

Final Report for Computational Astrophysics

Course of 3rd semester - Physics of Data

Savina Tsichli, Marco Foster

University of Padova

8th July, 2025

ABSTRACT

Context. The study of exoplanets has become an important field in modern astrophysics, combining data science, simulation modeling, and observational techniques.

Aims. This project aims to explore the detection of exoplanets and the analysis of their atmospheres using computational methods.

Methods. Support Vector Machines and Neural Networks were used to classify synthetic transit light curves, while the TauREx retrieval framework was employed to estimate atmospheric parameters from real observational data.

Results. Both machine learning models showed strong performance in identifying planetary systems, and the retrieval analysis provided consistent estimates for key parameters such as temperature, planetary radius, and molecular abundances.

Conclusions. These results demonstrate the effectiveness of combining machine learning and retrieval techniques for both the detection and atmospheric characterization of exoplanets.

Key words. support vector machine, neural network, exoplanets, batman, taurex, transit, atmosphere, planetary system

1. Introduction

The discovery of exoplanets has rapidly evolved over the past decades, driven by ground-based surveys and space telescopes. Among the known exoplanet classes, hot Jupiters are particularly favorable for observation because of their inflated atmospheres, strong transit signatures, and short orbital periods.

MASCARA-3Ab, also known as KELT-24b, is a hot Jupiter discovered by the MASCARA project (Multi-site All-Sky CAM-era), which is designed to detect transiting planets around bright stars (Snellen et al. 2013), along with complementary efforts like KELT and TESS. Bright host stars are particularly valuable, as they allow for detailed atmospheric and orbital characterization. MASCARA-3Ab, the fourth planet identified by the survey, orbits a late F-type star (HD 93148). The system was first identified photometrically and confirmed via radial velocity measurements and spectroscopic follow-up using SONG. Its brightness and deep transits make it an ideal candidate for high-resolution transmission spectroscopy and spin-orbit alignment studies.

Its large radius and deep transit depth make it an excellent candidate for transit detection techniques and allow for testing machine learning algorithms on realistic synthetic light curves. Furthermore, orbiting its bright host star HD 93148, makes it favorable for further characterization. Studying such systems contributes to the understanding of the physical processes governing highly irradiated gas giants, such as inflation mechanisms and atmospheric escape.

WASP-39b is a hot Saturn as well, discovered by the WASP (Wide Angle Search for Planets) survey (Faedi et al. 2011). It exhibits a significantly bloated atmosphere and has one of the lowest densities measured for a gas giant. It orbits a G-type main-sequence star that has a high visual magnitude; the deep transits of WASP-39b have enabled high-precision spectroscopic

studies, including transmission spectroscopy with the Hubble Space Telescope (HST) and, more recently, the James Webb Space Telescope (JWST). Its chemical composition in the atmosphere makes WASP-39b a benchmark target for atmospheric retrievals; especially SO₂ which indicates photochemistry, provides a wealth of information and tests the capabilities of atmospheric models. Finally, since its observation by JWST was carried out using multiple optical instruments (eg. NIRISS, NIR-Cam, NIRSpec), it produced high signal-to-noise data across a wide wavelength range (Lueber et al. 2024) and thus, allowing the testing of retrieval models like TauREx.

Both MASCARA-3Ab and WASP-39b were utilized across the three assignments: MASCARA-3Ab for training and evaluating machine learning classifiers on synthetic transits, and WASP-39b for atmospheric retrievals based on real observed data, in order to meet the criteria in Assignment 3b.

By combining detection techniques and retrieval analysis, this study covers complementary phases in the characterization of exoplanets, from the identification of transiting candidates to the inference of their atmospheric properties.

2. Planetary system parameters

The planetary and stellar parameters used in the simulations were taken from the Extrasolar Planets Encyclopaedia (exoplanet.eu) and verified against relevant literature. In the final assignment, WASP-39b was chosen over MASCARA-3Ab as it met the following selection criteria: availability of a transmission spectrum, JWST observations, and a published peer-reviewed article.

As mentioned earlier, MASCARA-3Ab is a hot Jupiter orbiting the F-type main-sequence star HD 93148 with $V = 8.33$ and it has a short period of approximately 6 days. The host star has a mass of approximately $1.9 M_{\odot}$ and a radius of 1.6

R_{\odot} , with an effective temperature of about 9100 K (Giovinazzi et al. 2024). Due to the high temperature and brightness of its host star, detecting transits and secondary eclipses is challenging, as the planet-to-star flux ratio is relatively low. Nevertheless, MASCARA-3Ab is an interesting candidate for atmospheric studies due to its inflated radius and short orbital period.

