# Exoplanet Atmospheric Retrieval on Ariel Big Data Challenge dataset and WASP 39-b using Machine Learning Techniques

Rashmi R[1*], Shobha T[1], Hamshith A M[1], Harimedha S C[1], Gagandeep Y[1], Dhruv H[1]

[1*]Department of Information Science and Engineering, B.M.S. College of Engineering, Bull Temple Road, Bengaluru, 560019, Karnataka, India .

*Corresponding author(s). E-mail(s): rashmir.ise@bmsce.ac.in; Contributing authors: shobha.ise@bmsce.ac.in; hamshitham.is23@bmsce.ac.in; harimedhas.is23@bmsce.ac.in; gagandeepy.is23@bmsce.ac.in; dhruvh.is23@bmsce.ac.in;

**Abstract**

The development of high resolution observatories such as the James Webb Space Telescope (JWST) have increased the demand for rapid and scalable techniques for exoplanet atmospheric characterization.Conventional Bayesian retrieval methods are computationally expensive which have inspired the use of supervised machine learning for the estimation of atmospheric abundances. In this study we evaluated the correctness of the predictions using synthetic spectra from the Ariel Big Challenge (ABC) dataset and processed JWST NIR-Spec/PRISM observations of WASP 39b. This study provides a systematic benchmarking of preprocessing strategies across three ML regressors, demonstrating that sample wise normalization is the dominant factor that governs model stability and prediction accuracy a contribution that has received limited attention in prior atmospheric retrieval literature. Three regression models (MLP, SVR, and XGBoost) were trained under similar conditions to isolate the influence of preprocessing choices. The optimal setting entailed normalization and logarithmization of the abundances, where XGBoost performed optimally suggesting that tree based models are more robust to generalize on real world JWST observations from synthetic compared to MLP. In addition, the model predictions correlated qualitatively with the actual atmospheric composition comprising the gases $H_2O$, $CO_2$, CO, $CH_4$, and $NH_3$, and temperature for WASP 39b dataset from JWST telescope, indicating its applicability. The estimated value of the temperature is 958.40 K, more detailed tables relating to temperatures and gas estimates can be seen in the results section. In general, the results illustrate the efficiency of machine learning along with the atmospheric retrieval process.

**Keywords:** XGBoost for Atmospheric Retrieval, Support Vector Regression for Atmospheric Retrieval, Multilayer Perceptron for Atmospheric Retrieval, Ariel Big Challenge, WASP 39b, Transmission Spectroscopy

# 1 Introduction

Exoplanetary atmosphere characterization has advanced rapidly over the past decade, driven by improvements in space based spectroscopic instrumentation and atmospheric retrieval techniques. Transmission spectroscopy has been crucial to this progress, enabling constraints on atmospheric chemical composition by measuring wavelength dependent variations in stellar flux during planetary transits [1]. Prior to the JWST, observations were limited to narrow spectral windows accessible with the Hubble Space Telescope (HST) and ground based facilities, imposing fundamental limitations on abundance estimation and introducing degeneracies in retrieval models[1]. With JWST now operational, high-precision broad-band spectra spanning the near and mid-infrared are achievable, enabling simultaneous detection of multiple atmospheric species including $H_2O$, $CO_2$, CO, $CH_4$, and $NH_3$ in benchmark targets such as WASP 39b[2–4]. However, the growing volume and quality of JWST spectra demand retrieval approaches that are both computationally scalable and robust to preprocessing choices, a challenge that conventional Bayesian methods, despite their physical interpretability, are ill suited to meet at population scale.

Conventional methods of atmospheric retrieval use Bayesian inference and radiative-transfer to diagnosis to retrieve the atmospheric structure and composition using observed spectra [5]. Although these techniques offer physically interpretable posterior distributions, they are costly to compute [6], and the scale of such computations is not favorable when using large datasets or large grids of models, especially when atmospheric complexity or non-equilibrium chemistry is taken into account. The growing amount and quality of spectroscopic measurements with JWST and upcoming missions like Ariel is driving the discussion of alternative algorithms that can minimize the level of computational burden and predictive error.

