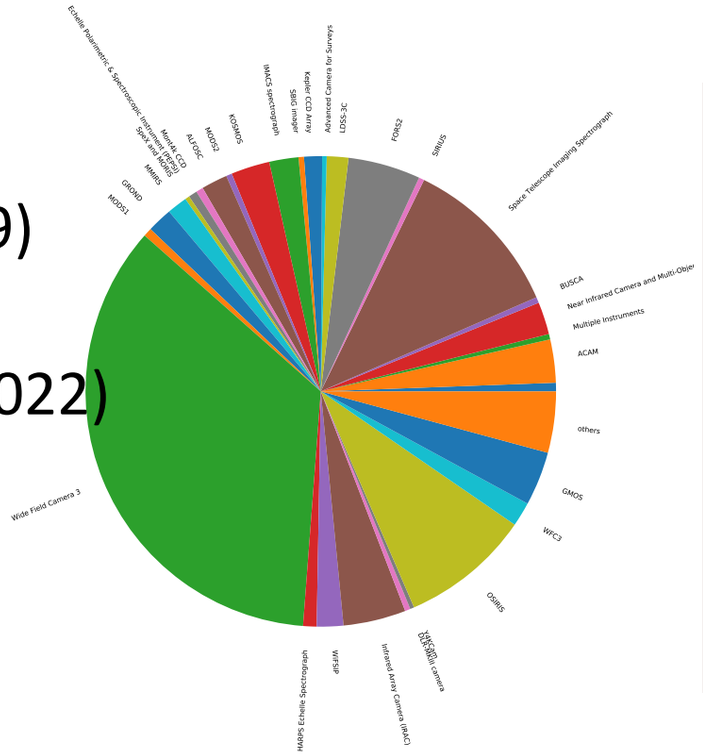


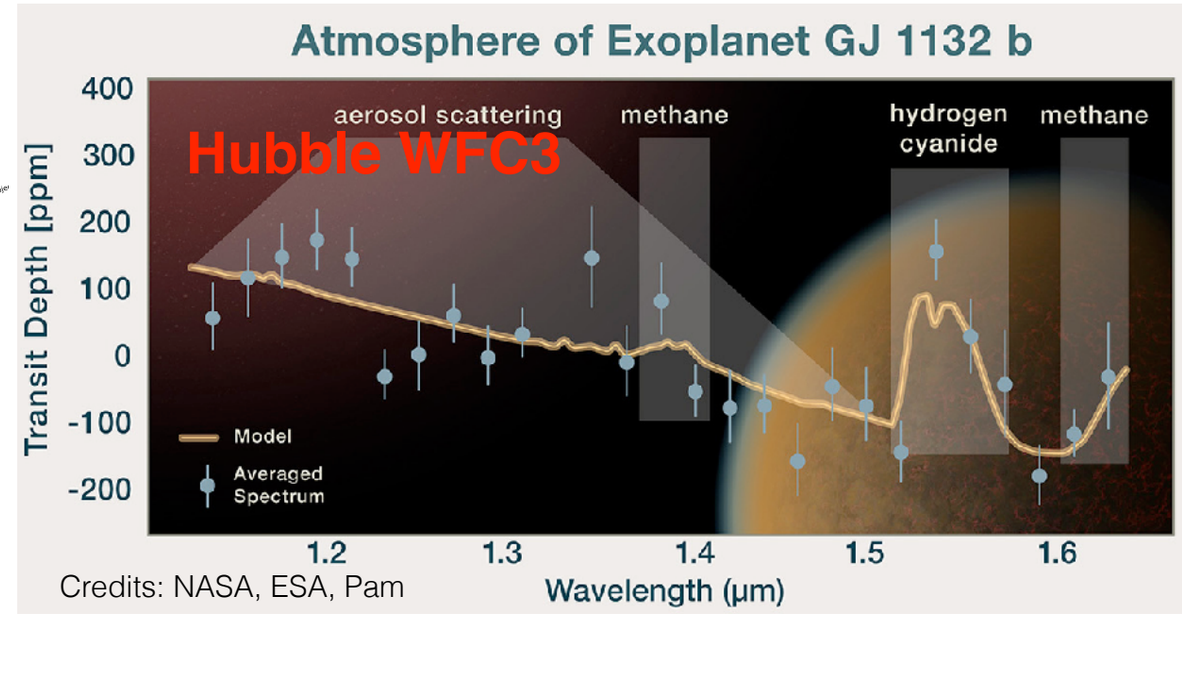
Eyes on the Sky



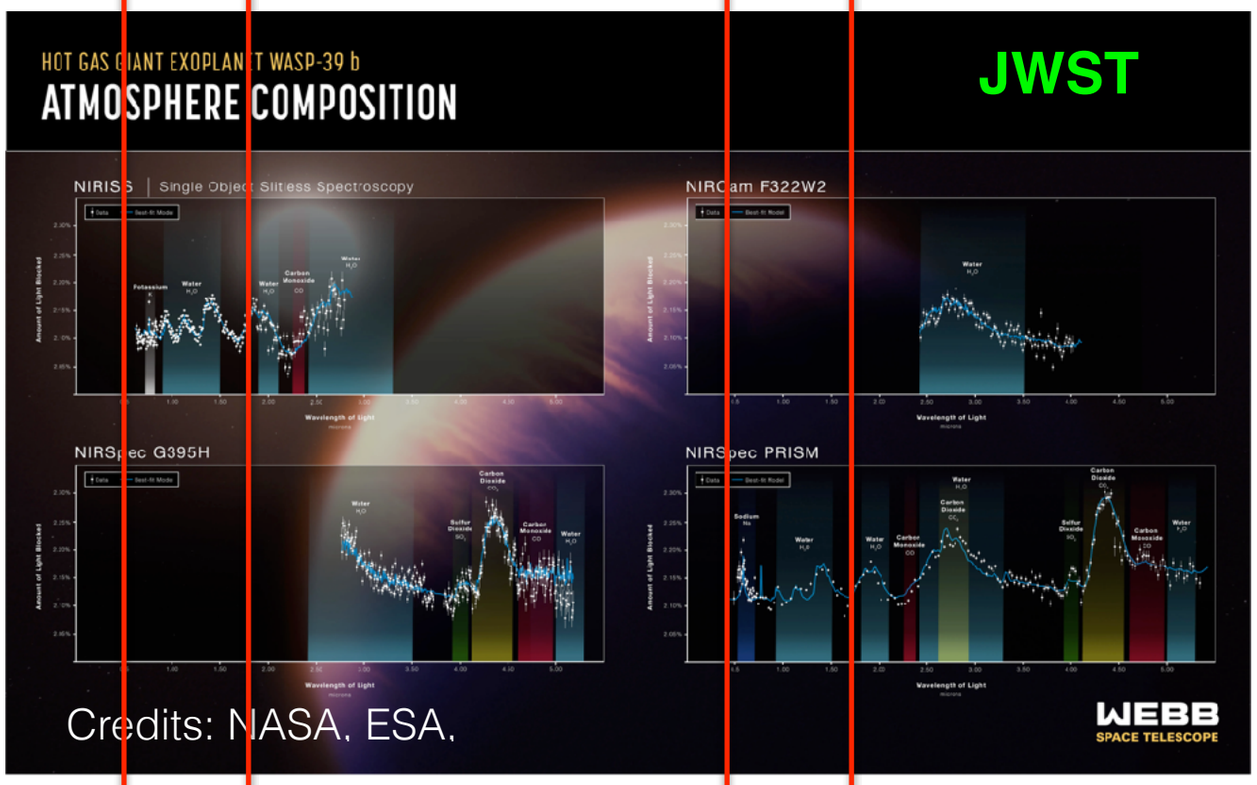
Expanding Horizons

- Hubble WFC3
  - 0.8-1.7 mic (2009)
- JWST
  - 0.5-5.2-28 mic (2022)
- Ariel
  - 0.5-7 mic (2029)

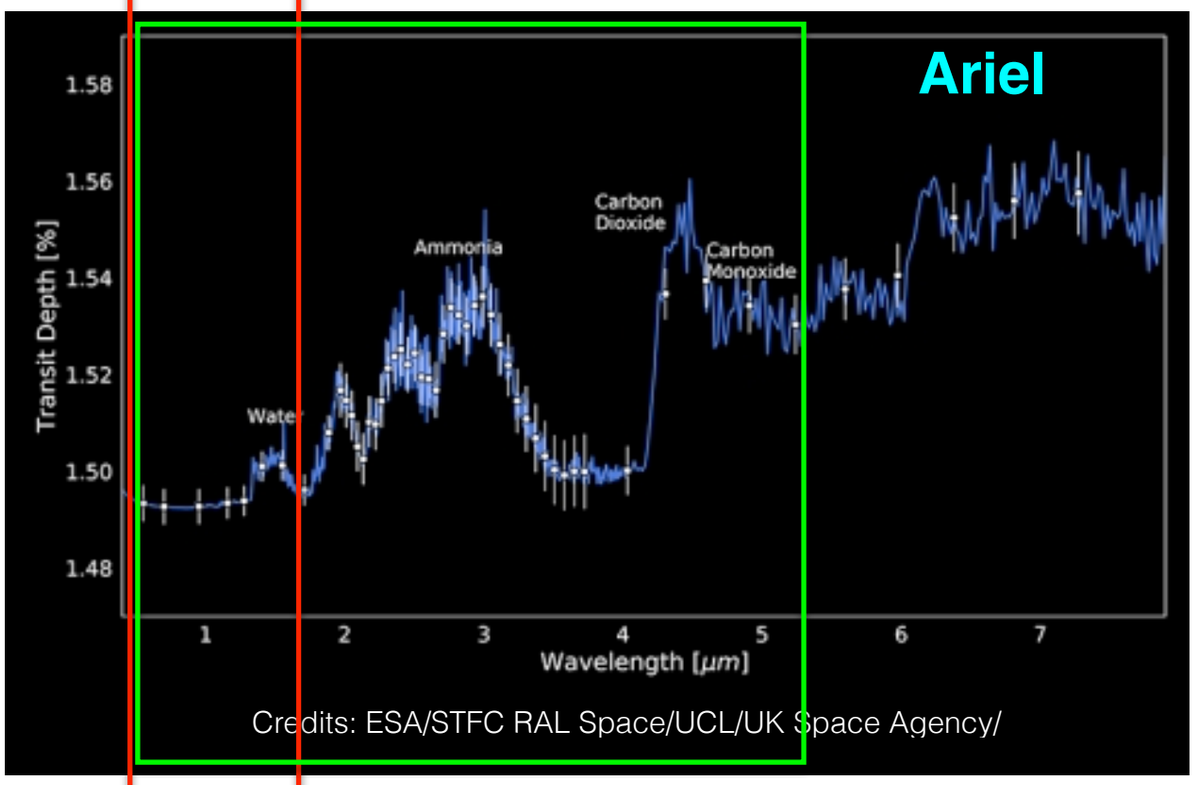




Credits: NASA, ESA, Pam

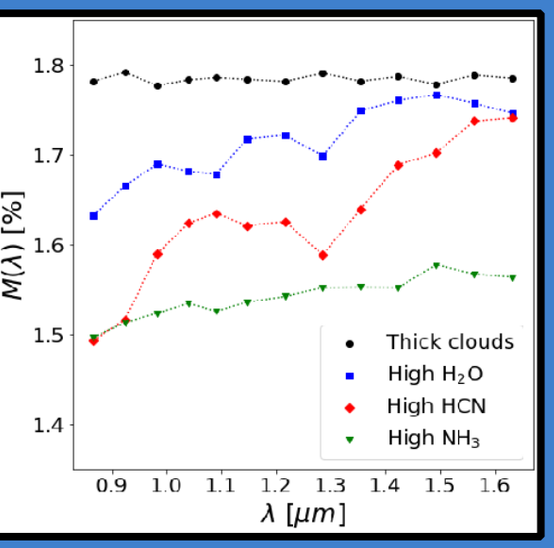


Credits: NASA, ESA,



Credits: ESA/STFC RAL Space/UCL/UK Space Agency/

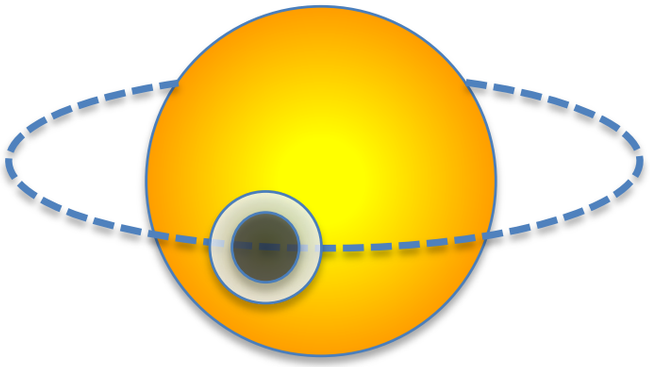
Transit Transmission Spectroscopy



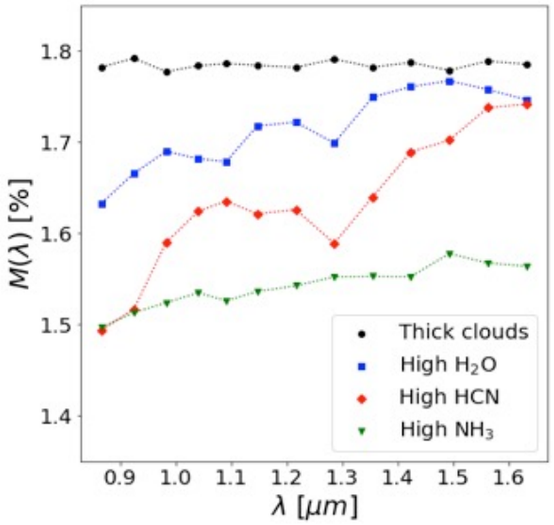
$$R_T(\lambda) = R_0 \left\{ 1 + \frac{H}{R_0} \left[\gamma_E + \ln\left(\frac{P_0 \kappa(\lambda)}{g} \sqrt{2\pi \frac{R_0}{H}} \right) \right] \right\}$$

Heng et al. 2017

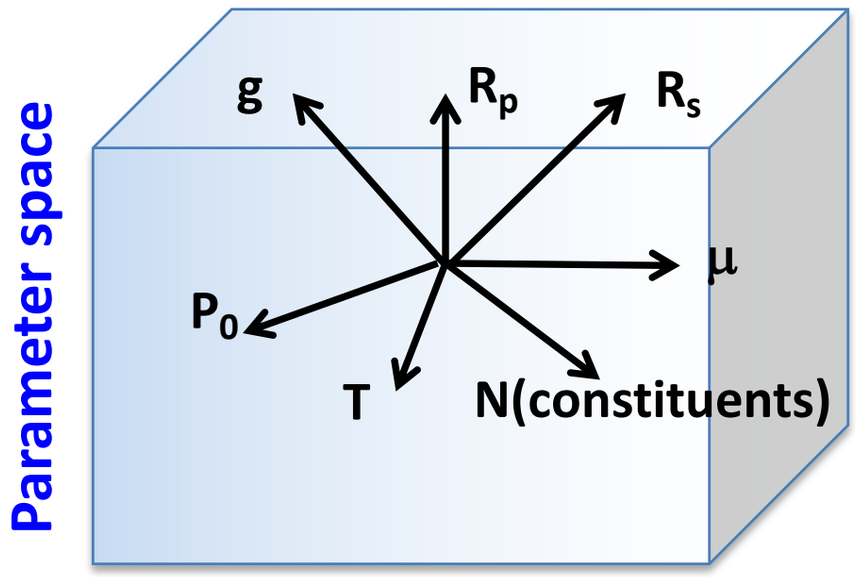
Exoplanet Atmospheric Retrievals



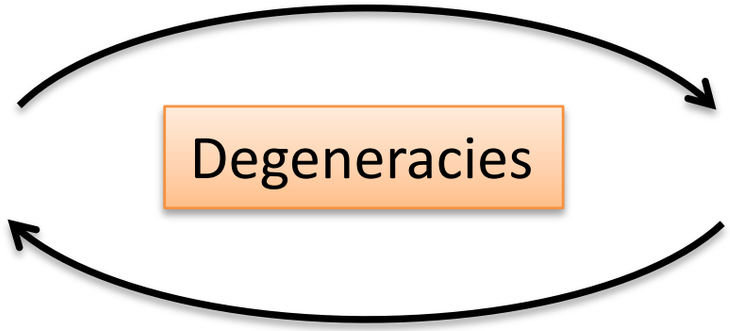
Observation



$$R_T(\lambda) = R_0\left\{1 + \frac{H}{R_0}\left[\gamma_E + \ln\left(\frac{P_0\kappa(\lambda)}{g}\sqrt{2\pi\frac{R_0}{H}}\right)\right]\right\}$$