Planet: MASCARA-3Ab

Parameter	Value
Status	Confirmed
Discovery Year	2019
Update Date	2025-02-11
Mass	$5.18 \pm 0.22 M_J$
Semi-Major Axis	$0.06971^{+0.0088}_{-0.0096}$ AU
Orbital Period	$5.5514926 \pm 1.8\text{e-}6$ d
Eccentricity	0.085 ± 0.023
ω (deg)	$41.0^{+14.0}_{-20.0}$
T_{peri} (JD)	2457146.6 ± 0.7
Radius	$1.272 \pm 0.022 R_J$
Inclination	$89.16^{+0.6}_{-0.77}$ °
Detection Method	Primary Transit
Mass Method	Radial Velocity
Radius Method	Primary Transit
Primary Transit	2457147.053 ± 0.002 JD
Secondary Transit	$2457144.422^{+0.006}_{-0.056}$ JD
Spin-Orbit Angle λ	$2.6^{+5.1}_{-3.6}$ °
Impact Parameter b	0.32 ± 0.01
Velocity Semiamp. K	403.0 ± 11.0 m/s
Calc. Temperature	1458 ± 16 K
Alt. Names	KELT-24 b, HD 93148 b

Star: MASCARA-3A

Parameter	Value
Distance	96.79 ± 0.03 pc
Spectral Type	F7 IV–V
Apparent Magnitude V	8.4
Mass	$1.309^{+0.052}_{-0.018} M_{\odot}$
Age	$2.81^{+4.0}_{-0.79}$ Gyr
Effective Temperature	6347^{+57}_{-64} K
Radius	$1.513^{+0.035}_{-0.031} R_{\odot}$
Metallicity [Fe/H]	$0.168^{+0.068}_{-0.058}$
RA (J2000)	10:47:38.0
Dec (J2000)	+71:39:21.0
Alt. Names	KELT-24, HD 93148

Table 1: Basic Parameters of the MASCARA-3 System

WASP-39b, on the other hand, orbits a G-type star every 4.05 days. The latter, has a temperature of approximately 5400 K and a radius of $0.9 R_{\odot}$. The planet is a hot Saturn with a mass of $0.28 M_J$ and a radius of $1.27 R_J$, leading to a low density and extended atmosphere that is particularly amenable to transmission spec-

troscopy. The system has been extensively studied, particularly following JWST observations that revealed prominent absorption features from H_2O , CO_2 , CO , and SO_2 (Carter et al. 2024). The lower stellar activity and favorable planet-star radius ratio make WASP-39b a benchmark target for atmospheric retrieval.

Planet: WASP-39b

Parameter	Value
Status	Confirmed
Discovery Year	2011
Update Date	2023-05-18
Mass	$0.28^{+0.03}_{-0.03} M_J$
Semi-Major Axis	0.0486 ± 0.0005 AU
Orbital Period	$4.055259 \pm 9\text{e-}6$ d
Eccentricity	0.0
ω (deg)	—
T_{peri} (JD)	—
Radius	$1.27^{+0.04}_{-0.04} R_J$
Inclination	87.83 ± 0.25 °
Detection Method	Primary Transit
Mass Method	—
Radius Method	—
Primary Transit	2455342.9688 ± 0.0002 JD
Secondary Transit	—
Spin-Orbit Angle λ	—
Impact Parameter b	—
Velocity Semiamp. K	—
Calc. Temperature	1120 K
Alt. Names	—

Star: WASP-39

Parameter	Value
Distance	230 ± 80 pc
Spectral Type	G8
Apparent Magnitude V	12.11
Mass	$0.93 \pm 0.03 M_{\odot}$
Age	—
Effective Temperature	5400 ± 150 K
Radius	$0.895 \pm 0.023 R_{\odot}$
Metallicity [Fe/H]	-0.12 ± 0.1
RA (J2000)	14:29:18.0
Dec (J2000)	-03:26:40.0
Alt. Names	—

Table 2: Basic Parameters of the WASP-39 System

3. Detection of transit light curves using Machine Learning

In this project, we explored the use of Machine Learning algorithms to detect exoplanetary transits from light curves, on synthetic and real observational data. The dataset was derived from the NASA Kepler mission and consisted of flux measurements

over time for thousands of stellar systems. Each system is labeled as either a "system with planet" (label 2) or a "system without planet" (label 1), enabling supervised learning.

The original dataset without injected planets (kepler/data_no_injection) includes 5087 training examples and 570 validation examples, with a strong class imbalance (only 37 systems with planets in the training set and 5 in the validation set). We also used a second version of the dataset with synthetic transits injected (kepler/data_injected) to evaluate model performance under more realistic transit frequencies.

Before feeding the data to the models, we applied several preprocessing steps to make the light curves machine-readable. This included: Fourier transformation to extract frequency-domain features, normalization to scale the flux values, Gaussian filtering to smooth the signal and suppress noise, and standardization to center and scale the features. These steps ensured that the models could learn more effectively from subtle variations caused by planetary transits.