Machine learning has become an alternative or even complementary avenue to atmospheric retrieval. Most recent works have shown that it can be used to map spectral features of physical atmospheric parameters with supervised regression models, emulators, and neural network-based uncertainty frameworks [7, 8]. These methods allow reducing the inference time by a considerable margin after training, allowing them to be used as well in population scale studies and high priority analysis[7, 9]. Nevertheless, even the work of ML models is susceptible to the formation of datasets, normalization, and the selection of architecture, which are under explored or poorly benchmarked in a variety of models.

We examined the predictive capability of various regression models of estimating an atmospheric abundance with synthetic spectra provided by the ABC dataset and actual JWST NIRSpec observations of WASP 39b. Particular attention was given to the contribution of preprocessing pipelines such as interpolation, normalization, feature augmentation, and log-scaling and the effect it had on the model stability and accuracy. The comparisons on benchmarks were carried out among three representative algorithms, which were MLP, SVR, and XGBoost. The trained models were then used to apply JWST spectra that had undergone processing to test their ability to generalize into the real world.

Contributions of this paper include:

- Benchmarking of five preprocessing approaches on three ML regressors under well-controlled experimental settings.
- Revealing sample-wise normalization to be the key parameter affecting the stability of the models, an issue not widely examined in previous retrieval works.
- Verification of XGBoost forecasts against those available in the literature for the JWST retrieval of WASP 39b.

# 2 Results

## 2.1 Impact of Preprocessing on Model Stability

A key result of this study is the strong dependence of model performance on preprocessing. Five Feature Scaling strategies were tested under identical training conditions for comparative study of temperature.

**Findings for temperature:**

- Models trained on Z-score Normalization ($N$) and Z-score Normalization with Feature Augmentation ($NMS$) produce stable high-accuracy models, with MLP achieving $R^2 \approx 0.99$.
- Models trained on Max Normalization ($NM$) and Max Normalization with Feature Augmentation ($NMM$) degrade performance substantially ($R^2$ between 0.6–0.9).
- Models trained on Standardisation ($S$) caused complete model collapse for all models, yielding near-zero or negative $R^2$, confirming that global scaling disrupts high dynamic range spectral features.

**Findings for gases:** M1 (Feature scaled on $N$) and M2 (Feature scaled on $NMS$) show the strongest and most stable performance across all models.

## 2.2 Model Performance Comparison

Table 1 depicts the predicted log abundances of the five atmospheric gases ($H_2O$, $CO_2$, CO, $CH_4$, and $NH_3$) for WASP 39b across three regression models trained on the $N$ and $NMS$ preprocessed spectra, the two configurations identified above as producing the most stable predictions. Table 2 extends this comparison to the planetary equilibrium temperature further evaluating all five feature scaling techniques ($N$, $NM$, $NMS$, $NMM$, and $S$) across the three models. Together, the obtained results illustrate the sensitivity of both gas abundance and temperature predictions to the choice

| Model | log_$H_2O$ | log_$CO_2$ | log_CO | log_$CH_4$ | log_$NH_3$ |
|---|---|---|---|---|---|
| XGB ($N$) | -5.99 | -6.35 | -4.76 | -3.99 | -6.46 |
| XGB ($NMS$) | -6.25 | -5.38 | -3.57 | -3.66 | -5.37 |
| SVR ($N$) | -1.22 | -6.85 | -4.73 | -3.92 | -3.92 |
| SVR ($NMS$) | -0.02 | -6.76 | -4.42 | -3.72 | -3.00 |
| MLP ($N$) | -4.96 | -4.62 | -2.37 | -0.84 | -1.57 |
| MLP ($NMS$) | 5.36 [1] | -4.57 | -3.35 | -2.17 | -4.06 |

**Table 1** Prediction of Atmospheric Composition Concentrations Across Models on the Preprocessed Spectra of WASP 39b

of preprocessing pipeline, with XGBoost showing the most consistent performance across configurations.