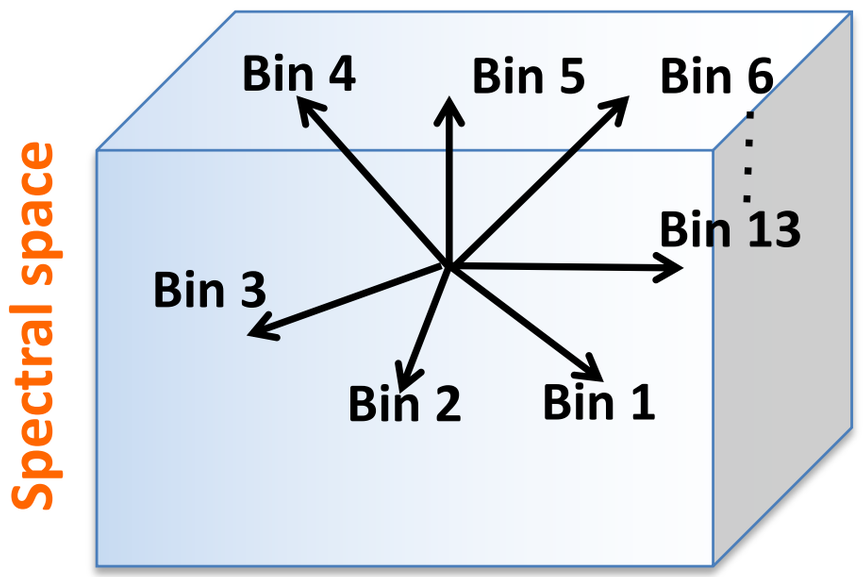


Forward model



Degeneracies

Retrieval model



Dimensionality Reduction

$\pi_1, \pi_2, \pi_3, \pi_4$

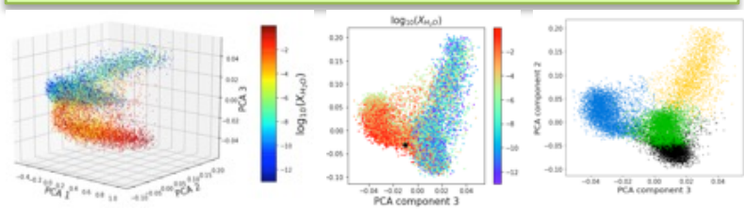
$$M(\lambda;\pi_1',\pi_2,\pi_3) = \pi_3^2\left\{1 + \frac{1}{\pi_1'}\left[\gamma_E + \ln\left(\pi_2\sqrt{2\pi\pi_1'}\right)\right]\right\}^2$$

Matchev et al 2021, 2022, 2022

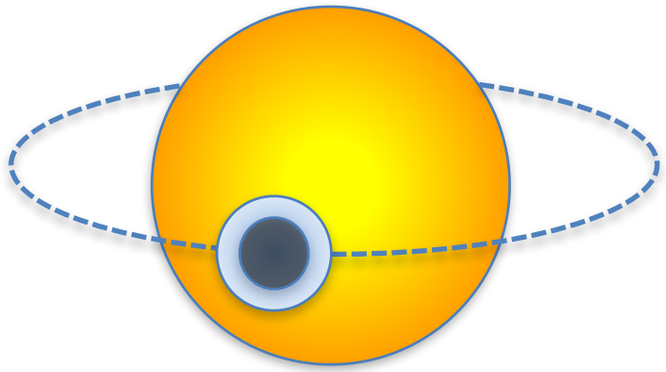
Forward model

Retrieval model

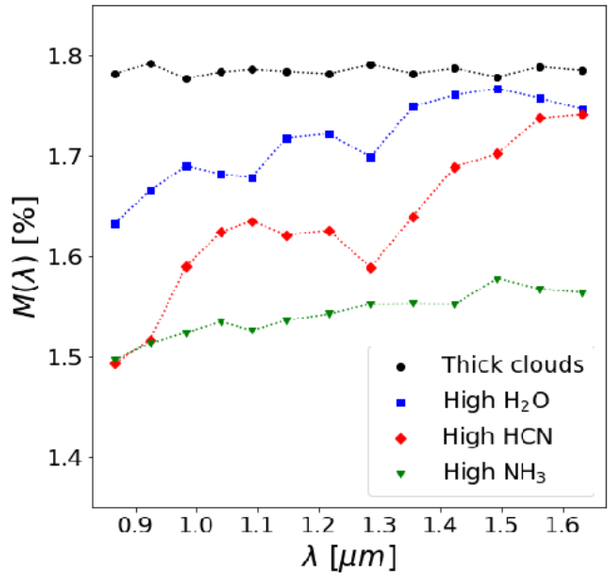
PCA, Isomap, AE, …



Forward Radiative Transfer Models



Observation



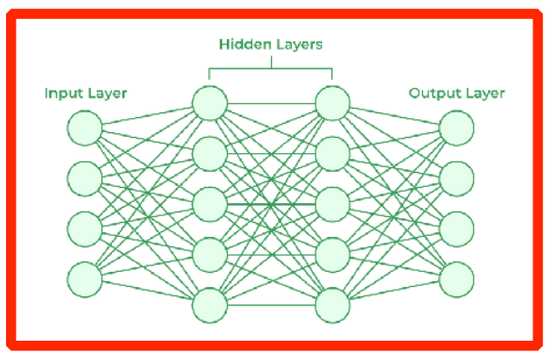
Inputs (features)

Planet	T (K)	R_p (R_J)	R_s (R_{sun})	X_{H2O}
1	1300	1.8	1.6	10^{-3}
2	650	0.9	1.4	10^{-2}
3	960	1.9	2.3	10^{-4}
4	1150	2.0	1.5	10^{-5}

Analytical

$$R_T(\lambda) = R_0\left\{1 + \frac{H}{R_0}\left[\gamma_E + \ln\left(\frac{P_0\kappa(\lambda)}{g}\sqrt{2\pi\frac{R_0}{H}}\right)\right]\right\}$$

RT model: TauREx



Outputs (targets)

Planet	M_1 %	M_2	M_3	M_4
1	1.41	1.44	1.42	1.52
2	0.52	0.55	0.61	0.58
3	0.92	1.03	1.11	0.95
4	1.85	1.94	1.99	1.82

- Generate a training database of spectra **M** by scanning over the **input parameters** for the forward model.

Meet the DATA!

- **HELA** database (Márquez-Neila P. et al., 2018, Nature,2, 719)

We use a public database[1] of 100,000 synthetic atmospheres created with an **analytical formula**:

- **Fixed parameters**: gravity, mean molecular mass, planetary radius, star radius, reference pressure (WASP-12b)
- **Scanned parameters**:
  - ✓ Temperature: 500 – 2900 K
  - ✓ H_2O volume mixing ratio: 10^{-13} – 1
  - ✓ HCN volume mixing ratio: 10^{-13} – 1
  - ✓ NH_3 volume mixing ratio: 10^{-13} – 1
  - ✓ Cloud opacity: 10^{-13} – 10^{2}
- Noise floor of 50 ppm on the transit depth (WFC3-like).
- **Spectral range**: 0.838-1.666 μm in 13 bins.

- **TRANSIT** database (with M. Himes, J. Harrington UCF)

We use full forward radiative transfer model (**TRANSIT**) with variable gravity , g, and self consistent mean molecular mass, μ.

- **Fixed parameters**: planetary radius, star radius, pressure grid of 100 layers
- **Scanned parameters**:
  - ✓ Temperature: 500 – 2900 K
  - ✓ H_2O volume mixing ratio: 10^{-13} – 10^{-2}
  - ✓ HCN volume mixing ratio: 10^{-13} – 10^{-2}
  - ✓ NH_3 volume mixing ratio: 10^{-13} – 10^{-2}
  - ✓ Cloud opacity: 10^{-13} – 10^{2}
  - ✓ Rayleigh Scattering and CIA
- No noise
- **Spectral range**: 0.838-1.666 μm in 13 bins

- **Ariel 2022** challenge database (Changeat and Yip, RASTI, 2023).

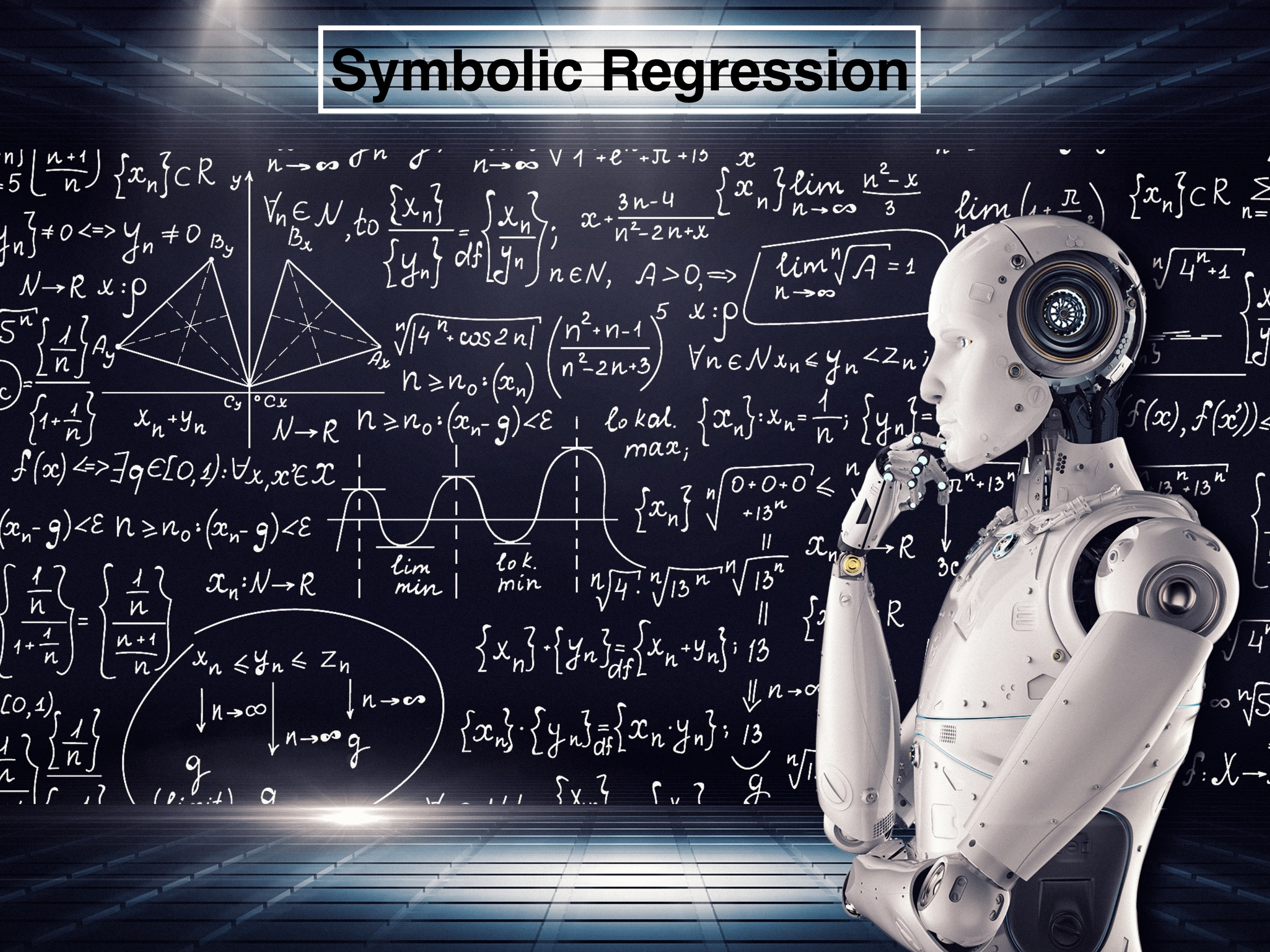
Ariel database (**TauRex**) with variable gravity , g

- **Fixed parameters**: pressure grid, mean molecular mass, μ
- **Varying parameters:** target planet/star: R_s, R_p, M_p, g, T_p.
- **Scanned parameters**:
  - ✓ H_2O volume mixing ratio: 10^{-9} – 10^{-3}
  - ✓ CO_2 volume mixing ratio: 10^{-9} – 10^{-4}
  - ✓ CH_4 volume mixing ratio: 10^{-9} – 10^{-3}
  - ✓ CO volume mixing ratio: 10^{-6} – 10^{-3}
  - ✓ NH_3 volume mixing ratio: 10^{-9} – 10^{-4}
  - ✓ No clouds
  - ✓ Rayleigh Scattering and CIA
- Noise
- **Spectral range**: 0.5-7.5 μm in 52 bins

You can make your own database!

- spectral range, resolution, and noise.
- what are the fixed/varying parameters(T, R, g, m, clouds,…)?
- what are the ranges?
- what type of sampling?
- what optical processes are included?
- what are the physics approximations?

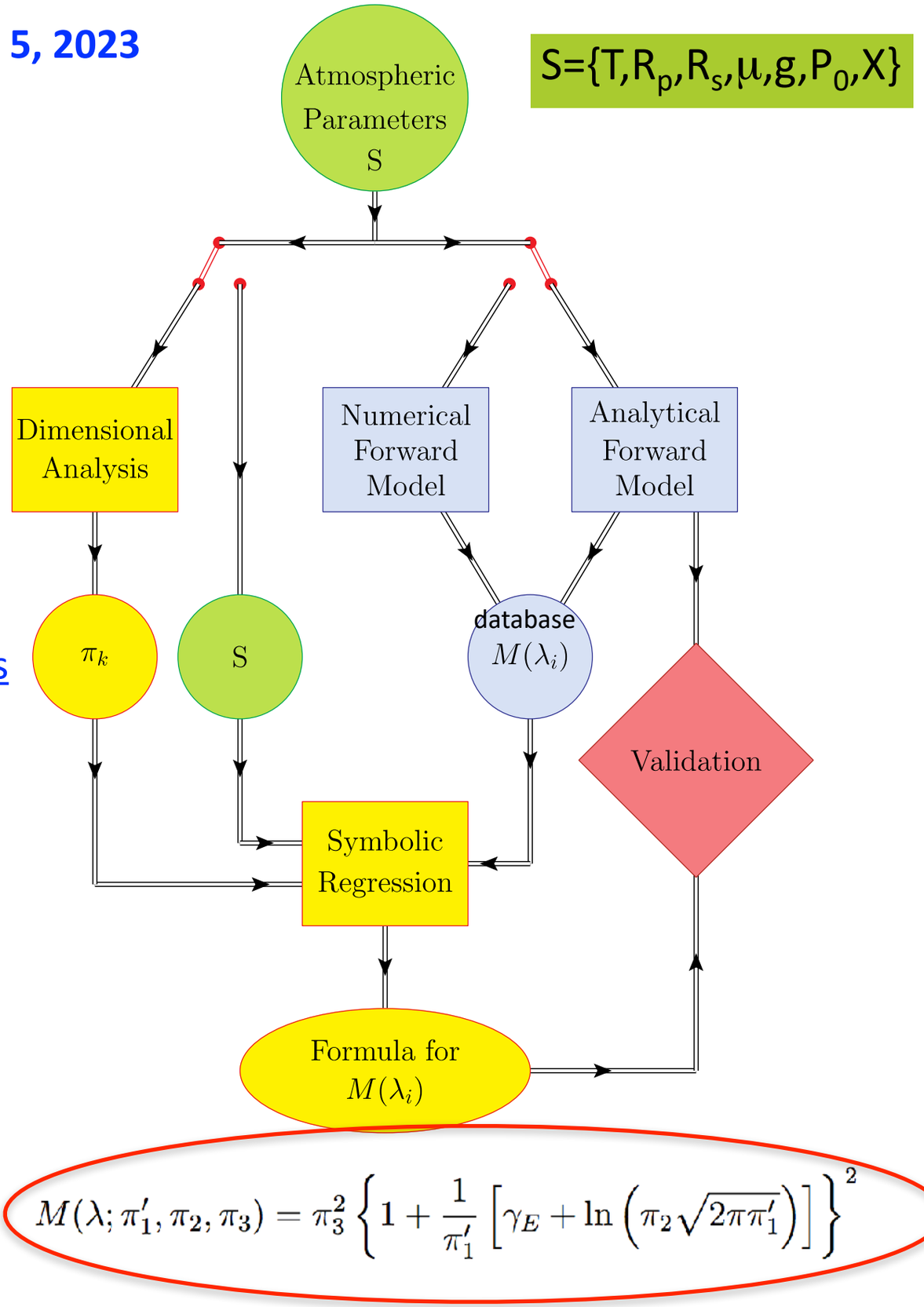
Symbolic Regression



Symbolic Regression: Learning From Data

Matchev, Matcheva, Roman, ApJ, v 930, n 1, 2022
Dong, Kong, Matchev, Matcheva, Phys Rev D, v 107, n 5, 2023

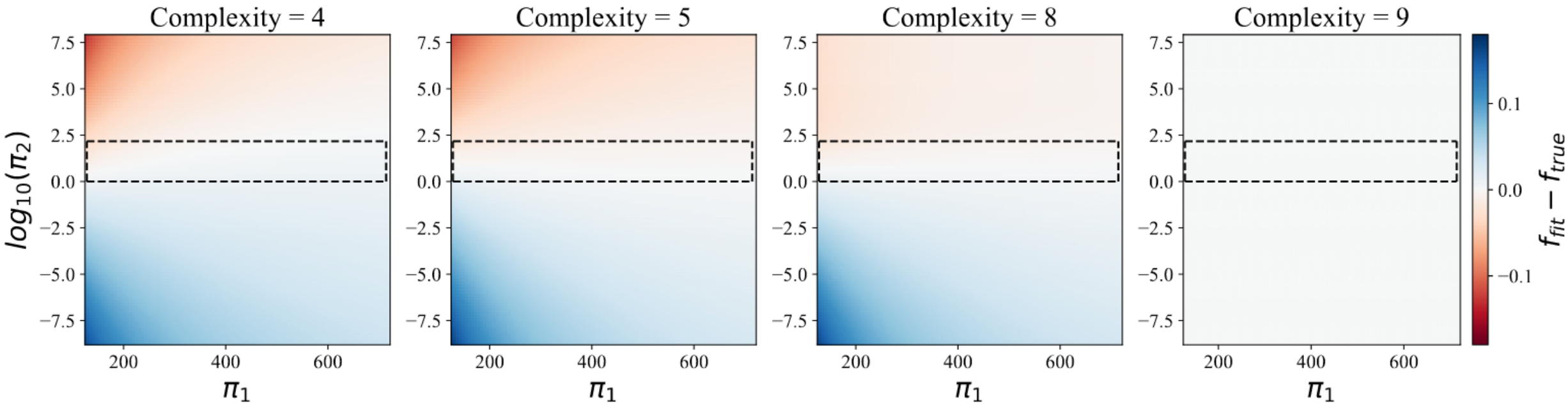
- Let's ask the AI to play the role of the theorist
  - Use the results from the simulations to derive an analytical formula
- Use symbolic regression as implemented in PySR
  - https://github.com/MilesCranmer/PySR
    https://www.youtube.com/watch?v=q6tjKXmhiMs



Following the **yellow branch** of the flowchart, use the Pi representation **(π_1,π_2, π_3, π_4)** and apply **Symbolic Regression** to derive the analytical form of **M(π_1,π_2, π_3, π_4)** from a synthetic database generated from the **Analytical Forward Model**.