Both SVMs and NNs were trained on the preprocessed data. Despite the challenges posed by class imbalance, we evaluated model performance using confusion matrices and precision scores, showing that NNs typically outperformed SVMs, especially when Dropout layers and ReLU activations were added to mitigate overfitting and improve learning stability.

3.1. Methodology

The primary challenge in exoplanet detection is the high rate of false positives. Many datasets, such as `data_no_injection`, are highly imbalanced, containing far more non-transiting systems than confirmed exoplanets. This imbalance can lead to poor model performance on the minority (planet) class. False positives often come from phenomena like binaries, stellar variability (eg. flares or rotational modulation), or instrumental noise that mimic transit signals.

To address these issues, we implemented two machine learning approaches; SVMs and NNs. Both methods are well-suited for transit photometry, as they can detect subtle, complex, and non-linear patterns in noisy light curves.

SVMs are effective classifiers, especially in high-dimensional spaces, and can handle small datasets well. We evaluated them using different kernel functions (linear, radial basis function, and polynomial) to assess their ability to separate transit-like events from noise. NNs, on the other hand, are more flexible and capable of learning hierarchical representations directly from raw input features. In this project, we enhanced the neural network's performance using ReLU activations, dropout layers to prevent overfitting, and SMOTE (Synthetic Minority Oversampling Technique) to balance the classes in the training data.

SVMs were applied to light curve data in order to distinguish planetary transits from non-transit signals. To account for non-linear separability in the data, different kernel functions were employed: linear, radial basis function (RBF), and polynomial kernels. These kernels allow the SVM to map input features into higher-dimensional spaces, where complex boundaries can be formed to better distinguish between classes. However, as deduced in this report, they can struggle with class imbalance, where one class is severely underrepresented. In such cases, the model may favor the dominant class, resulting in high accuracy but poor recall for detecting actual exoplanets.

To evaluate model performance across kernels, confusion matrices and precision scores were computed for both the training and development (dev) datasets.

On the other hand, NNs can learn hierarchical features directly from pixel-level light curves, helping reduce noise and improve classification accuracy (Tey et al. 2023). In this project, the NN was implemented in TensorFlow using a fully connected architecture. It began with an input layer followed by two hidden layers: the first with 128 neurons and the second with 64, both using the ReLU activation function. ReLU introduces non-linearity while avoiding vanishing gradient issues that can hinder learning with older functions like sigmoid.

The second hidden layer also reduced the dimensionality of features learned in the first, making the model more efficient while retaining its ability to learn complex patterns. Dropout layers with a 0.3 rate were added after each hidden layer to randomly deactivate neurons during training, helping the network generalize better and avoid overfitting.

The output layer had a single neuron with a sigmoid activation function. The model was trained using the Adam optimizer, which adapts learning rates to converge faster in noisy data. Binary cross-entropy was used as the loss function, penalizing incorrect predictions based on their confidence. Finally, to address the significant class imbalance in the dataset (where real planetary transits were underrepresented), the SMOTE algorithm was applied to generate synthetic positive examples, helping the model learn from both classes more effectively.

3.2. Support Vector Machines Results

SVMs were applied to the Kepler datasets to classify light curves as either containing a planetary transit or not. Three different kernels were tested: linear, gaussian and polynomial, using the scikit-learn SVC implementation.

For all models, default hyperparameters were used except for the polynomial kernel where the degree was set to 4. The evaluation was performed using confusion matrices and precision scores, computed on both training and validation sets. Precision was chosen as the key metric due to the strong class imbalance in the dataset, where only a small fraction of systems contain planets.

The table below presents the performance metrics for SVM models with linear, RBF, and polynomial kernels. It includes columns for Kernel, Accuracy, Precision, Recall, and Confusion Matrix.

Kernel	Accuracy	Precision	Recall	Conf. Matrix
Linear	0.45	0.5586	0.5421	$\begin{pmatrix} 201 & 245 \\ 262 & 310 \end{pmatrix}$
RBF	0.59	0.5632	0.9965	$\begin{pmatrix} 4 & 442 \\ 2 & 570 \end{pmatrix}$
Polynomial	0.58	0.5620	0.9982	$\begin{pmatrix} 1 & 445 \\ 1 & 571 \end{pmatrix}$

Table 3: Comparison of SVM kernel.

The RBF and polynomial kernels achieved high recall but low precision, indicating a tendency to over-classify non-transit signals as transits, resulting in a large number of false positives. While the linear kernel performed slightly better in terms of precision, it did not do well in overall classification accuracy. Although the Gaussian and polynomial kernels were somewhat better at capturing non-linear patterns, all SVM models showed clear limitations when applied to the injected dataset.

These results show that there is a trade-off between how complex the SVM kernel is and how well it generalizes to new data.