| Feature Scaling Technique | XGB | SVR | MLP |
|---|---|---|---|
| $N$ | 605.38 | 102.91 | -0.99 |
| $NMS$ | 580.21 | 604.92 | -0.99 |
| $NM$ | 627.30 | 1965.44 | 3.77 |
| $NMM$ | 958.40 | 824.84 | -0.99 |
| $S$ | 1331.42 | 939.89 | $1.97\times10^{51}$ |

**Table 2** Comparison of Feature Scaling Techniques Across Models for Predicting the Planetary Temperature from the Preprocessed Spectra of WASP 39b

## 2.3 Discussion of Model Behavior

Across all experiments, several consistent patterns emerge:

- **MLP** consistently outperforms other architectures once an appropriate normalization is applied. Although the performance metrics were promising, the model fails to generalize on the observational dataset.
- **SVR** provides robust performance with minimal preprocessing, though computational cost is high for large samples. It generalized accurately on temperature; however, it cannot be deemed credible as the performance metrics were not promising. Similar behavior to MLP was observed for gases.
- **XGB**'s tree-based structure, when trained with appropriate normalization, produced promising performance metrics and also generalized accurately on the observational dataset.
- **Preprocessing dominates model behavior:** versions trained with standard scaling universally collapsed [10], confirming that the spectral shape is the primary source of abundance information. Versions trained with feature augmentation tend to produce better generalizations.

[1] This value falls outside the physically plausible range of for a volume mixing ratio which is consistent with MLP's documented failure to generalize on observational data

These findings emphasize that model comparisons must be made after controlling for preprocessing effects [10]; otherwise, architectural differences become overshadowed by data representation artifacts.

### 2.4 Model Selection

| $X_i$ \Model | XGB ($N$) | Retrieval | Reference |
|---|---|---|---|
| log $X_{H_2O}$ | $-5.99$ | $-4.85 \pm 0.38$ | Constantinou et al. [11] |
| log $X_{CO_2}$ | $-6.35$ | $-6.59$ to $-4.16$ | Constantinou et al. [11] |
| log $X_{CH_4}$ | $-4.76$ | $< -5.3$ | Ahrer et al. [2] |
| log $X_{CO}$ | $-3.99$ | $-4.25$ to $-2.58$ | Constantinou et al. [11] |
| log $X_{NH_3}$ | $-6.46$ | $< -6$ | Alderson et al. [3] |

**Table 3** Most Accurate Predictions of Atmospheric Composition from XGB trained on data preprocessed using $N$

The most accurate planetary equilibrium temperature of WASP 39b was estimated by XGBoost which was trained on the data scaled using $NMM$, yielding a temperature prediction of 958.40K, compared to the actual equilibrium temperature of 1100K [12].This is an absolute error of 141.6 K (12.9 %) from the true value due to the skewed distribution of temperature in the ABC data set, where temperatures beyond 2000 K are not well-represented.

## 3 Discussion

This study provides a structured evaluation of machine-learning regressors for exoplanet atmospheric retrieval and demonstrates that preprocessing choices exert a dominant influence on model performance. Across all experiments, the most stable configuration combined sample wise normalization of spectral features with logarithmic scaling of gaseous abundances and temperature. This approach preserved the relative spectral shape of critical features for molecular identification while mitigating the effects of skewness in the underlying distributions.

The evaluation also underscores several broader considerations for machine learning retrieval. First, feature scaling had to be tuned to the unique properties of exoplanet spectra , global standardization compresses the dynamic range of absorption features and can may lead to catastrophic performance collapse of the models. Second, synthetic datasets such as the ABC archive, do not yet capture the full complexity of exoplanet atmospheres, including clouds, hazes[13–15] and non-equilibrium chemical process commonly observed in real exoplanet atmospheres. This limits models ability to generalize for species such as $CH_4$, which exhibits more complex and weakly represented spectral signatures.Third,lack of quantified uncertainties hinders comparisons to Bayesian approaches that offer posterior distributions. Fourth, the overall predictions have been performed taking a fixed single seed (42). These are the essential aspects that should be considered in future research regarding retrieval via machine learning.