Symbolic Regression

We demonstrate the implementation of **Symbolic Regression** to derive (recover) the analytical form of $M(\pi_1,\pi_2, \pi_3, \pi_4)$ from a synthetic database generated using equation (*). Below are the derived analytical fits to the data with different complexity and corresponding error, MSE. For the correct fit (**complexity 9**) the MSE is limited by the machine accuracy, **MSE~10^{-15}**.



$$f_{\text{fit}}^{(4)} = \frac{0.456}{\sqrt{|\pi_1|}}, \qquad MSE = 8.314751 \times 10^{-5},$$

$$f_{\text{fit}}^{(5)} = \frac{1.9413}{0.3009\,\pi_1} = \frac{6.452}{\pi_1}, \qquad MSE = 6.008056 \times 10^{-5},$$

$$f_{\text{fit}}^{(8)} = \frac{\ln(|\pi_2 + 2.88\,\pi_1|)}{\pi_1}, \qquad MSE = 5.702897 \times 10^{-5},$$

$$f_{\text{fit}}^{(9)} = \frac{\ln(4.4645\,|\pi_2|\,\sqrt{|\pi_1|})}{\pi_1}, \qquad MSE = 1.003324 \times 10^{-15}.$$

$$M(\lambda;\pi_1',\pi_2,\pi_3) = \pi_3^2\left\{1 + \frac{1}{\pi_1'}\left[\gamma_E + \ln\left(\pi_2\sqrt{2\pi\pi_1'}\right)\right]\right\}^2 \quad *$$

Degeneracies

A **degeneracy** arises when a suitable scaling of the input planetary parameters leads to an identical observed spectrum. The dimensional analysis reveals three families of degeneracies.

Variable	R_0 scaling	P_0 scaling	κ scaling
R_0	L_{R_0}	—	—
P_0	—	L_{P_0}	—
κ	—	—	L_κ
T	$L_{R_0}^3$	L_{P_0}	—
m	$L_{R_0}^2$	—	L_κ^{-1}
g	—	L_{P_0}	L_κ
R_S	L_{R_0}	—	—

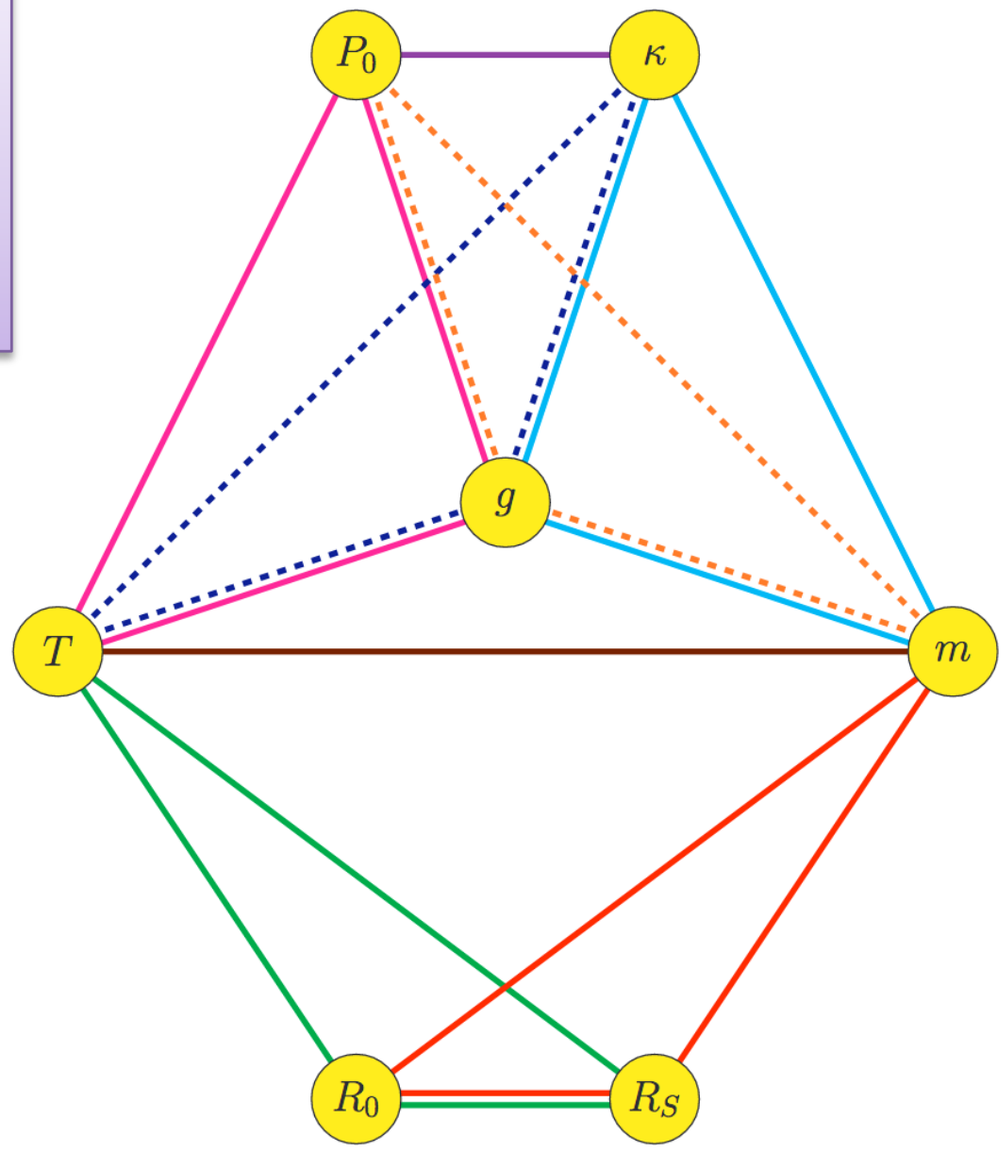
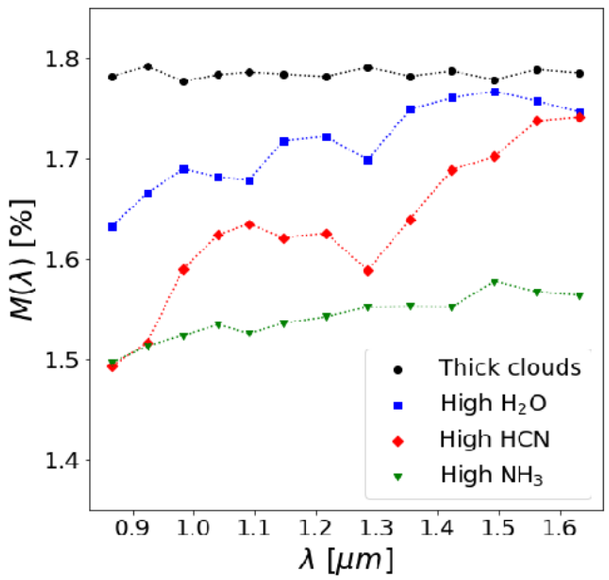


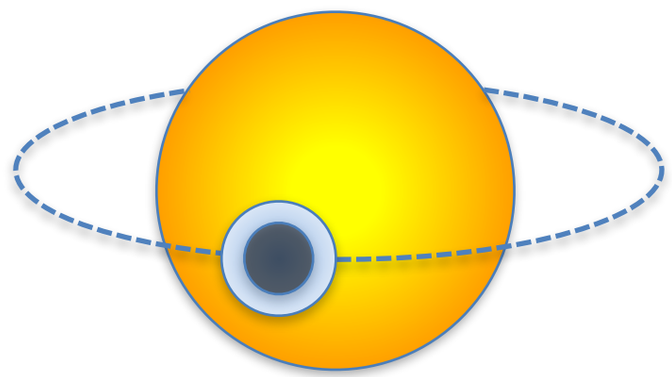
Table: Explicit parameterization of the three guaranteed degeneracies. The three scaling factors ***L*** are for R_0, P_0, κ, respectively.

Diagram: Graph representation of the two-level (**line segments**) and three-level (**triangles**) degeneracies.

Inverse Problem: parameter retrievals



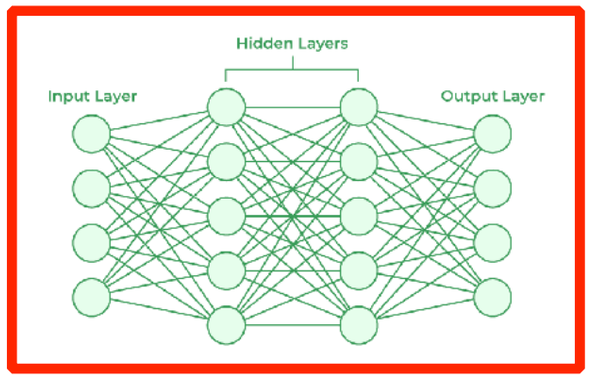
Retrieval Model



Inputs (features)

Planet	*M_1 %*	*M_2*	*M_3*	*M_4*
1	*1.41*	*1.44*	*1.42*	*1.52*
2	*0.52*	*0.55*	*0.61*	*0.58*
3	*0.92*	*1.03*	*1.11*	*0.95*
4	*1.85*	*1.94*	*1.99*	*1.82*

Retrieval: TauREx

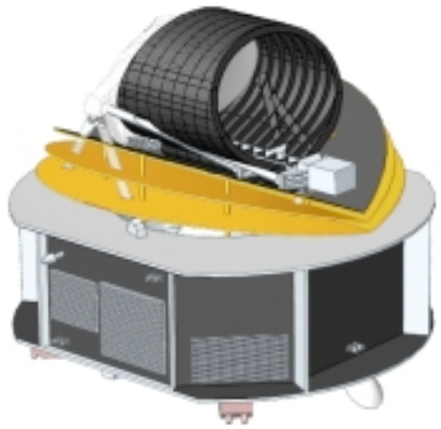


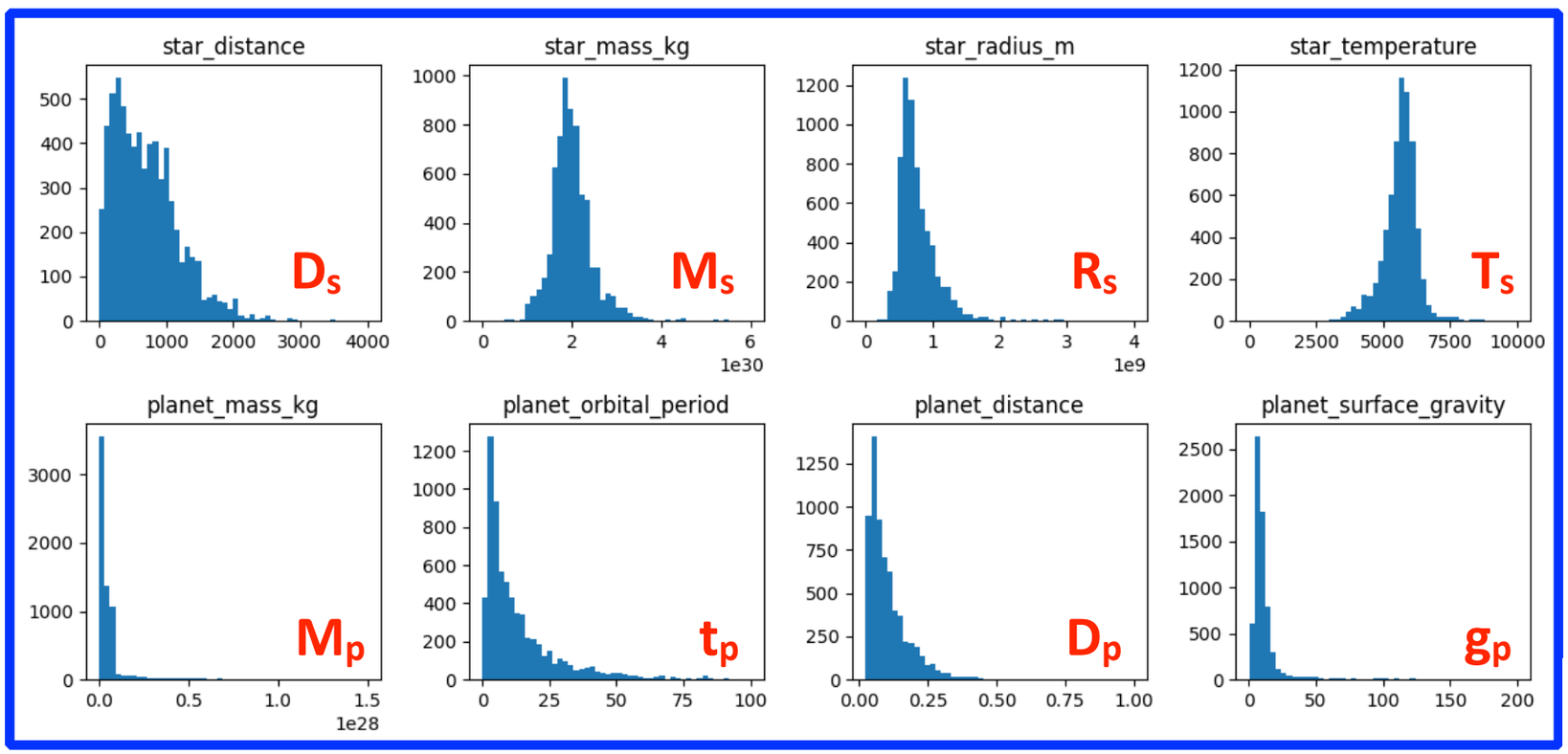
Outputs (targets)

Planet	*T (K)*	*R_p (R_J)*	*R_s (R_{sun})*	*X_{H20}*
1	*1300*	*1.8*	*1.6*	*10^{-3}*
2	*650*	*0.9*	*1.4*	*10^{-2}*
3	*960*	*1.9*	*2.3*	*10^{-4}*
4	*1150*	*2.0*	*1.5*	*10^{-5}*

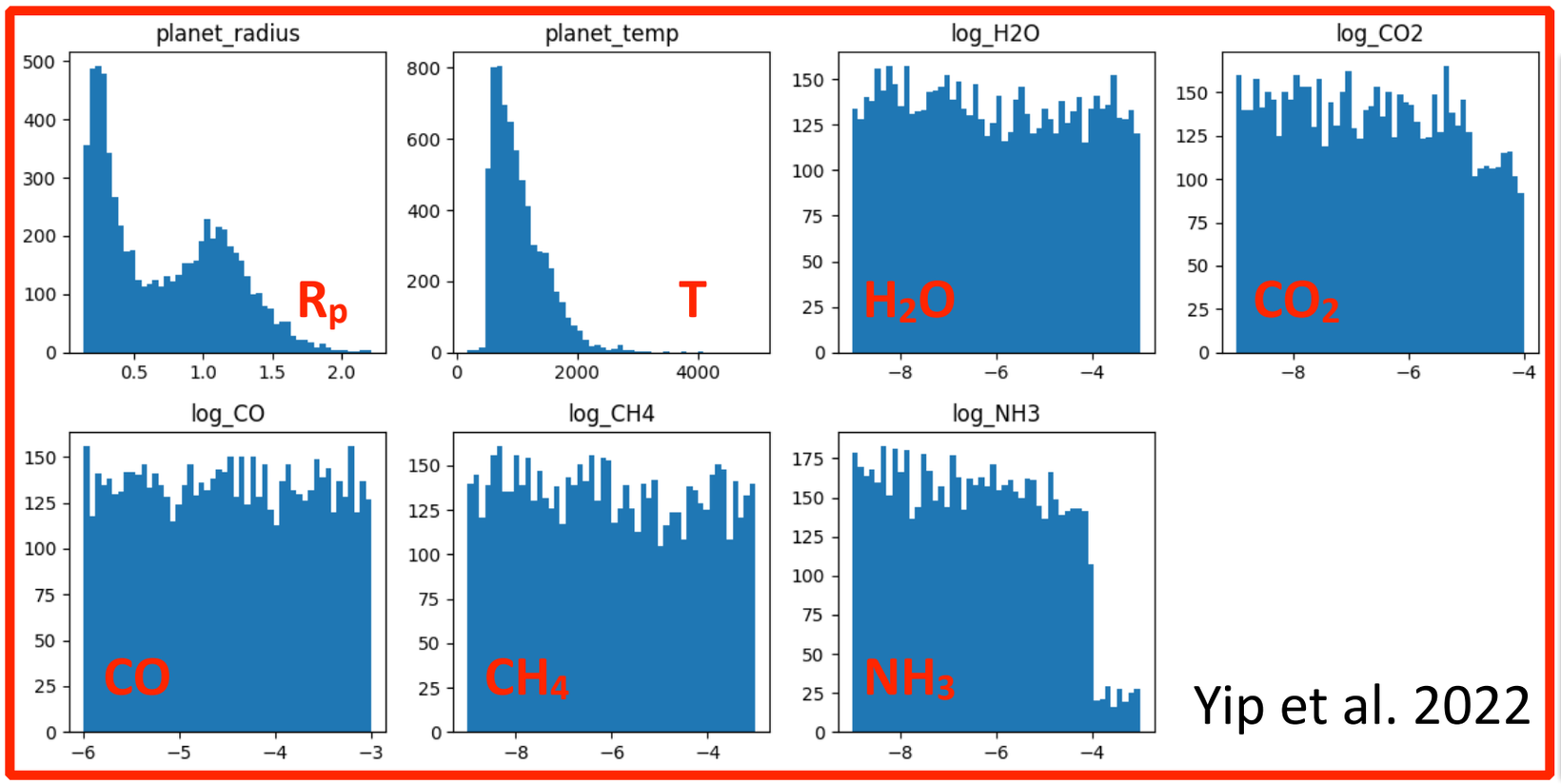
- The Ariel Data challenge is **ML as a substitute for the Bayesian model**: Train on a database of solutions from TauREx with the goal of reproducing the TauREx predictions.