3.3. Artificial Neural Networks

The NN was implemented using TensorFlow and consisted of a fully connected architecture with three layers. The input layer was followed by two hidden layers with 128 and 64 neurons, respectively, each followed by a ReLU activation function. ReLU was chosen due to its efficiency in training deep networks and its ability to mitigate vanishing gradient problems. To improve generalization and mitigate overfitting, dropout layers with a rate of 0.3 were added after each hidden layer. The final output layer consisted of a single neuron with a sigmoid activation function, suitable for binary classification.

The model was compiled using the Adam optimizer and trained using the binary cross-entropy loss function, appropriate for imbalanced binary classification tasks. Training was performed over 50 epochs with a batch size of 32, a commonly effective configuration that balances convergence speed and stability. To address the severe class imbalance in the training set, the SMOTE algorithm was applied to oversample the minority class (planetary systems), helping the model better learn from underrepresented examples.

This configuration was evaluated on the dataset with injected synthetic transits `data_injected`. Evaluation metrics included accuracy, precision, recall, and confusion matrices for both training and validation sets.

The table below presents the performance metrics for NN.

Dataset	Accuracy	Precision	Recall	Conf. Matrix
Train	0.5821	0.5789	0.9423	$\begin{pmatrix} 265 & 1961 \\ 165 & 2696 \end{pmatrix}$
Dev	0.5544	0.5544	1.0000	$\begin{pmatrix} 0 & 254 \\ 0 & 316 \end{pmatrix}$

Table 4: NN Performance on Dataset with Injected Transits

The NN showed strong sensitivity to planet transit signals, achieving high recall across both training and validation sets. However, this came at the cost of precision, with the model overpredicting positives and generating many false ones. As the confusion matrices indicate, further improvements could include threshold tuning or structural refinements to reduce false positives.

3.4. Results

Both SVMs and NNs were used on the `data_no_injection` and `data_injected` datasets to assess how well transit-like patterns could be recognized in the data.

On `data_no_injection`, the SVM with a linear kernel achieved a dev precision of 0.56, while the NN reached a slightly lower precision of 0.42. However, the NN perfectly recalled all true positives in the training set, suggesting strong memorization but poorer generalization. On the other hand, SVMs with polynomial and Gaussian kernels produced extremely skewed predictions, classifying nearly all examples as negative or positive, leading to dev precision values around 0.56 and limited practical usefulness.

On `data_injected`, the NN achieved a dev precision of 0.55 and detected all positive cases (recall = 1), although at the cost of high false-positive rates, reflected in a dev confusion matrix with no true negatives. This result highlights the neural network's sensitivity to transit-like patterns, even in noisy light curves, but also its vulnerability to overprediction. The SVMs performed poorly on this dataset, often failing to balance precision and recall and showing signs of underfitting or model inflexibility.

Overall, the NN handled the complex signals in the injected dataset better than the SVM. It showed a more balanced performance between correctly finding planets (recall) and avoiding false ones (precision). Both models had trouble with false positives, which is common in exoplanet detection as seen in the mentioned papers previously, but the NN's design with ReLU activations, dropout layers, and SMOTE, which helped it learn transit patterns more effectively, by balancing the two classes. Future improvements could include adjusting the prediction threshold, changing the network structure, or trying convolutional networks to better detect time-based features.

4. Atmospheric Studies on MASCARA-3Ab & WASP-39b

4.1. MASCARA-3Ab

In this report, MASCARA-3Ab was used in order to work on synthetic spectra. Since it lacks atmospheric spectra, a synthetic transmission spectrum was generated based on published planetary and stellar parameters. A clear, isothermal atmosphere with solar composition was assumed. The inputs used are the following:

- **Planet radius:** $1.272 R_J$
- **Planet mass:** $5.18 M_J$
- **Stellar radius:** $1.513 R_\odot$
- **Temperature:** 1458 K (equilibrium temperature)
- **Pressure range:** 10^{-4} to 10^6 Pa
- **Chemistry:** H_2O , CH_4 , CO_2 , CO

Each parameter plays a specific role in shaping the synthetic transmission spectrum. The planetary radius, mass, and stellar radius define the scale height and transit depth. The assumed molecules H_2O , CH_4 , CO_2 and CO , were selected based on their expected abundance. The model assumes a clear, isothermal atmosphere with solar composition, which simplifies the chemistry and radiative transfer while still producing realistic spectral features.

4.2. WASP-39b

WASP-39b has been among the first exoplanets observed by JWST. The planet was discovered by the groundbased transit survey SuperWASP in 2011 and their analysis indicated WASP-39b to be a highly inflated Saturn-like exoplanet. Various molecules were reported in the literature (Ma et al. 2025), to be present in its atmosphere, including H_2O , CO , CO_2 , K , Na , H_2S , CH_4 and tentative detections of SO_2 . The latter publication also suggests that in addition to H_2O , CO_2 , Na and K , silicon-based chemistry plays a major role in shaping the chemistry and condensates of the atmosphere of WASP-39b. The parameters used in this report are the following:

- **Planet radius:** $1.27 R_J$

- **Planet mass:** $0.28 M_J$
- **Stellar radius:** $0.895 R_\odot$
- **Surface gravity:** computed from mass and radius
- **Pressure range:** 10^{-4} to 10^6 Pa
- **Chemistry:** H_2O , CO_2 , CO , SO_2
- **Temperature profile:** isothermal, with temperature as a free parameter

4.3. Model

Atmospheric models used in exoplanet studies aim to simulate the transmission of starlight through a planet’s atmosphere during transit.