While the models achieve high $R^2$ scores on the synthetic test set, indicating strong capture of variance in the target distributions, the absolute error metrics (MAE, RMSE) reflect the inherent difficulty of predicting abundances across a wide dynamic range of approximately 6 dex. For abundances in log scale, even a perfectly tuned model will have significant absolute errors at the edges of the distribution, especially for less abundant species like $CH_4$ and $NH_3$, due to their weaker spectroscopic signatures. This demonstrates that $R^2$ as a sole metric is not a good performance measure of retrieval.

These findings collectively highlight the potential of machine learning regressors as optimal tools for atmospheric characterization while emphasizing the importance of carefully designed pipelines based on data preprocessing, dataset construction, and evaluation strategies. As high quality spectra continue to emerge from JWST and future missions, incorporating noise aware training, physics informed normalization, symmetry aware neural architectures, and comprehensive uncertainty estimation will be essential for transitioning ML based retrieval from proof-of-concept to standard practice in exoplanet science.

# 4 Methods

This section outlines the machine learning framework developed for atmospheric abundance estimation from exoplanet transmission spectra. The methodology spans database description, preprocessing rationale, architecture design, optimization strategy, and evaluation criteria. Each methodological choice is motivated both from a computational perspective and from the physical properties of exoplanetary spectra. All algorithms will be implemented using scikit-learn machine learning library, however the XGB model will be implemented using the XGBoost machine learning library. The supervised regression algorithms implemented and tested include XGB,SVR,and MLP.