Ariel 2022 Database





100,000 synthetic spectra



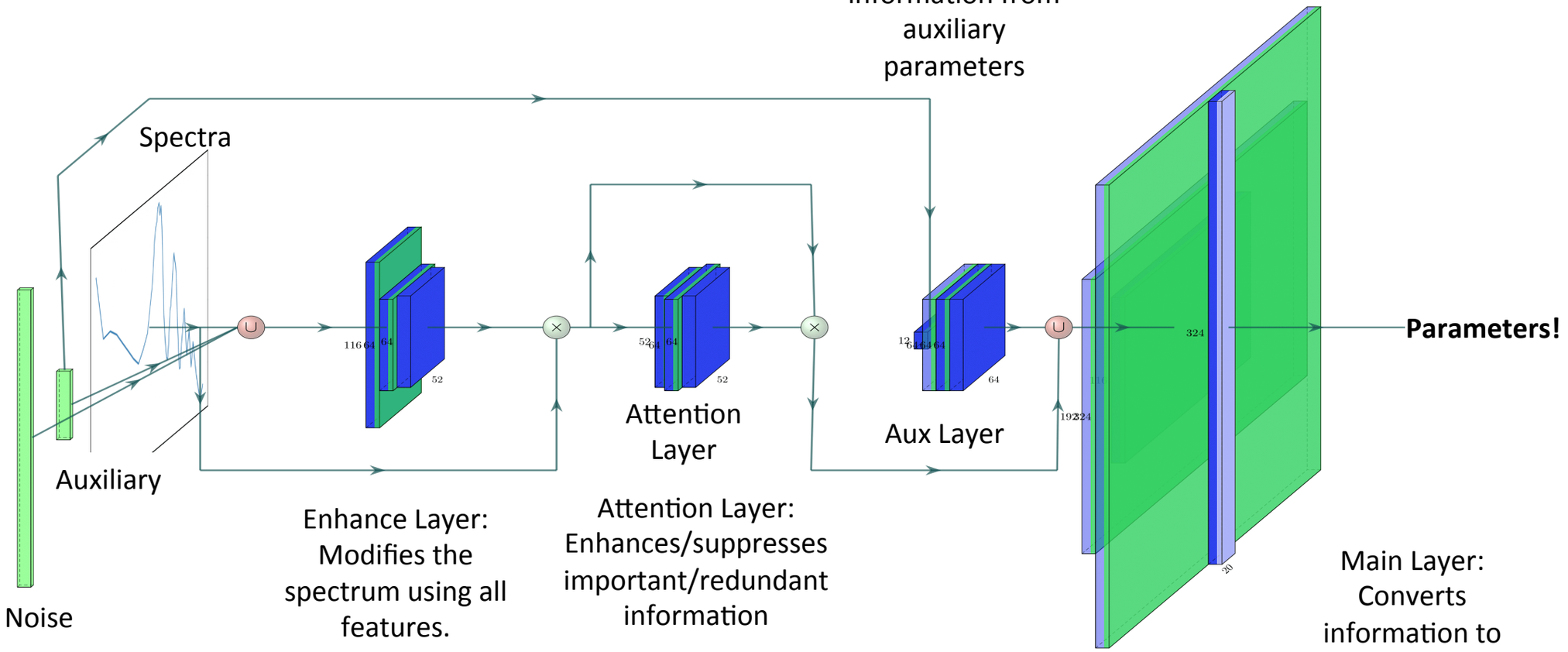
Yip et al. 2022

Ariel database (**TauRex**) with variable gravity , g

- **Fixed parameters:** pressure grid, mean molecular mass, μ
- **Varying parameters:** target planet/star: R_s, R_p, M_p, g, T_p.
- **Scanned parameters**:
  - ✓ H_2O volume mixing ratio: 10^{-9} – 10^{-3}
  - ✓ CO_2 volume mixing ratio: 10^{-9} – 10^{-4}
  - ✓ CH_4 volume mixing ratio: 10^{-9} – 10^{-3}
  - ✓ CO volume mixing ratio: 10^{-6} – 10^{-3}
  - ✓ NH_3 volume mixing ratio: 10^{-9} – 10^{-4}
  - ✓ No clouds
  - ✓ Rayleigh Scattering and CIA
- Noise
- **Spectral range**: 0.5-7.5 μm in 52 bins

Model Architecture and Training

- We only trained on 21,988 labeled samples. We didn't use the provided unlabeled data.
- We used 80/20 train-test split to determine the training hyperparameters.
- The model was trained with the Adam optimizer.
- We use different learning rates at different stages of the training.
- We used a dropout layer during training.



Parametrization of the posterior distribution

- The labeled six-dimensional population was parametrized with the following ansatz.

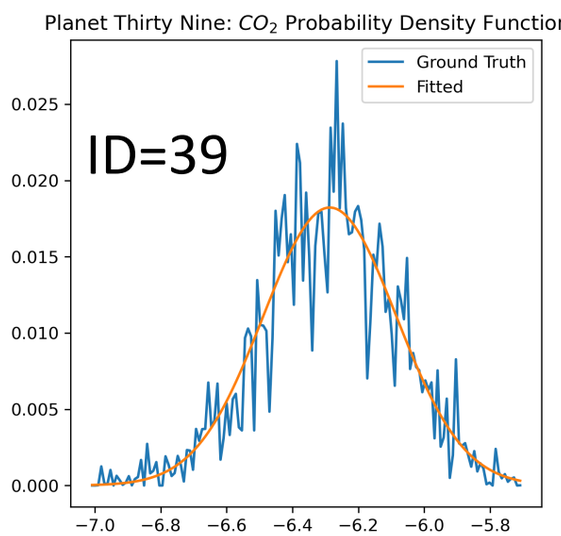
$$\rho(T,\vec{x};T_p,\mu_i,\sigma_i,A_i,m_i) = \left[\frac{\Theta(T_p-T)}{\sigma_{T_1}\sqrt{2\pi}}e^{-\frac{(T-T_p)^2}{2\sigma_{T_1}^2}} + \frac{\Theta(T-T_p)}{\sigma_{T_2}\sqrt{2\pi}}e^{-\frac{(T-T_p)^2}{2\sigma_{T_2}^2}}\right]$$

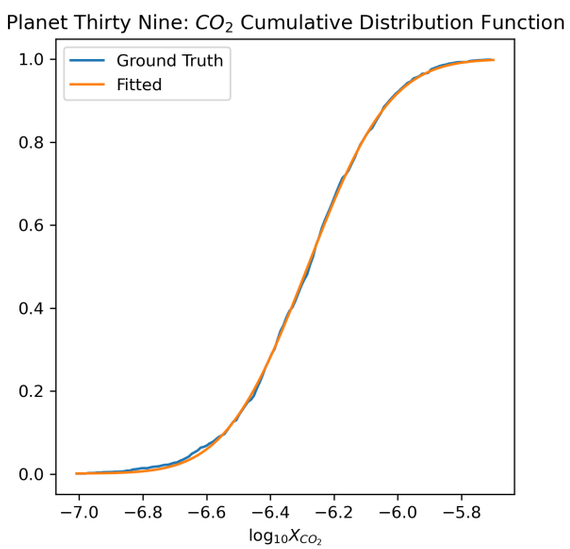
Bigaussian

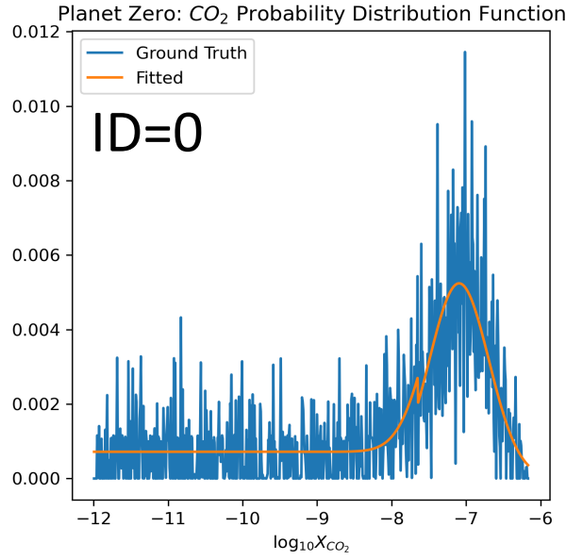
Gaussian

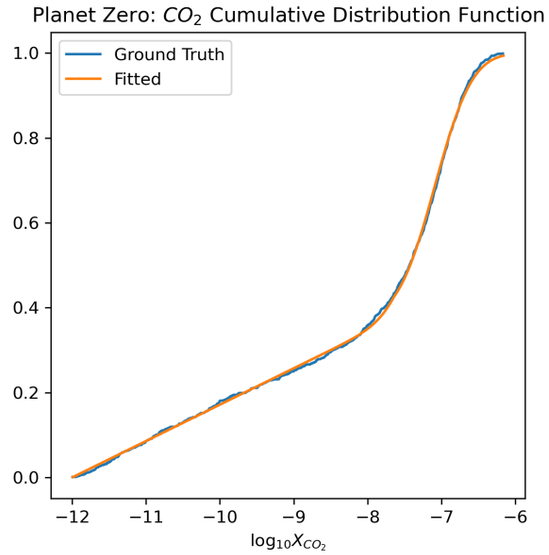
$$\times \prod_{i=1}^{5}\left[\frac{A_i}{\sigma_i\sqrt{2\pi}}e^{-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}} + (1-A_i)\frac{\Theta(x_i+12)\Theta(m_i-x_i)}{m_i+12}\right]$$

Uniform









Preprocessing and Feature Engineering

- **Cleaning** the data by modifying unphysical values for the flux
- **Subtracting** the opaque disk of the planet

$$M'_{i\lambda} = M_{i\lambda} - \left(\frac{R_p}{R_s}\right)^2$$

- **Rescaling** of the spectral features

$$M''_{i\lambda} = \frac{M'_{i\lambda}}{\max_\lambda M'_{i\lambda}}$$

- **Rescaling** of the noise

$$\epsilon'_{i\lambda} = \frac{\epsilon_{i\lambda}}{\max{}_\lambda M'_{i\lambda}}$$

- **Feature engineering** - add new dimensionless variables (Pi-groups)

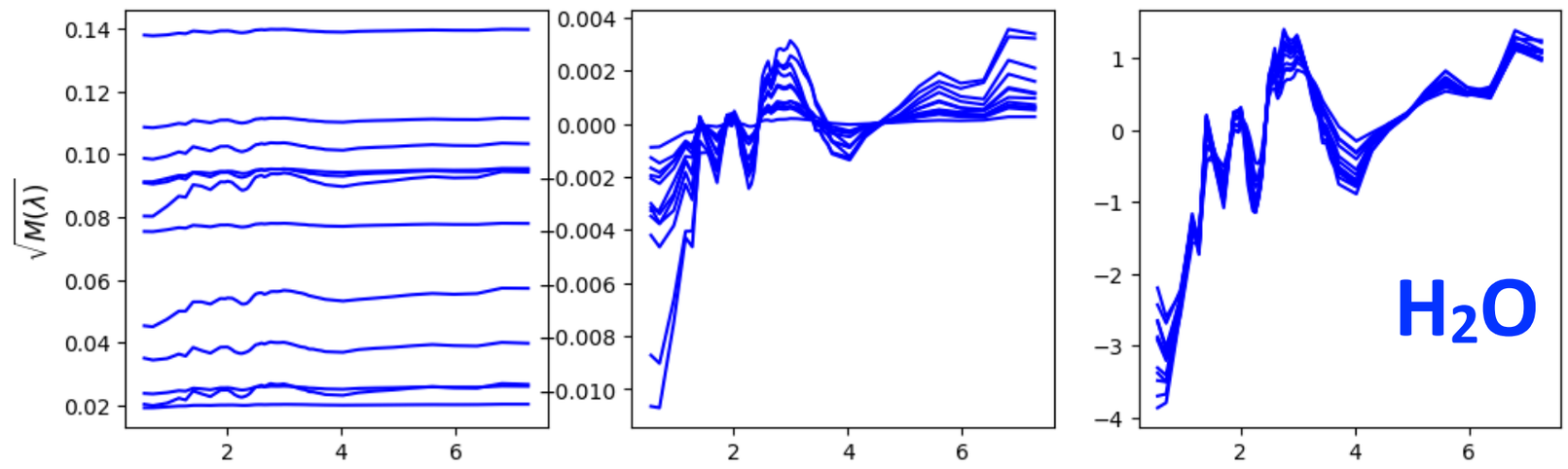
$$Aux \rightarrow (Aux, \frac{R_p}{R_s}, \frac{D}{H}, \frac{R_s}{H}\max_\lambda M'_\lambda)), \quad H = \frac{k_b T_p}{mg}, \quad T_p = T_s\sqrt{\frac{R_s}{2D}}$$