This has been approached using various machine learning techniques, including supervised models like SVMs and NNs. These typically require labeled datasets with both transit and non-transit examples. An alternative strategy, would be the use of a One-Class SVM trained solely on positive (transit) (Roche 2024). This approach is especially helpful when there are few or no labeled negative examples, and it performs well in terms of both precision and speed. Although this study focuses on supervised SVMs and NNs, unsupervised methods can also help by quickly spotting potential signals and narrowing down large datasets early on.

These supervised models mentioned above typically assume a 1D, plane-parallel geometry, hydrostatic equilibrium, and local thermodynamic equilibrium. A commonly used configuration is an isothermal temperature profile, which simplifies the radiative transfer but may not capture vertical temperature gradients present in real atmospheres.

In the case of WASP-39b, in Carter et al. 2024 TauREx-3 was used with a parametric temperature profile with uniform abundances throughout the atmosphere. The model included key absorbers (H_2O , CO_2 , CO , CH_4 , Na , K , SO_2), opacity sources from ExoMol, HITRAN, and CIA databases, and accounted for Rayleigh scattering and cloud/haze opacity. It assumed a 1D, hydrostatic atmosphere in chemical equilibrium. Limitations include the lack of photochemistry, atmospheric mixing, and 3D structure, which are important given the likely photochemical origin of SO_2 .

Additionally, in Rustamkulov et al. 2023 other models of interpretation of the spectra were used; ScCHIMERA, PICASO 3.0, ATMO, and PHOENIX. These models treat the atmosphere as a single vertical slice and assume it is stable, with energy flowing through it by radiation and convection. They also assume chemical equilibrium and hydrostatic balance, without accounting for winds or mixing. ScCHIMERA and PICASO 3.0 used Bayesian retrievals, while ATMO and PHOENIX relied on grid search fitting. SO_2 and cloud opacity were added during post-processing to better match observations. The main limitations of these models are their 1D structure, which prevents them from capturing horizontal and vertical variations in the atmosphere, and the absence of non-equilibrium processes such as photochemistry and vertical mixing.

Another study on WASP-39b (Nikolov et al. 2016) presents a detailed optical to near-infrared transmission spectrum using mainly HST data. A forward atmospheric model was used assuming solar composition and chemical equilibrium. The model includes Rayleigh scattering, molecular absorption (mainly from H_2O), and collision-induced absorption from H_2 - H_2 and H_2 -He. The atmosphere is assumed to be isothermal with a constant gravity profile and no clouds or hazes initially. The comparison between model predictions and observed data revealed a significant broad water absorption, consistent with a solar-abundance,

clear atmosphere. However, some discrepancies, especially the omission of broad sodium and potassium absorption, implied possible presence of clouds or sub-solar elemental abundances. The key assumptions of the model include chemical equilibrium, solar composition, and a homogeneous, cloud-free atmosphere, all of which simplify the physics but may not capture real atmospheric complexity. The limitations concern the model’s 1D nature, lack of vertical mixing or photochemistry, and absence of clouds or aerosols in its baseline configuration. While it provides a reasonable fit for the infrared data, it struggles to reproduce certain optical features, meaning that there is need for more advanced models that incorporate the values omitted.

For this study, both the synthetic and observed spectrum retrievals were carried out using TauREx-3, a Bayesian retrieval framework. We used:

- An isothermal temperature profile
- H_2 -He dominated atmosphere with trace absorbers (H_2O , CH_4 , CO , CO_2 , SO_2)
- 100 logarithmically spaced pressure layers from 1 to 10^6 Pa
- CIA and Rayleigh scattering contributions
- Nested sampling via *nestle* with 100 live points

This configuration strikes a balance between computational efficiency and physical accuracy. However, it introduces limitations: real atmospheres may have temperature gradients, clouds/hazes, and vertical chemical variation, which are not captured by our simplified assumptions. Moreover, fixing the abundances of several molecules (e.g., CH_4 , CO , CO_2) reduces the dimensionality of the fit but also risks biasing the retrieved parameters if those species have significant spectral contributions.

Future models may benefit from more flexible temperature profiles, cloud parameterizations, or even 2D/3D modeling, though these approaches require higher signal-to-noise data and substantially more computational power.

5. Retrieval

Spectral retrieval methods have often been used in recent years to interpret exoplanetary atmospheric spectra. These algorithms use statistical techniques and high-performance computers to sample a broad parameter space and to search for optimised solutions through a considerable number of iterations. Robust convergence, especially in high-dimensional retrievals, requires a large number of iterations and steps, making the computing power requirements more demanding.