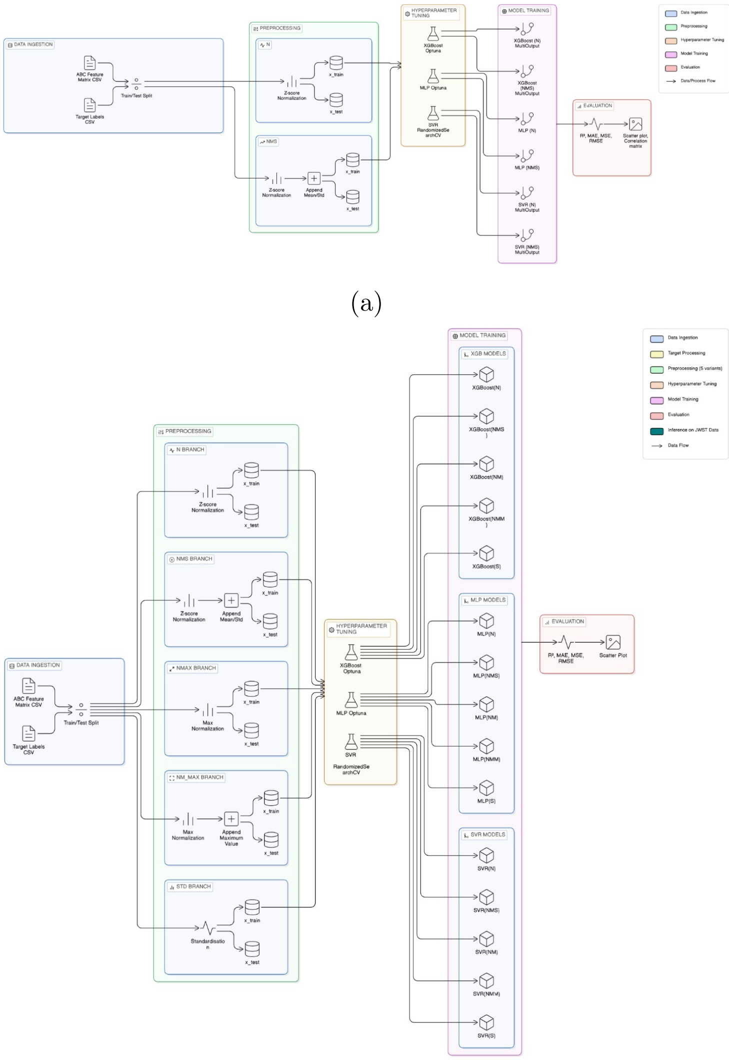
## 4.1 Database Description

subsec4

This study utilizes a combination of synthetic atmospheric spectra and real JWST observational data. Each dataset plays a specific role in the machine-learning pipeline, including model training, distribution verification, preprocessing analysis, and real-world evaluation. This section provides a complete description of all datasets used and their respective contents.

### 4.1.1 Synthetic Dataset: ABC Database

The synthetic spectra used in this work are derived from the ABC database [16], originally released for the 2022 Ariel Machine Learning Data Challenge . The ABC archive contains forward-modelled transmission spectra generated using the TauREx radiative-transfer code, assuming hydrogen–helium dominated, isothermal atmospheres and including line absorption, Rayleigh scattering, and CIA processes.



(a)

(b)

**Fig. 1** System Architecture for (a) Prediction of Atmospheric Gases (b) Prediction of Temperature .

Each simulation corresponds to a unique planet–star system, with stellar and planetary properties drawn from the Ariel target catalog. This catalog consisted of 5900 actual planets.

Each planet retained its stellar properties and planetary parameters, e.g., radius, mass, distance, etc., whereas the planet's atmospheric composition was generated randomly and the equilibrium temperature of the planet was chosen to produce the spectral data.

The same forward-modelled parameters and spectral data were used as originally in the ABC dataset without taking instrument-specific noise into consideration.

This database includes trace amounts of five different atmospheric gases ($H_2O$, $CH_4$, $CO_2$, CO, $NH_3$) represented as log-concentration, with the volume mixing ratio being less than $10^{-3}$ [16]. The other parameters included in the dataset are star–Earth distance, star temperature, star radius, star mass, star–planet distance, planet radius, planet mass, planet surface gravity, and the planet orbital period.

#### 4.1.2 Observational Dataset: WASP 39b

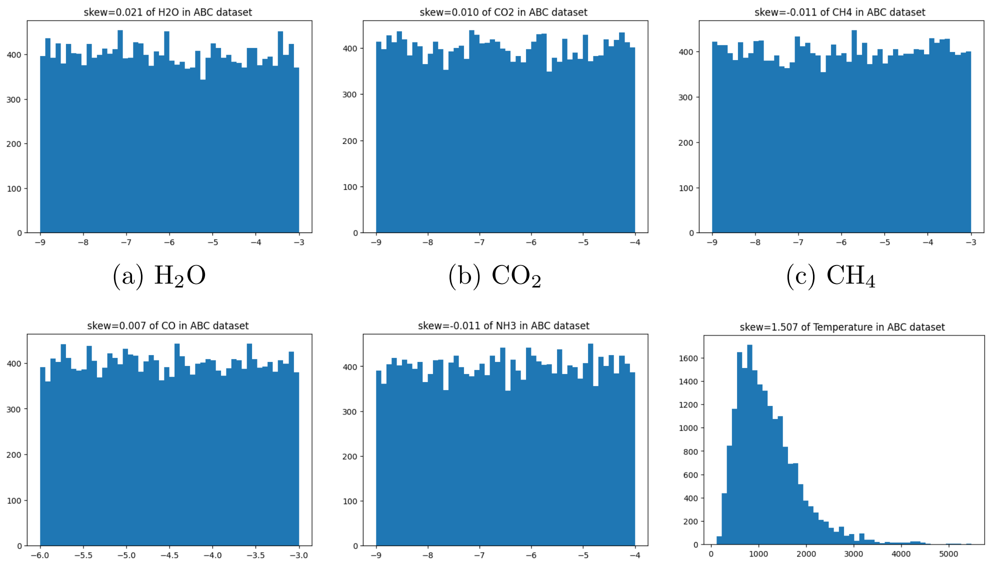
This dataset includes the spectral data of WASP 39b from the JWST observed using the NIRSpec/PRISM instrument [2–4]. This contained spectral range from 0.6 nm to 5.3 nm which was extracted from the `EXTRACT_1D` FITS files. This includes readings of wavelength, flux, flux-error, noise, background signal surface brightness, and Modified Julian Date at different time periods of the observation. We used the same wavelength and the flux observed by JWST.