- **Standardizing** the auxiliary features

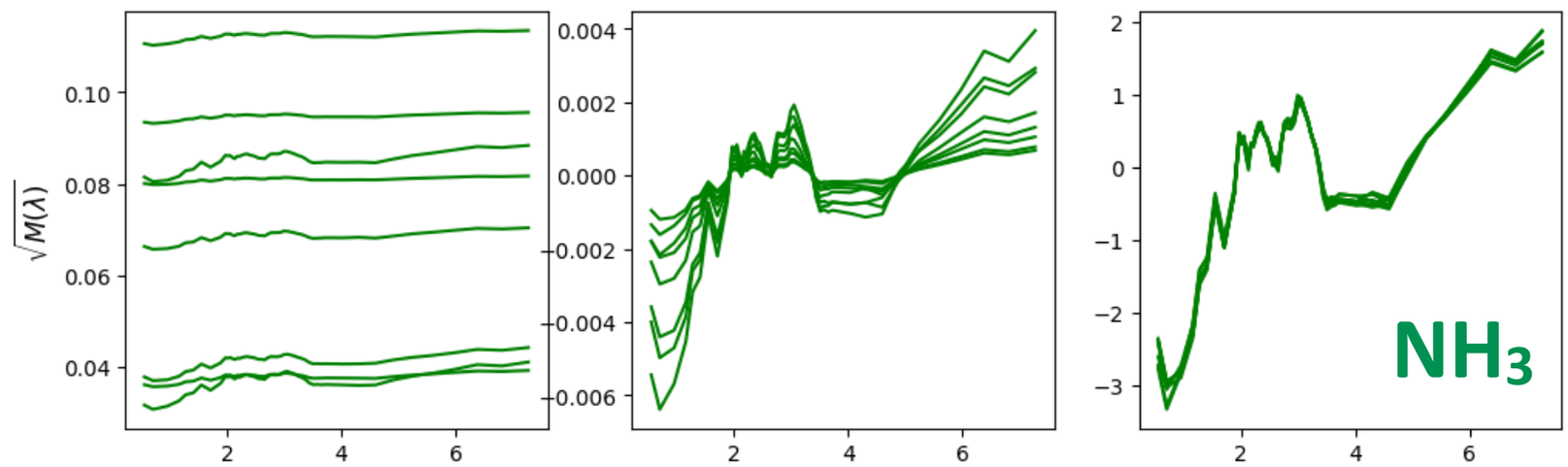
Preprocessing: Rescaling the spectra

- Original
- Centered
- Normalized

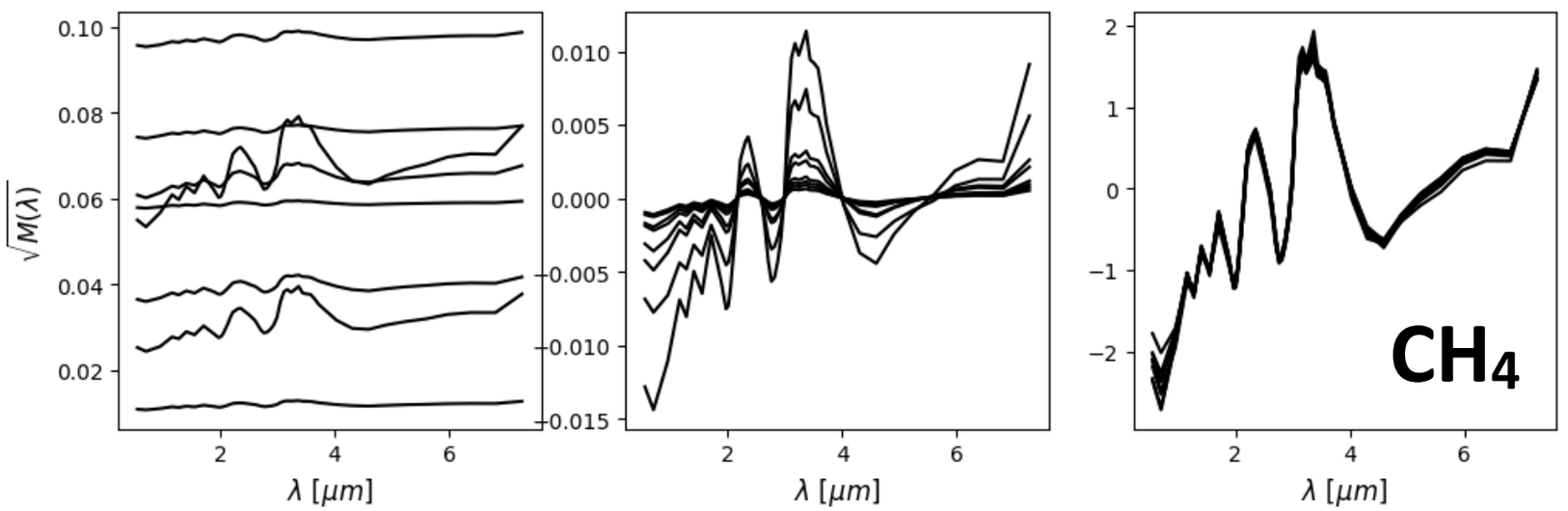
H_2O



NH_3



CH_4



Different

R_s

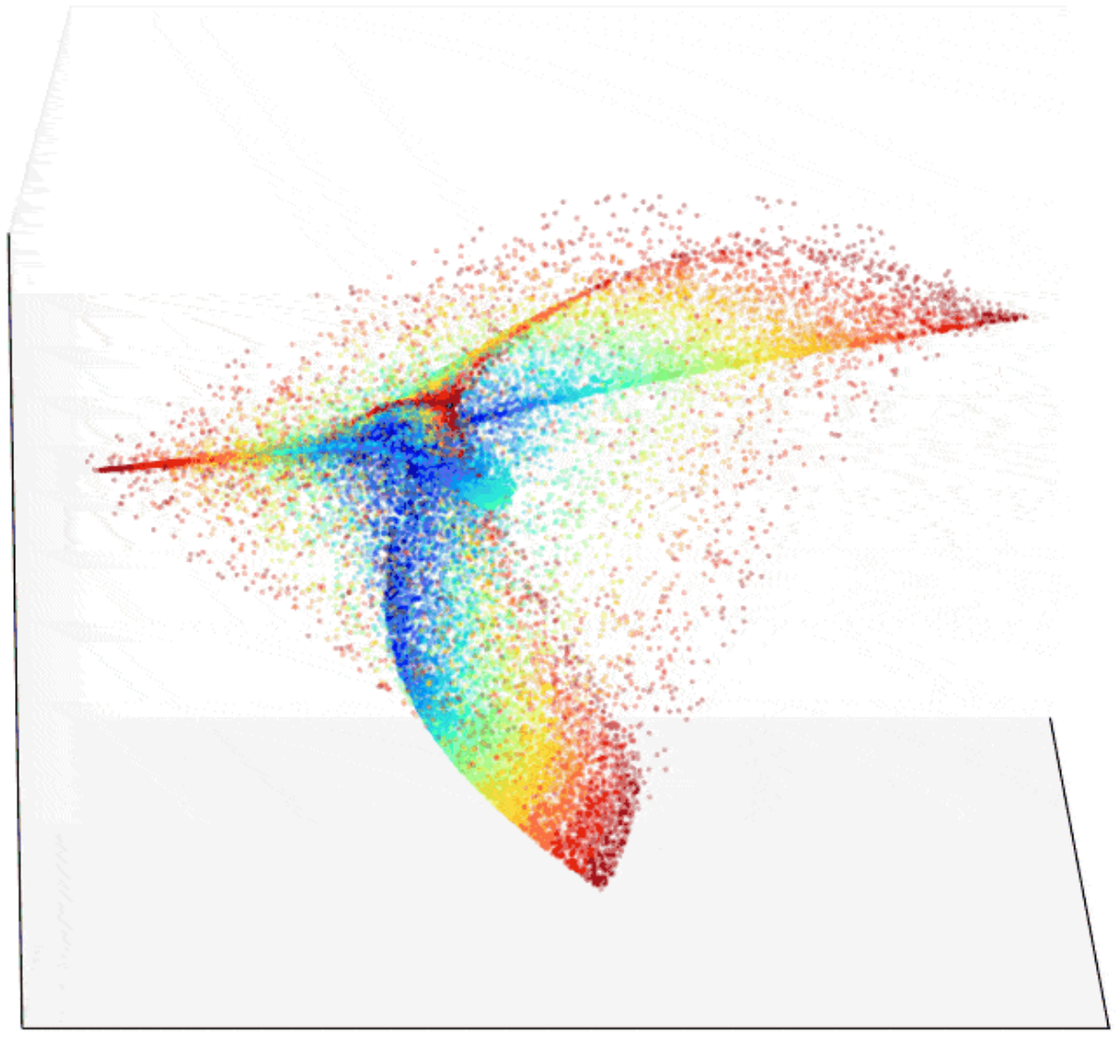
R_p

T_p

- SQRT
- Subtract the mean
- Divide by the Std Dev

3-D representation of the first 3 Principal Components

- **TRANSIT** database (with M. Himes, J. Harrington, unpublished)



Color coding by temperature, T=500-2900K

Spectral classes of chemical regimes

- ✓ H_2O branch
- ✓ NH_3 branch
- ✓ HCN branch
- ✓ Cloud branch

Distinct branch for each extinction

- ✓ Absorption due to distinct absorber
- ✓ Scattering
  - ✓ Grey clouds
  - ✓ Rayleigh scattering
- ✓ CIA (H_2-H_2, H_2-He)

Novelty Detection for Exoplanets

- **Science Questions**:
  - Can we identify planets with unusual or **unexpected chemical** composition?
  - Can we identify **new physics**?
  - Can we spot **glitches** with the instrument?
  - Can we spot **alien life** as we do not know it?

- **ML question**: Can we detect **anomalous** spectra?

(Forestano et al. ApJ 2023)

Outlier versus Novelty Detection

- **Outlier detection:** useful when we have an idea what anomalies might look like.

Training data

Testing data

- **Novelty detection:** useful when we do not know what the potential anomalies look like.

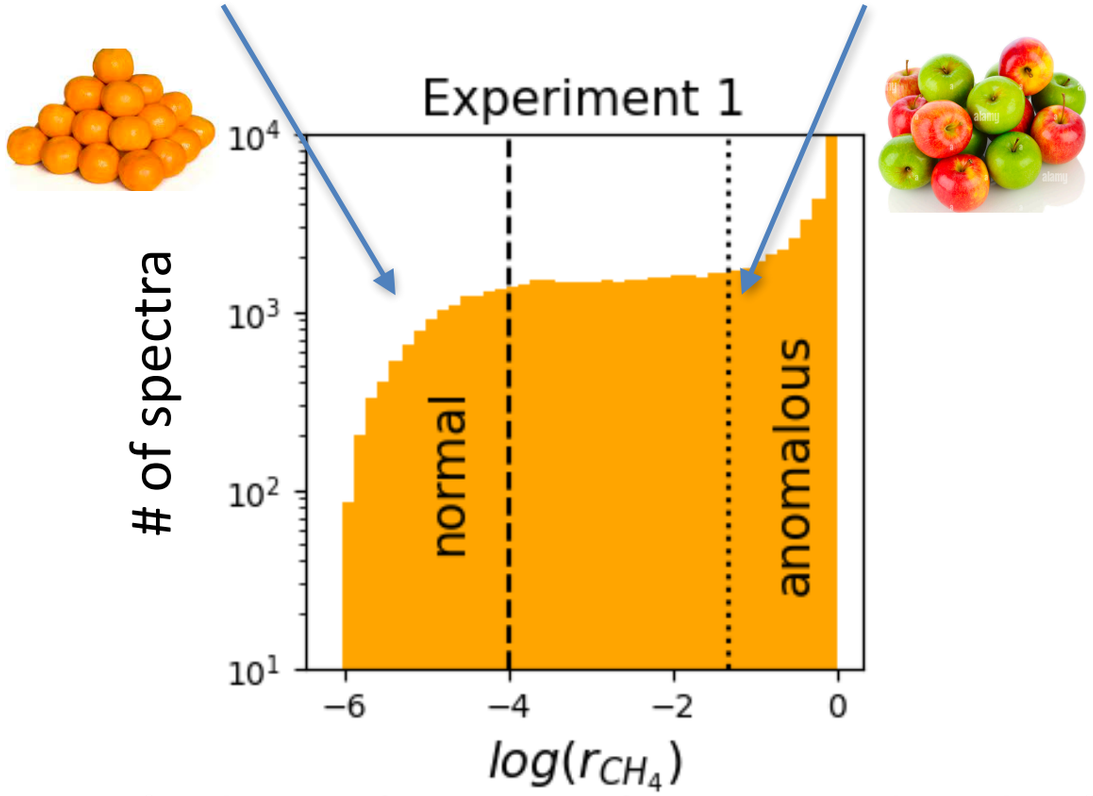
Training data

Testing data

Defining "anomalous" atmospheres

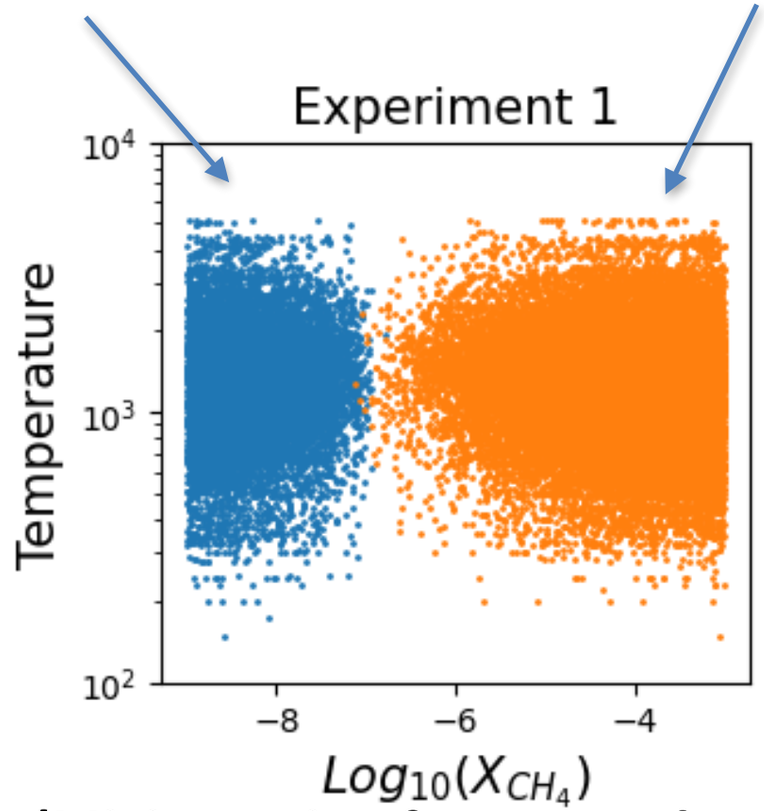
- Since we do not know what types of surprises we can get, we want to **train** the model on **normal** samples only : **"novelty detection"**
- The **testing** is done on both **normal** and **anomalous** samples
- **Anomalous**: having an unexpected **mystery absorber**
  - Experiment 1: **CH_4**
- **Normal**: a mixture of the remaining four absorbers in the database, no mystery absorber
  - Experiment 1: **CO_2, H_2O, NH_3, CO**

Normal **Anomalous**



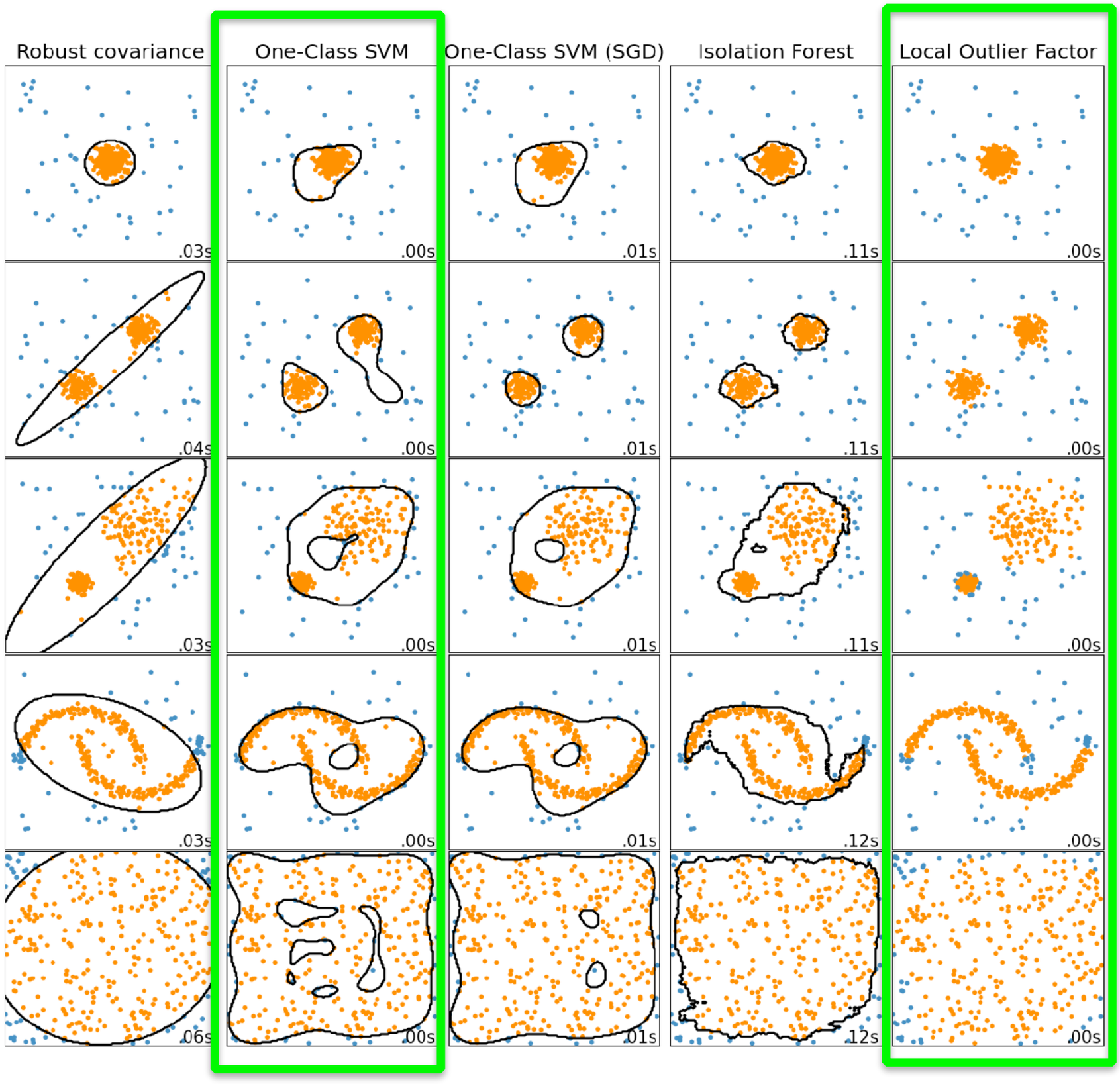
Log(Relative fraction of mystery absorber)

Normal **Anomalous**



Log(Mixing ratio of mystery absorber)