TauREx is a state-of-the-art modeling framework to simulate exoplanetary atmospheres and to interpret exoplanet atmospheric data collected with different techniques through inverse models based on Bayesian statistics. Several plugins can be activated to simulate, eg. atmospheric chemistry, cloud microphysics, stellar activity and phase-curves interpretation. Parameters that are included and tested in retrievals with TauREx-3 include but are not limited to instrumental, atmospheric thermal, chemical, and cloud profiles, as well as planetary and stellar parameters.

To infer atmospheric parameters from transmission spectra, we performed atmospheric retrievals using TauREx-3. Two cases were considered; a synthetic spectrum for MASCARA-3Ab and an observed spectrum for WASP-39b. Both retrievals assumed an isothermal atmosphere in hydrostatic equilibrium. The opacity database included contributions from key molecules, collision-induced absorption (CIA), and Rayleigh scattering. For each planet, the configuration specified:

- **Pressure layers:** 100, logarithmically spaced from 1 to 10^6 Pa
- **Chemistry:** H_2 and He dominated, also H_2O , CH_4 , CO , CO_2 , SO_2
- **Temperature profile:** Isothermal
- **Optimizer:** Nested sampling via *nestle*, using 100 live points
- **Retrieved parameters:** R_p , T , H_2O volume mixing ratio

All other gases were included but kept fixed to reduce computational complexity.

5.1. MASCARA-3Ab

The synthetic transmission spectrum for MASCARA-3Ab was generated assuming solar abundances. The transmission spectrum used for the retrieval is shown in Fig. [1]. The parameters were fixed and only H_2O , the radius and the temperature of the planet were retrieved, to minimize computational power. TauREx-3 successfully retrieved the main atmospheric parameters:

- **Planetary radius:** $R_p = 1.27^{+0.00}_{-0.00}, R_J$
- **Temperature:** $T = 1457.6^{+10.7}_{-8.7}$ K
- **$\log(\text{H}_2\text{O})$:** $-3.92^{+0.02}_{-0.02}$

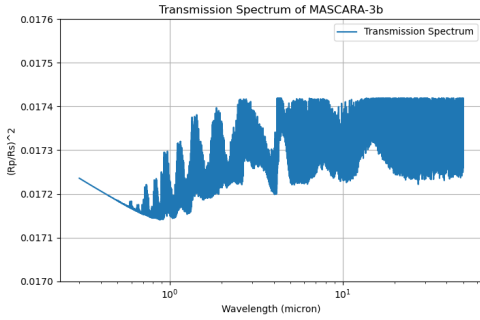


Fig. 1: Synthetic transmission spectrum of MASCARA-3Ab.

As shown in Fig. [2] of the posterior distributions, the TauREx retrieval gave confident estimates for the planet’s temperature and water content. These estimates are not completely independent—changing one slightly affects the other—but the retrieval still separated them reasonably well.

Since the derived spectrum closely matched the input, confirming the robustness of the forward model and the retrieval. The retrieved radius was consistent with the input, and the water signal was accurately recovered.

5.2. WASP-39b

For WASP-39b, we retrieved atmospheric parameters from the JWST spectrum using a similar setup. Only H_2O was fitted, while CO_2 , CO , and SO_2 were fixed to literature values. The retrieval produced:

- **Planetary radius:** retrieved within bounds
- **Temperature:** close to 1100 K
- **$\log(\text{H}_2\text{O})$:** consistent with literature value ~ 2.5 to -3

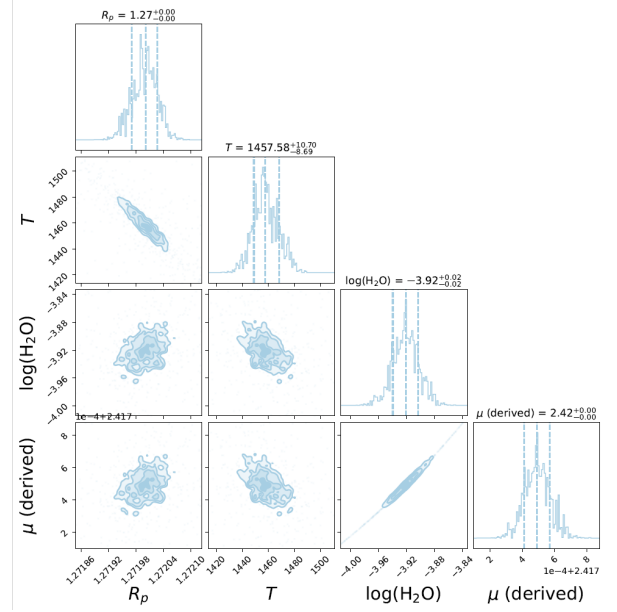


Fig. 2: Posterior distributions for retrieval of MASCARA-3Ab.