### 4.2 Data Preprocessing

#### 4.2.1 Dataset Validation and Structure Examination

Each planet contains a transmission spectrum, wavelength grid, and associated gas abundances.

Spectral arrays are uniform in length, but the temperature parameter was found to be highly skewed (Figure 2), unlike for gas abundances which was logarithmically scaled in the ABC dataset, therefore the temperature parameter was scaled in a similar fashion.




(a) $H_2O$ (b) $CO_2$ (c) $CH_4$

(d) CO (e) $NH_3$ (f) Temperature

**Fig. 2** Skew in the target parameters where the x-axis represents the atmospheric properties and the y-axis represents the number of instances.

A random sample of 20,000 planets was extracted using a fixed seed for reproducibility. The representativeness of the sample was verified using methods such as Comparative Distribution Plot (Figure 3),Kolmogorov-Smirnov, Mean and Std comparison, Outlier comparison (values beyond 3 std), Earth Mover's Distance (Wasserstein distance). The results show that the sample was truly representative of the original data.

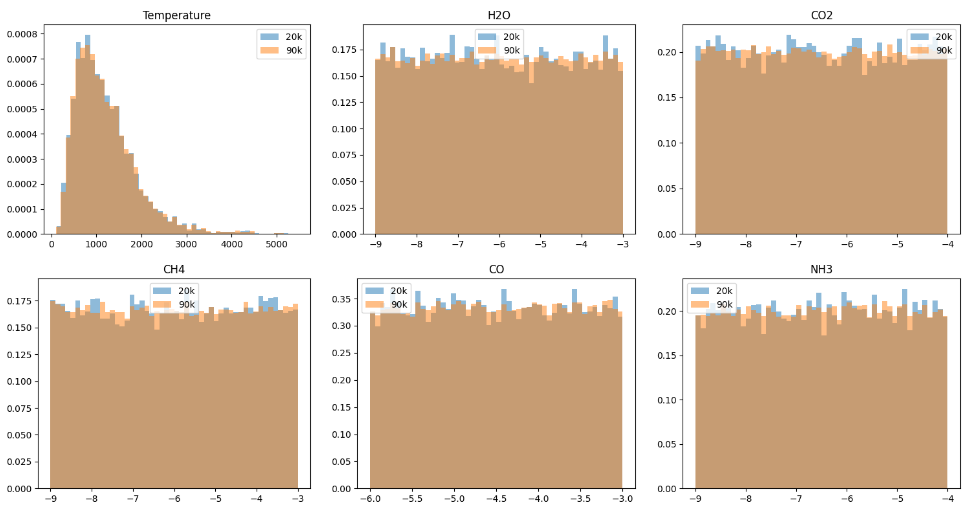



**Fig. 3** Comparison of target parameter distribution between the sample subset and the complete dataset where x-axis represents the atmospheric properties and y-axis represents probability density.

### 4.2.2 Wavelength Interpolation and Spectral Alignment

The JWST NIRSpec/PRISM observations of WASP 39b do not share the same wavelength grid as the ABC synthetic spectra. To ensure compatibility, all spectra were mapped onto a common 52-bin wavelength grid, the interpolation procedure involved using Cubic spline interpolation onto the JWST extracted wavelength grid to preserve local spectral (Figure 4). This step ensures that both synthetic spectra and the JWST case study can be given to the same ML model without architectural modifications.