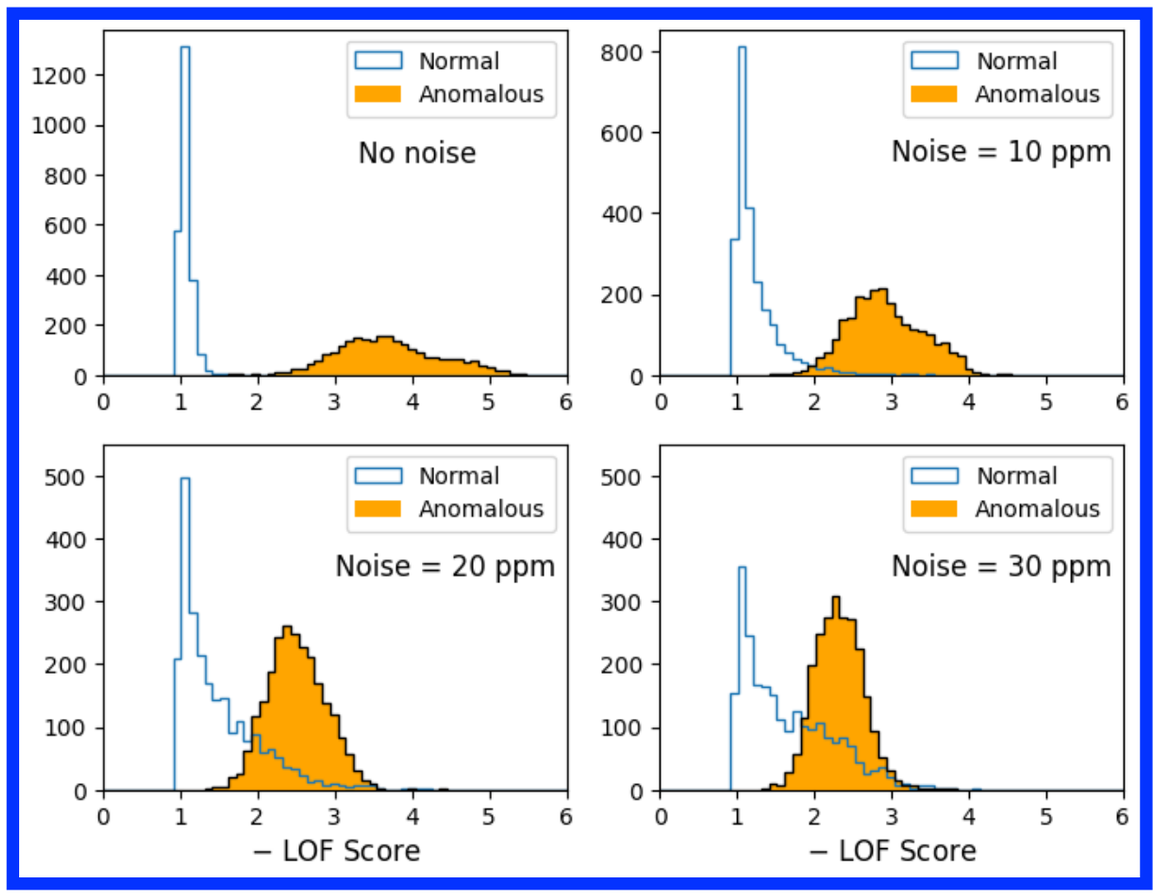
Anomaly Detection Methods



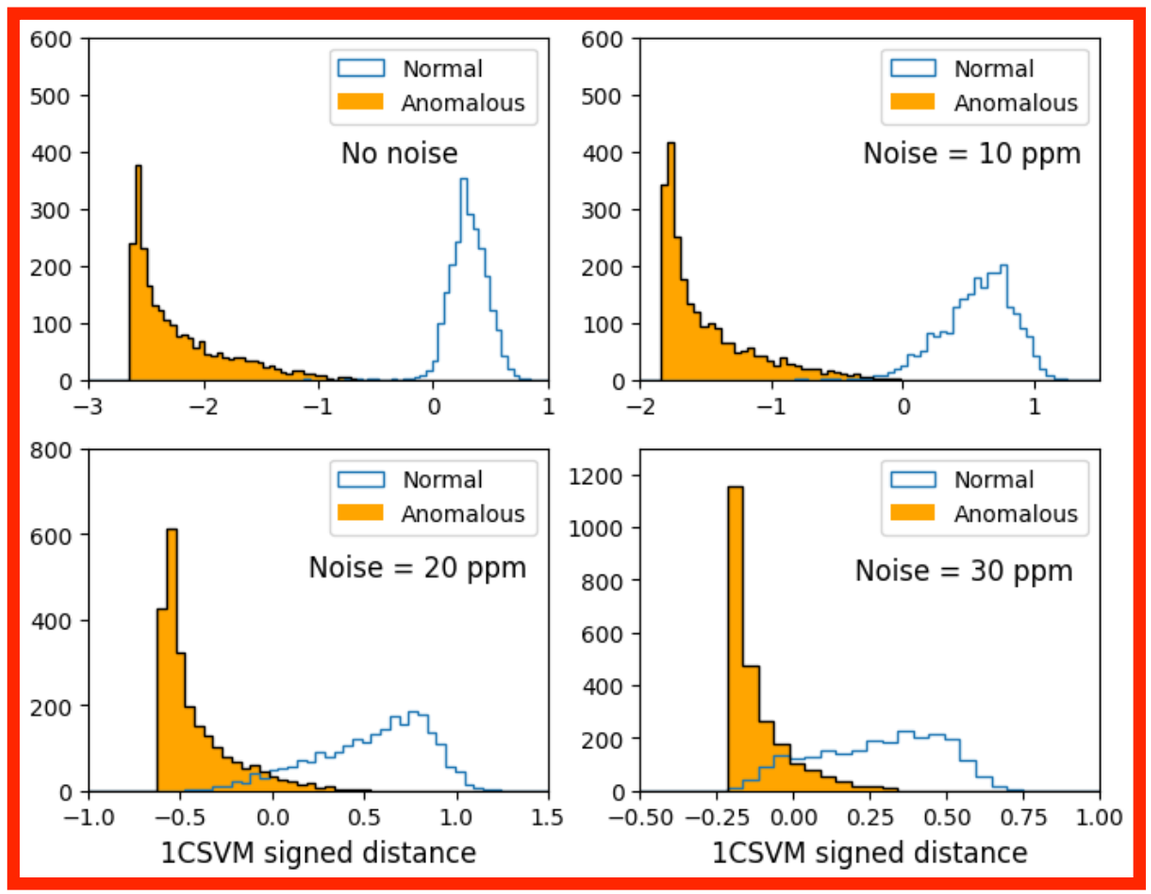
Results: LOF v/s 1CSVM

Blind testing on a unseen before spectra that are mixture of **normal** and **anomalous** samples

- **LOF**



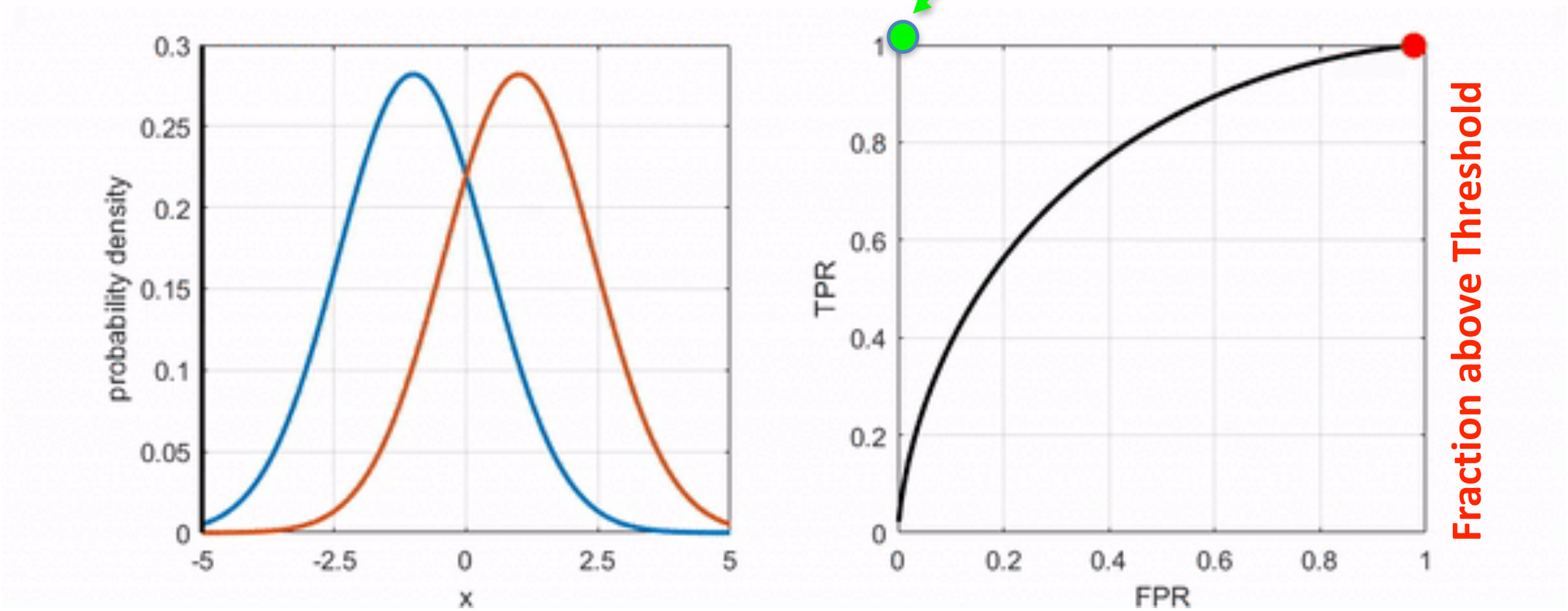
- **1 Class SVM**



- 1 class SVM seems to be more robust against noise.
- LOF has only one tunable hyper-parameter.

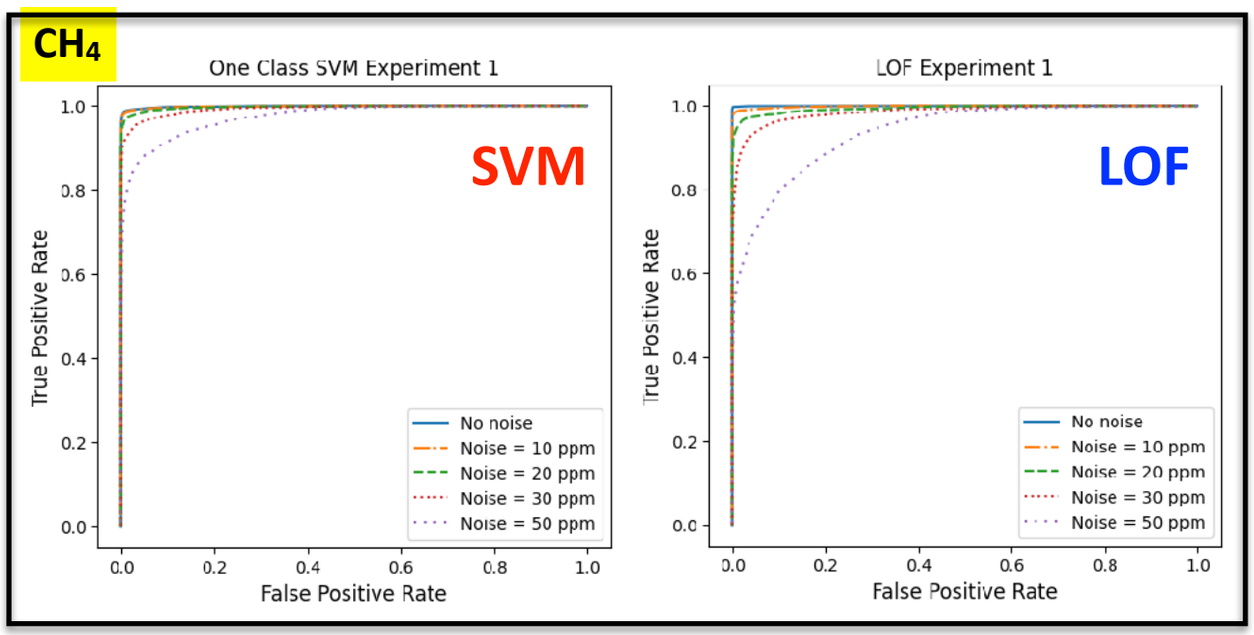
ROC Curve

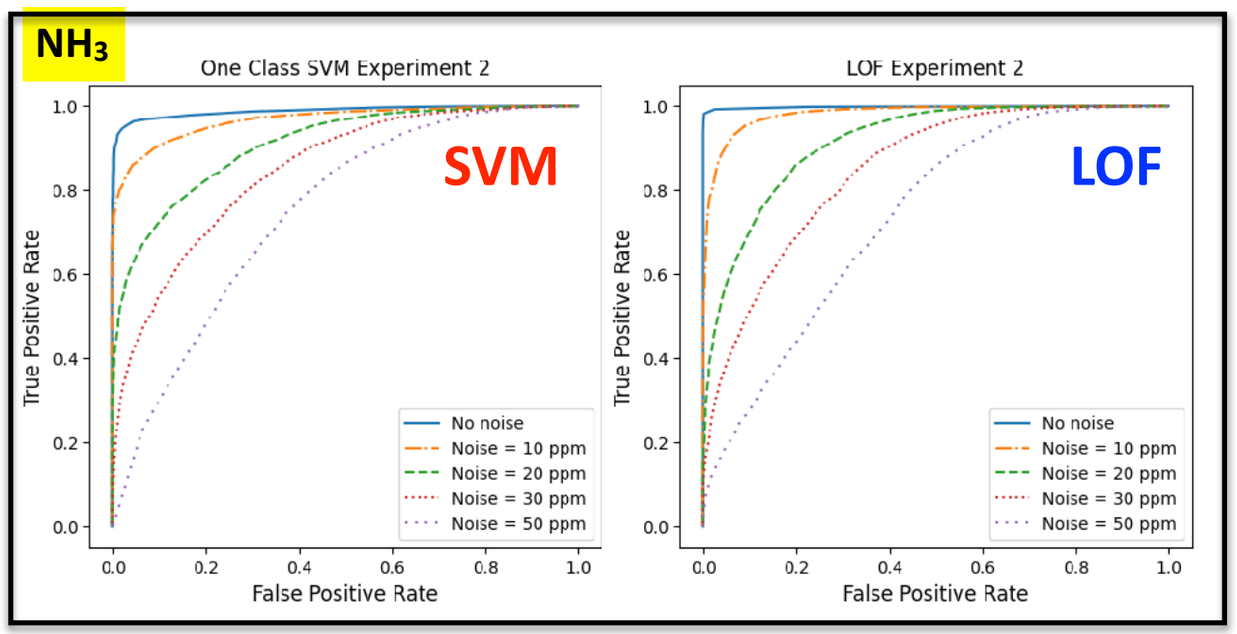
- A graph showing the **performance of a classifier** at all thresholds.
- Count the number of samples of each type to the right of the threshold

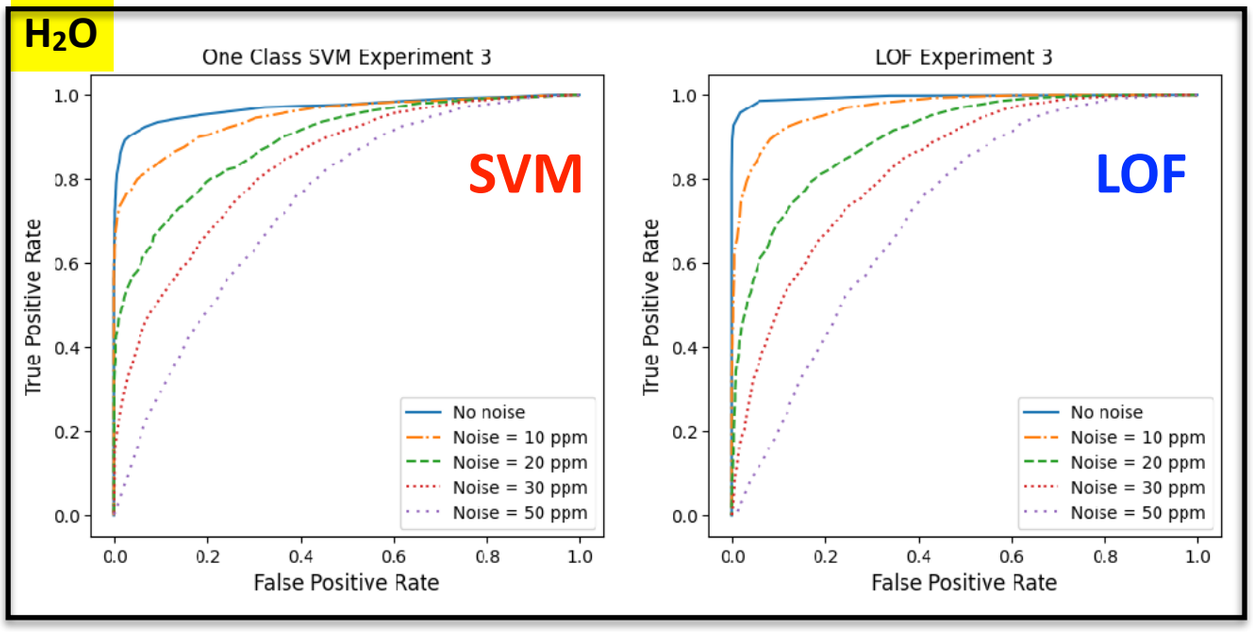


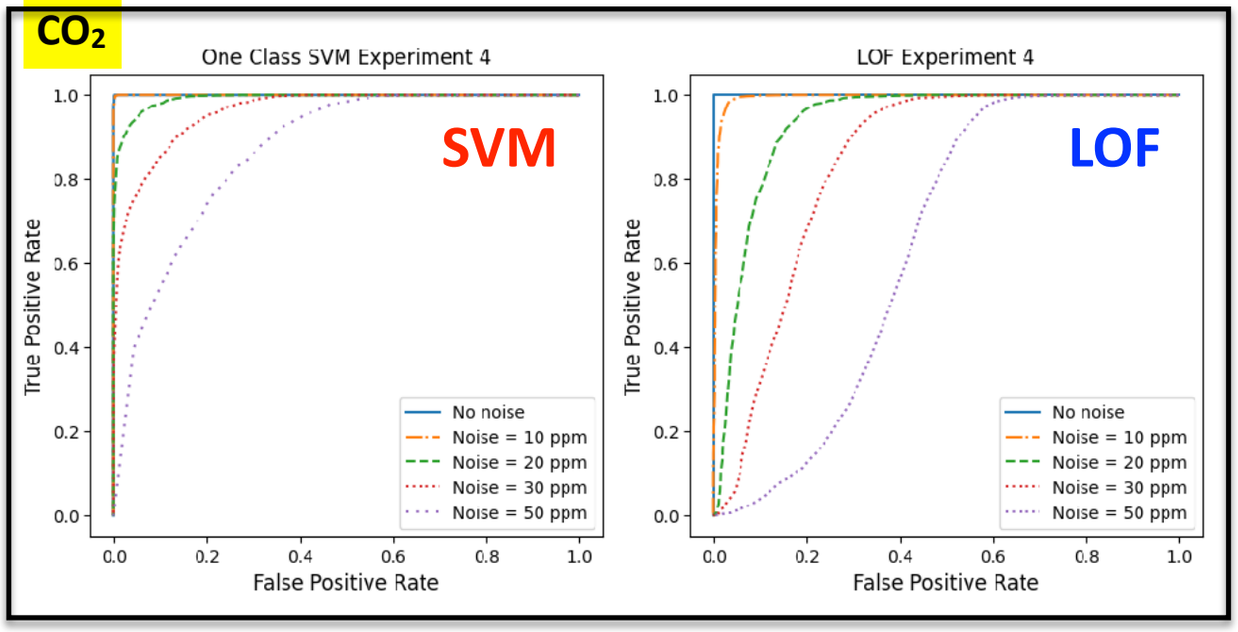
SVM vs LOF ROC curves

True Positive/False Positive rates



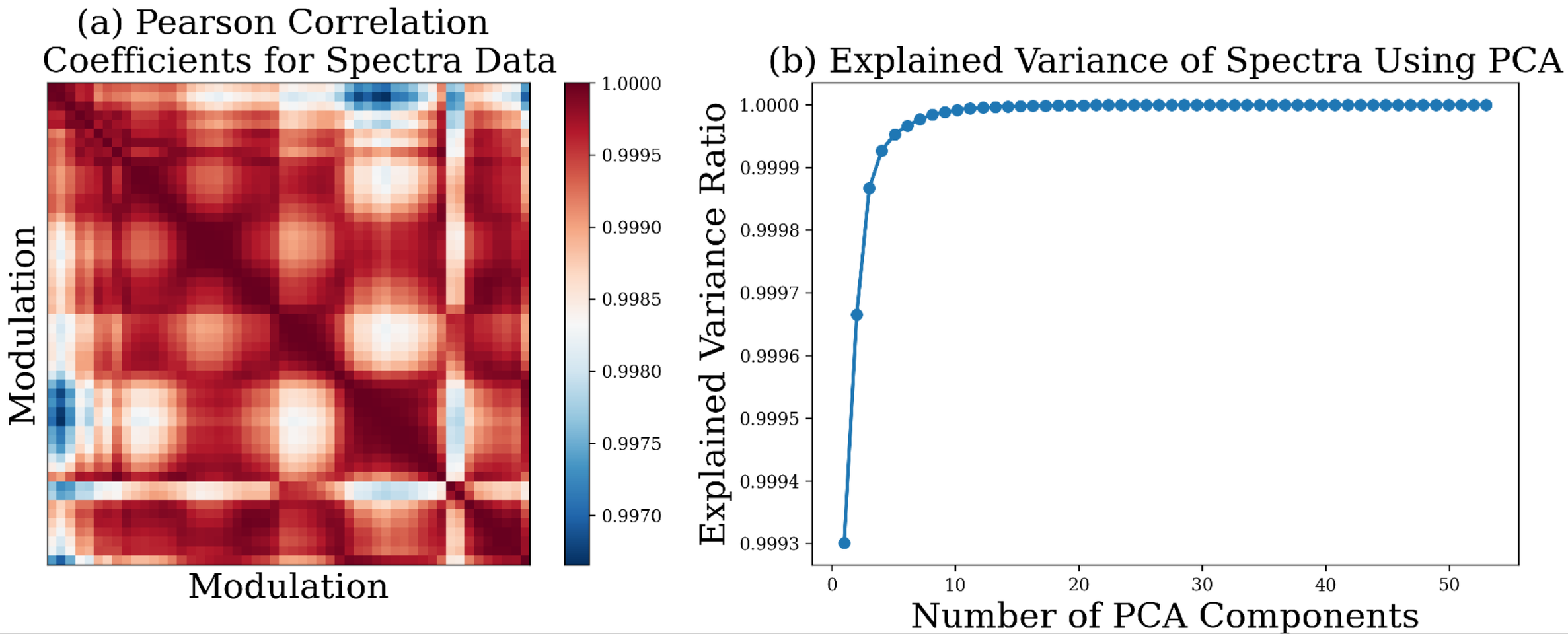






- 1 class SVM seems to be more robust against noise.
- LOF has only one tunable hyper-parameter and easy to interpret.