First, the transmission spectrum was produced, shown in Fig.[3]. Then, the atmospheric retrieval of its real spectrum was performed. The retrieved spectrum Fig.[??] captured the broad H_2O absorption features visible in the observed data. Even when fixing some molecular abundances, the model provided a good fit, illustrating TauREx’s sensitivity to H_2O even under constrained setups.

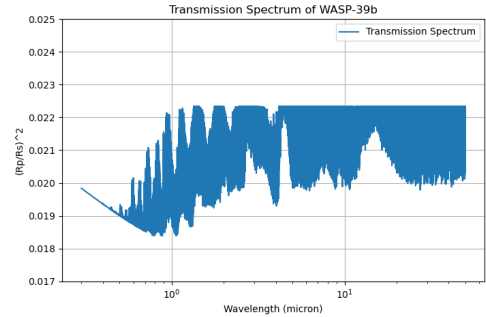


Fig. 3: Transmission spectrum of WASP-39b.

The retrievals demonstrate that even simplified setups can extract key parameters such as temperature and H_2O abundance with good accuracy. For synthetic data, the retrieval recovered the input parameters with high precision. For observed data, parameter correlations and fixed species introduce uncertainty, but TauREx-3 still retrieved physically consistent values.

Sensitivity to initial assumptions (e.g., isothermal profile, fixed gas abundances) can impact the fit quality and uncertainty bounds. More advanced profiles or inclusion of clouds could improve realism but increase model complexity.

The retrieval results depend on how the model is set up. Simpler assumptions like a constant temperature and fixed gases make it easier to compute, but they may oversimplify the real atmosphere. More advanced models (eg. including clouds, or non-isothermal temperature profiles), could give more realistic results, but are harder to manage and may introduce more uncertainties, if not constrained by high-quality data.

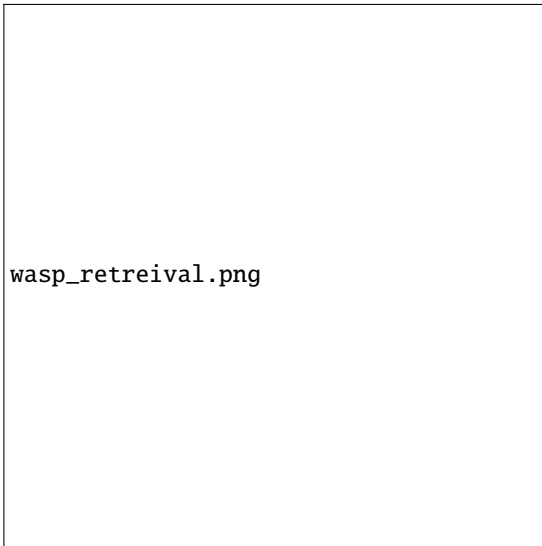


Fig. 4: Posterios Distributions for retrieval of WASP-39b.

6. Conclusions

This project explored the integration of machine learning techniques and atmospheric modeling for exoplanet detection and characterization. The workflow consisted of two main components: transit detection using SVMs and NNs, and atmospheric retrieval using TauREx.

In the detection part, both SVMs and NNs were applied to the Kepler light curve datasets. The NN demonstrated better adaptability to noisy and imbalanced data, achieving higher recall on synthetic transit injections. However, both models struggled with precision due to false positives, a known challenge in transit detection from light curves. The application of data processing techniques (eg. Fourier transforms, Gaussian smoothing, normalization) and SMOTE oversampling proved essential to improving model sensitivity.

In the atmospheric analysis, transmission spectra (both synthetic and observed) were modeled using TauREx. For MASCARA-3Ab, a synthetic spectrum was generated under the assumption of a solar-composition, isothermal atmosphere. For WASP-39b, real JWST transmission data were used for atmospheric retrieval. The retrieval successfully estimated parameters like the planetary radius, atmospheric temperature, and H₂O abundance. Posterior distributions showed well-constrained results, with sensitivity to assumptions such as fixed gas abundances and the isothermal profile.

Key scientific insights include the potential of ML to automate and accelerate exoplanet vetting, as well as the importance of physically informed models in interpreting atmospheric signals. The combination of ML classifiers for detection and Bayesian retrievals for characterization forms a robust pipeline for exoplanet science.

Future improvements could include:

- Using larger and more diverse training datasets (eg. TESS, PLATO)
- Incorporating convolutional or recurrent neural networks for temporal feature extraction
- Expanding atmospheric models to include clouds, thermal gradients, or 3D effects

As upcoming missions like PLATO, Ariel, and continued JWST observations provide higher-precision data, these techniques will become increasingly valuable. They offer a scalable

way to process large amounts of light curves and spectra, helping to study a wide range of planetary atmospheres and search for possible biosignatures. Projects like ExoClock, which help keep transit timings up to date, can also benefit from fast and accurate detection methods like One-Class SVMs, making follow-up observations more efficient.