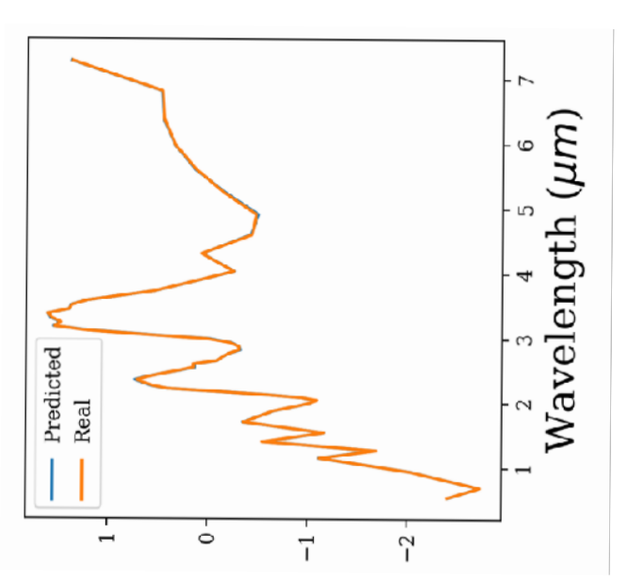
Correlations

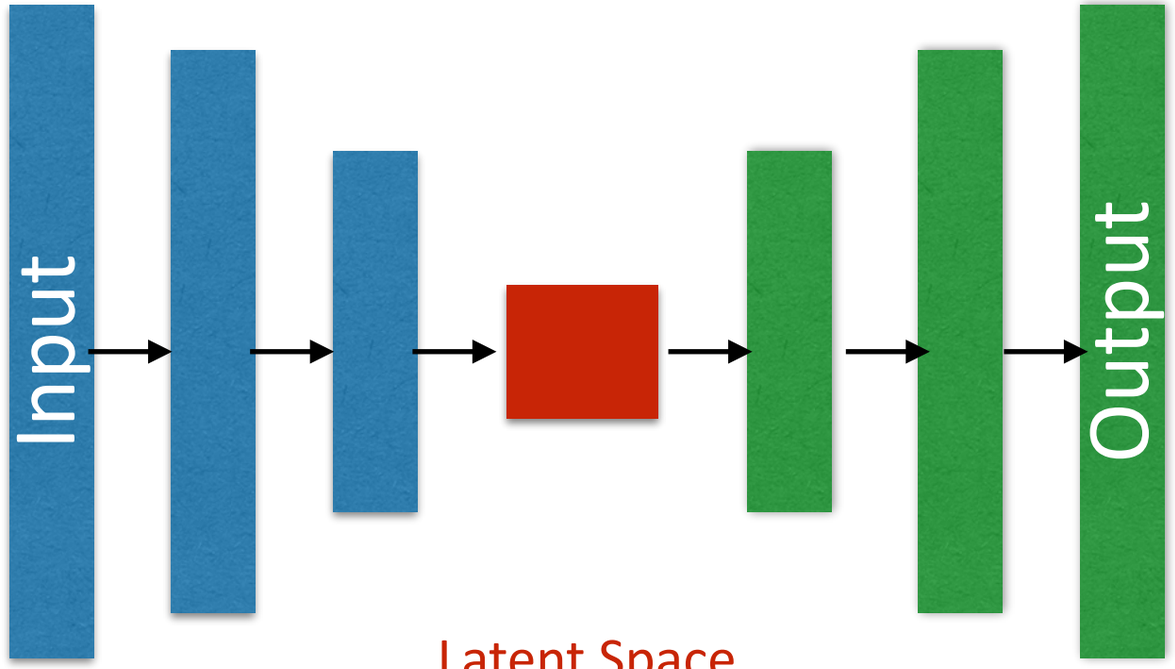


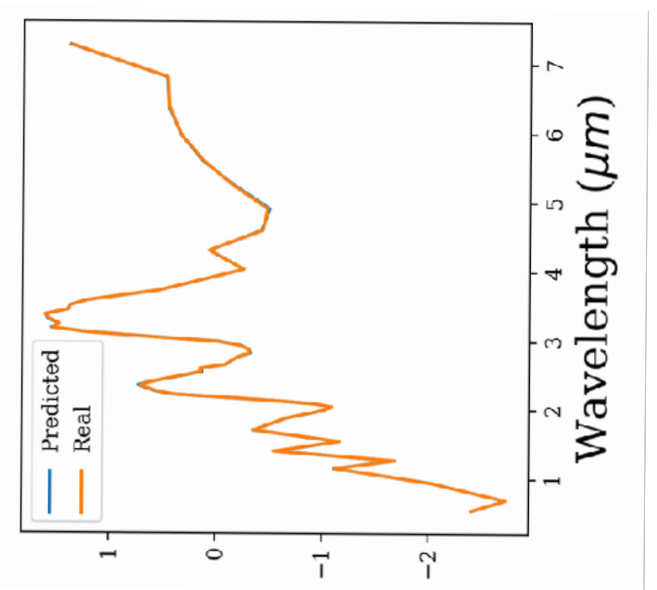
- The 52-D spectral database shows a high degree of **correlations**.
- PCA analysis illustrates that the data can be represented with **~10 variables**.
- This motivates the use of **dimensionality reduction** techniques.

Autoencoder

Encoder (52-32-16) **Bottleneck(8)** **Decoder (16-32-52)**







Spectral Space

52-dimensional

X_i

Latent Space

8-dimensional

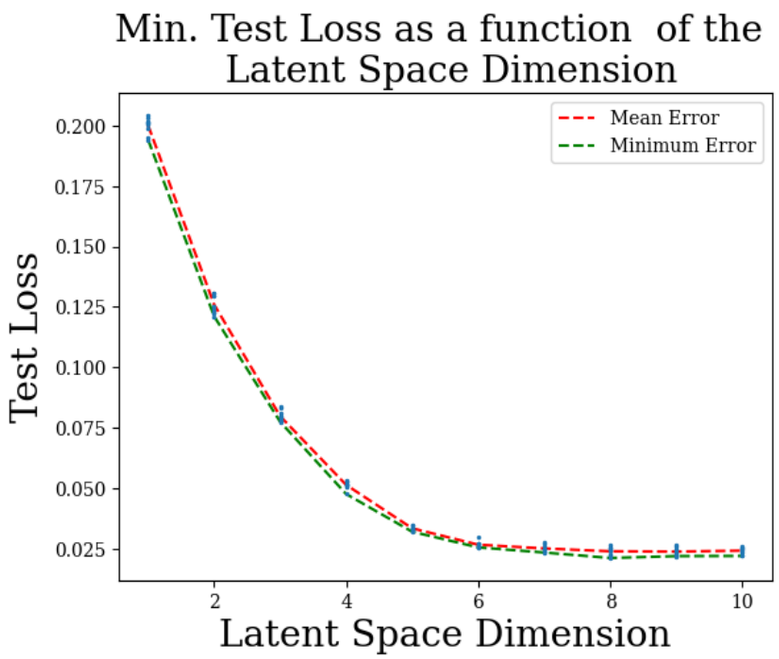
Compressed data

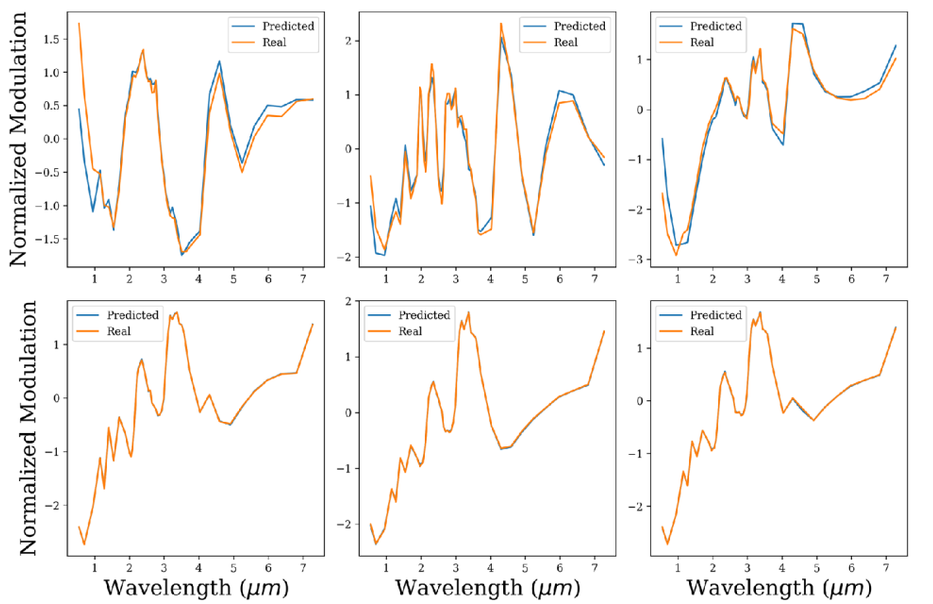
Reconstructed Space

52-dimensional

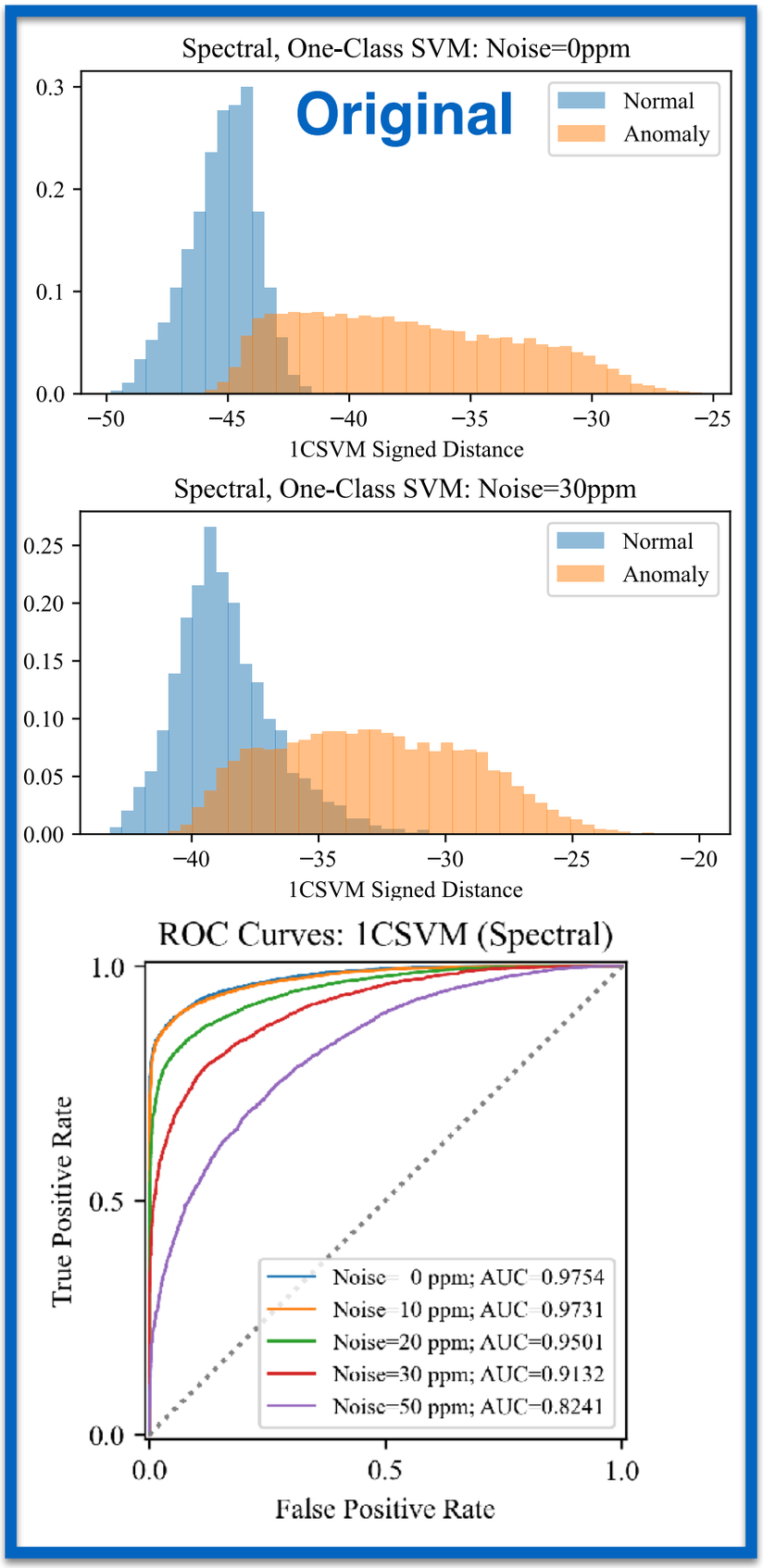
Y_i

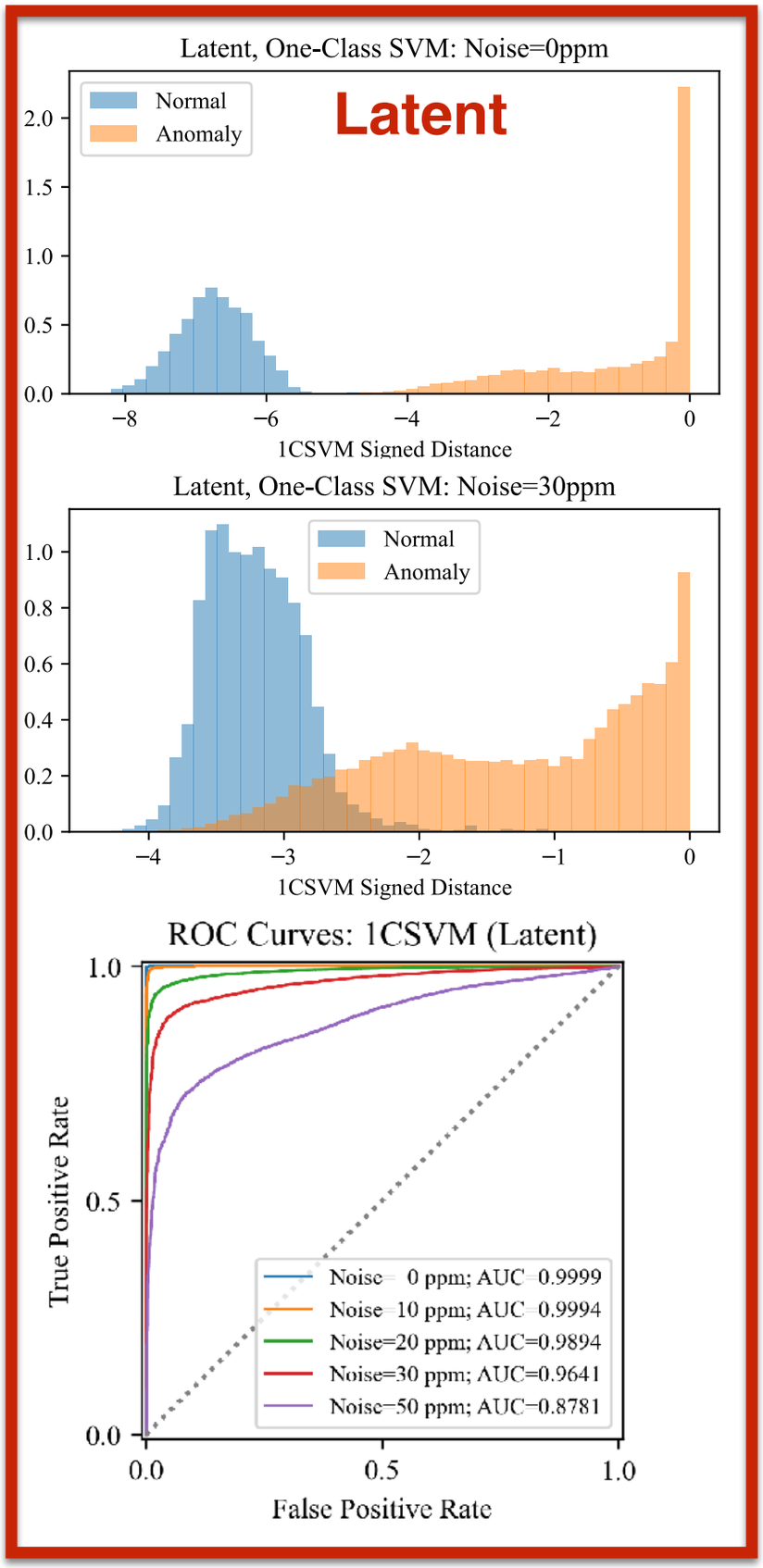
For **anomalous** spectra the **reconstruction loss** is large (the fit is bad).