References

- Carter, A. L., May, E. M., Espinoza, N., Welbanks, L., Ahler, E., Alderson, L., and ... (2024). A Benchmark JWST Near-Infrared Spectrum for the Exoplanet WASP-39b. *Nature Astronomy*. Includes spectroscopic detection of H₂O, CO₂, SO₂, Na, K, CO.
- Faedi, F., Barros, S. C. C., Anderson, D. R., ... Simpson, E. K., et al. (2011). WASP-39b: a highly inflated Saturn-mass planet orbiting a late G-type star. *Astronomy Astrophysics*, 531:A40.
- Giovannazzi, M. R., Cale et al., B., Eastman, J. D., Rodriguez, J. E., Blake, C. H., Stassun, K. G., Vanderburg, A., Kunitomo, M., Kraus, A. L., Twicken, J., Beatty, T. G., Dedrick, C. M., Horner, J., Johnson, J. A., Johnson, S. A., McCrady, N., Plavchan, P., Sliski, D. H., Wilson, M. L., Wittenmyer, R. A., Wright, J. T., Johnson, M. C., Rose, M. E., and Cornachione, M. (2024). Trials and tribulations in the reanalysis of keck-24 b: A case study for the importance of stellar modeling. *The Astronomical Journal*, 168(3):118.
- Lueber, A., Novais, A., Fisher, C., and Heng, K. (2024). Information content of jwst spectra of wasp-39b.
- Ma, S., Saba, A., Al-Refaie, A. F., Tinetti, G., Yurchenko, S. N., Tennyson, J., and Pestellini, C. C. (2025). A new look into the atmospheric composition of wasp-39 b.
- Nikolov, N., Sing, D. K., Gibson, N. P., Fortney, J. J., Evans, T. M., Barstow, J. K., Kataria, T., and Wilson, P. A. (2016). Vlt foris2 comparative transmission spectroscopy: Detection of na in the atmosphere of wasp-39b from the ground. *The Astrophysical Journal*, 832(2):191.
- Roche, J. (2024). Using a one-class svm to optimize transit detection. *The Open Journal of Astrophysics*, 7.
- Rustamkulov, Z., Sing, D. K., Mukherjee, S., May, E. M., Kirk, J., Schlawin, E., Line, M. R., Piaulet, C., Carter, A. L., Batalha, N. E., Goyal, J. M., López-Morales, M., Lothringer, J. D., MacDonald, R. J., Moran, S. E., Stevenson, K. B., Wakeford, H. R., Espinoza, N., Bean, J. L., Batalha, N. M., Benneke, B., Berta-Thompson, Z. K., Crossfield, I. J. M., Gao, P., Kreidberg, L., Powell, D. K., Cubillos, P. E., Gibson, N. P., Lecote, J., Molaverdikhani, K., Nikolov, N. K., Parmentier, V., Roy, P., Taylor, J., Turner, J. D., Wheatley, P. J., Aggarwal, K., Ahler, E., Alam, M. K., Alderson, L., Allen, N. H., Banerjee, A., Barat, S., Barrado, D., Barstow, J. K., Bell, T. J., Blečić, J., Brande, J., Casewell, S., Changeat, Q., Chubb, K. L., Crouzet, N., Daylan, T., Decin, L., Désert, J., Mikal-Evans, T., Feinstein, A. D., Flagg, L., Fortney, J. J., Harrington, J., Heng, K., Hong, Y., Hu, R., Iro, N., Kataria, T., Kempton, E. M.-R., Krick, J., Lendl, M., Lillo-Box, J., Louca, A., Lustig-Yaeger, J., Mancini, L., Mansfield, M., Mayne, N. J., Miguel, Y., Morello, G., Ohno, K., Palle, E., Petit dit de la Roche, D. J. M., Rackham, B. V., Radica, M., Ramos-Rosado, L., Redfield, S., Rogers, L. K., Shkolnik, E. L., Southworth, J., Teske, J., Tremblin, P., Tucker, G. S., Venot, O., Waalkes, W. C., Welbanks, L., Zhang, X., and Zieba, S. (2023). Early release science of the exoplanet wasp-39b with jwst nirspec prism. *Nature*, 614(7949):659–663.
- Snellen, I. A. G., Stuijk, R., Navarro, R., Bettonvil, F., Kenworthy, M., de Mooij, E., Otten, G., and le Poole, R. (2013). MASCARA: The multi-site all-sky camera. 8444:844401.
- Tey, E., Moldovan, D., Kunitomo, M., Huang, C. X., Shporer, A., Daylan, T., Muthukrishna, D., Vanderburg, A., Dattilo, A., Ricker, G. R., and Seager, S. (2023). Identifying exoplanets with deep learning. v. improved light-curve classification for tess full-frame image observations. *The Astronomical Journal*, 165(3):95.