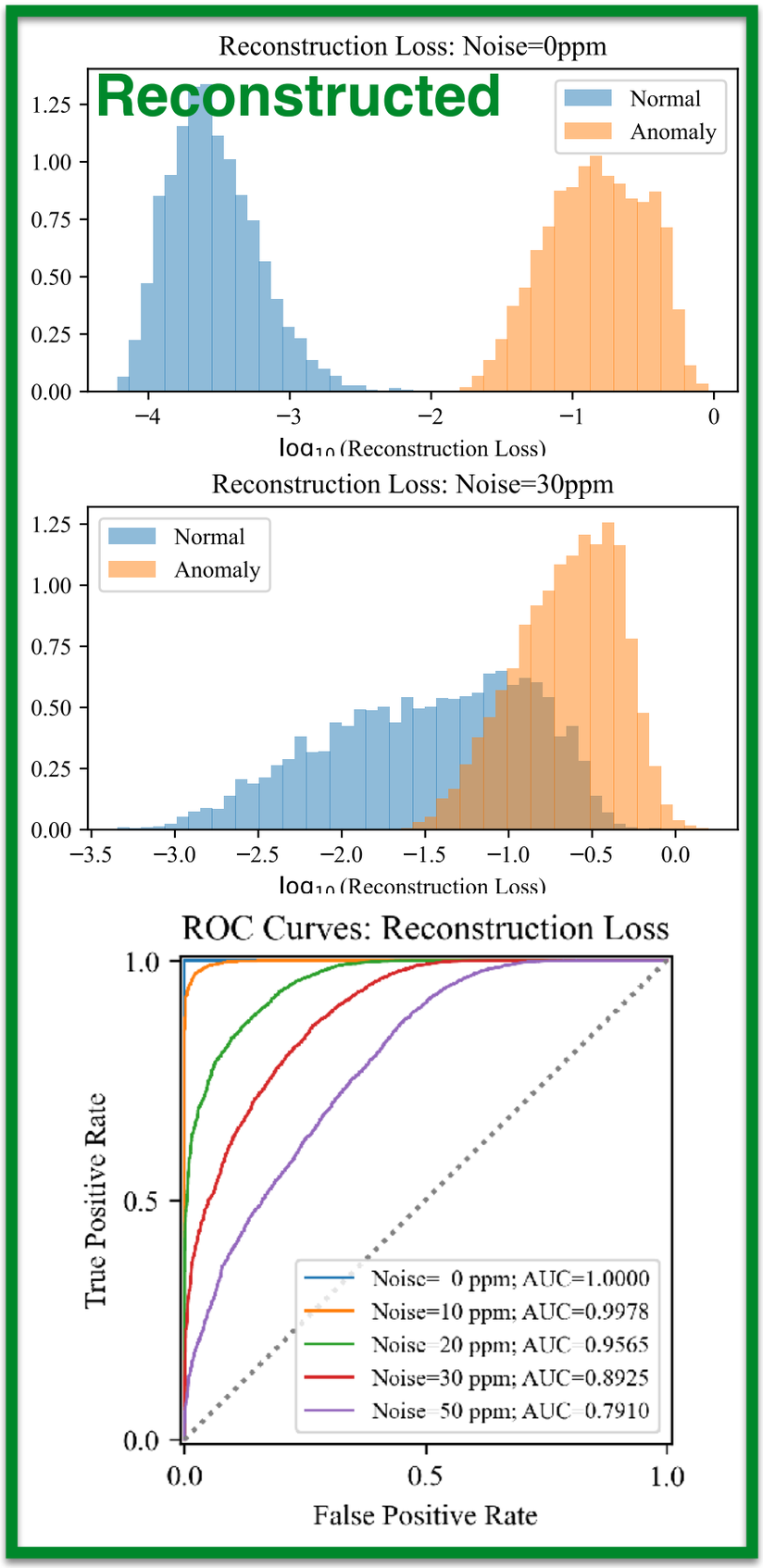




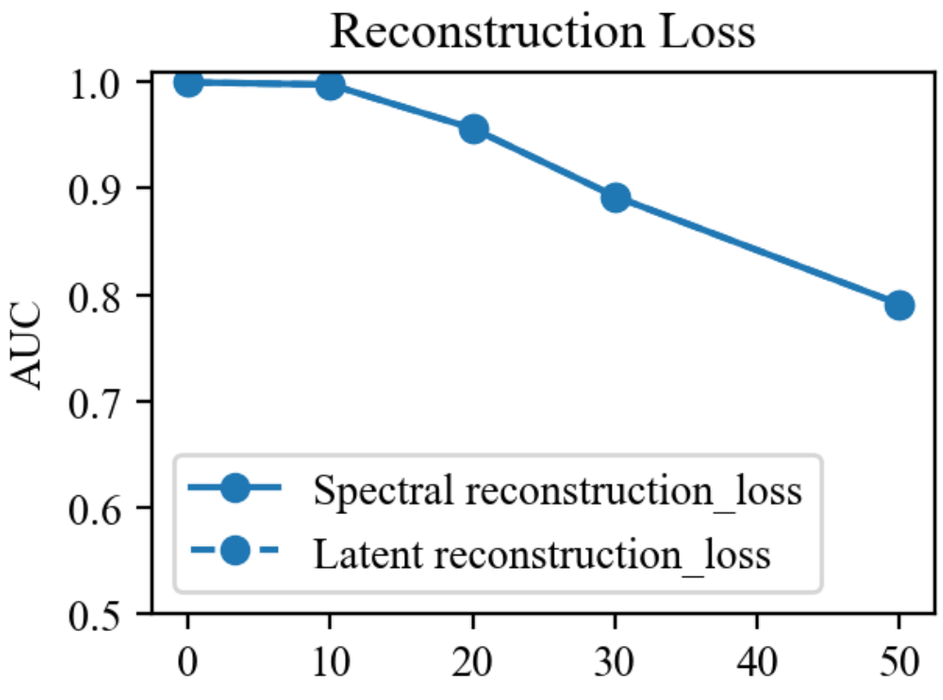
Input vs Latent vs Output

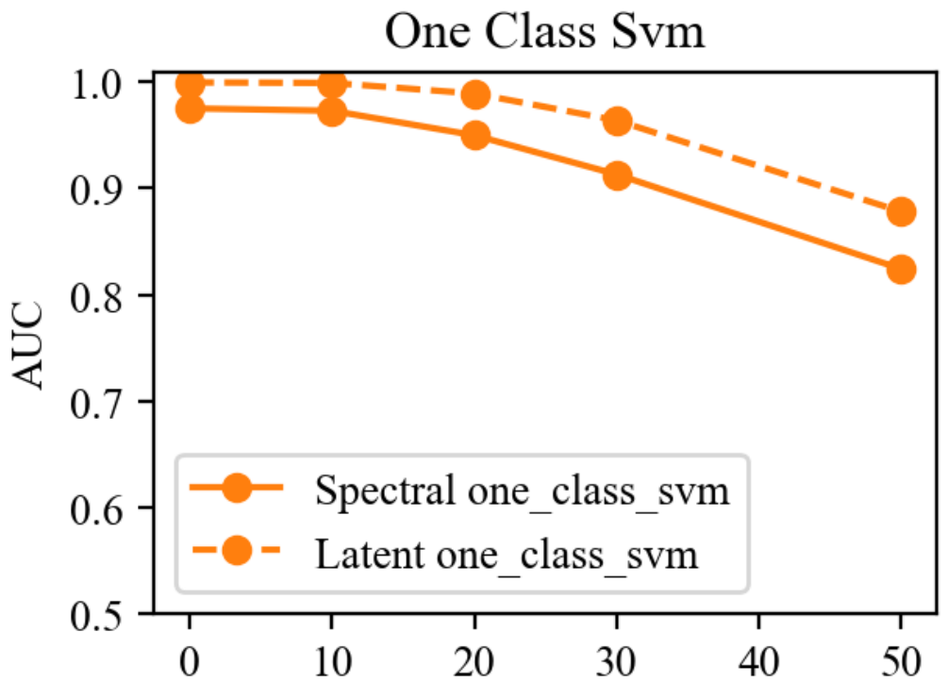


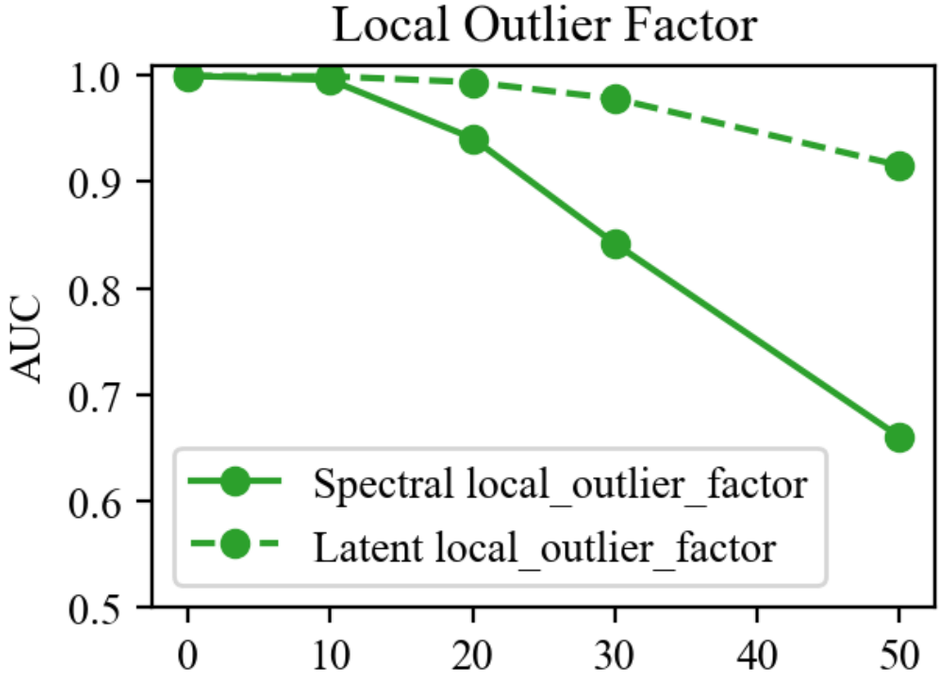


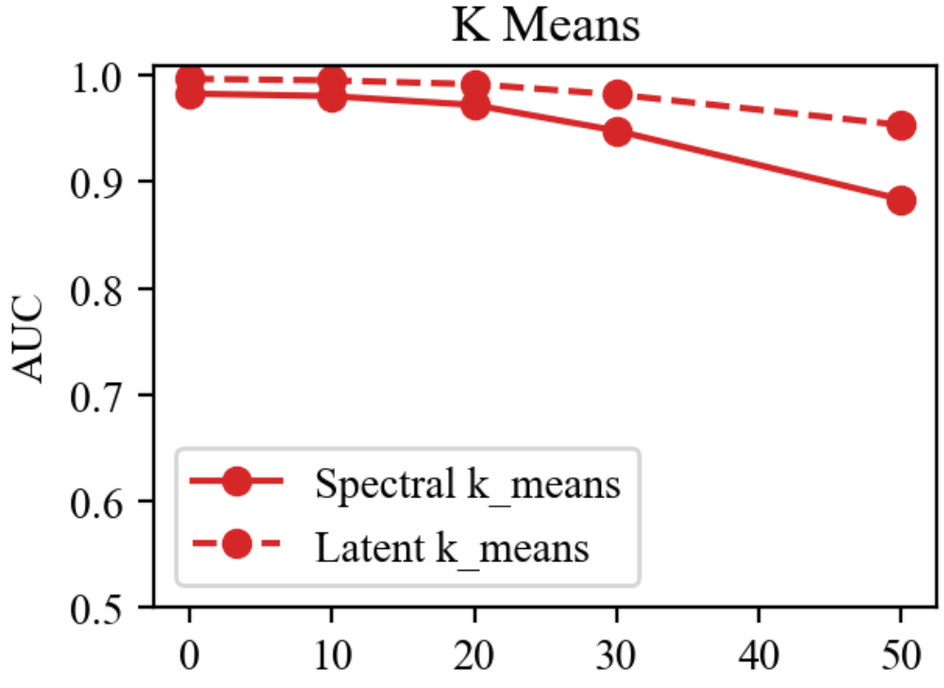


Performance vs Noise

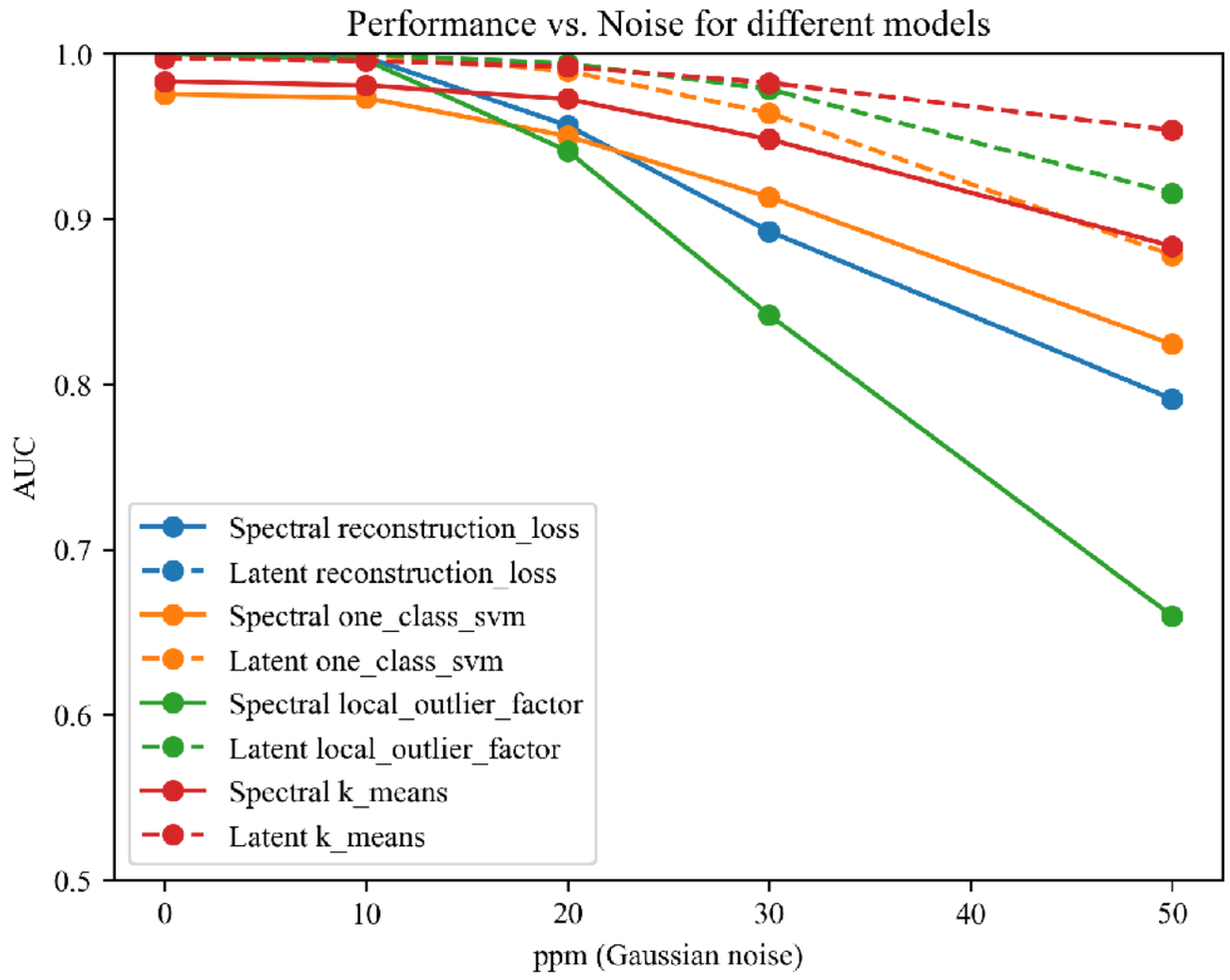








The Money Plot



- Latent space provides a low dimensional representation of planet spectra.
- Strong spectroscopic signal can be easily identified in all representations: original, latent, reconstruction.
- Weak or noisy spectroscopic signal benefits from analysis in latent space representation.

Summary

Transit spectroscopy works with **high-dimensional**, **highly-correlated** data. ML dimensionality reduction methods recast the data with minimal information loss.

Low-dimensional representation in terms of **PCA components** nicely resolves the individual classes of atmospheres with different **chemical composition**.

A **symbolic regression** trained on synthetic spectroscopic data is able to extract the analytical model used to generate the dataset.

Machine learning **anomaly detection** techniques can be used to flag exoplanet atmospheres with unusual **chemical composition**.

In the era of large planetary surveys, ML offers **fast and robust characterization** of the **planet spectra** and marks interesting cases for follow up in-depth studies.

Traditional **dimensional analysis** identifies the relevant dimensionless input variables and reveals the complete family of **degeneracies** among the inputs.