### 4.2.3 Feature Scaling and Feature Augmentation

- Normalization
  - $N$ : This technique was used to rescale each sample by subtracting its mean from each feature and dividing it by its standard deviation.

$$x^{(j)}_{Ni} = N[x^{(j)}_i] = \frac{x^{(j)}_i - \bar{x}i}{\sigma xi} \quad (1)$$

  - $NM$ : This technique was used to rescale each sample by dividing all its features by the maximum absolute value within the sample.

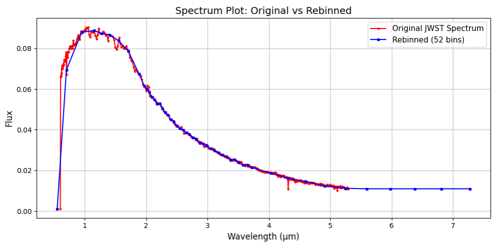



**Fig. 4** Originally observed JWST NIRSpec/PRISM spectrum of WASP 39b alongside the rebinned 52-bin version used as model input. Cubic spline interpolation was applied to align the observational wavelength grid with the synthetic ABC spectra, enabling both datasets to be processed by the same model without architectural modifications.

$$N_i^{(j)} = N\left[x_i^{(j)}\right] = \frac{x_i^{(j)}}{M^{(j)}} \tag{2}$$

– $NMS$ : $N$ was applied, and mean and standard deviation of each sample were augmented to the normalized data as additional features.
– $NMM$ : $NM$ was applied, and maximum absolute value of each sample was augmented to the normalized data as additional features.

Feature augmentation was performed in order to add information about the absolute intensity that would be lost during feature scaling [17].

- $S$ : This technique transforms each feature to have zero mean and unit variance by applying z-score scaling.

$$x_i^{(j)}\text{std} = S[x_i^{(j)}] = \frac{x_i^{(j)} - \bar{x}^{(j)}}{\sigma x^{(j)}} \tag{3}$$

Sample-wise normalization was used because of the promising results of this configuration [10].

On not having found any concrete machine learning induced predictions on temperature, the following comparative study was conducted. This study involved the study of five preprocessing techniques mentioned above were used.

#### 4.2.4 Transit Period Selection

Studying the variations in the relative flux as the planet transits the star, the spectral features were extracted during this transit (Figure 5)[4].

### 4.3 Model Architectures

Three models were selected to represent distinct regression paradigms commonly applied to spectral inference tasks:

#### 4.3.1 MLP

An MLP is a feedforward neural network composed of multiple layers of interconnected neurons with nonlinear activation functions. It learns complex function

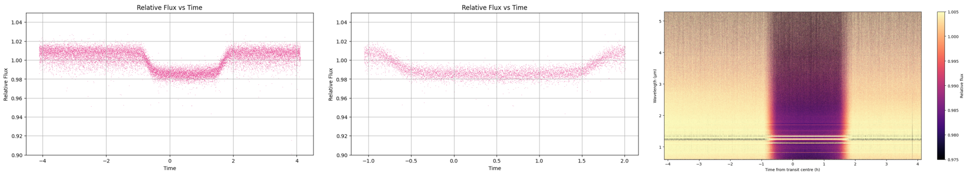



**Fig. 5** Selection of the transit period for WASP 39b using relative flux time series from JWST NIRSpec/PRISM observations. The in transit window used for spectral feature extraction is identified from the characteristic flux decrease during planetary transit

mappings through gradient-based optimization, making it suitable for modelling high-dimensional and nonlinear relationships. A fully connected neural network was implemented to model nonlinear relationships between absorption features and atmospheric composition. ReLU activation was adopted to preserve non-negative feature gradients, while dropout was incorporated to approximate Bayesian uncertainty in predictive inference [8]. This choice reflects the need for flexible function approximation given overlapping spectral signatures and pressure-dependent line shapes.

### 4.3.2 SVR

SVR is a kernel-based regression method that models nonlinear relationships by projecting data into a higher-dimensional feature space. It learns a function that fits the data within a tolerance margin while maximizing generalization through margin-based optimization. SVR with a radial basis function (RBF) kernel was included to benchmark non-neural, kernel-based methods capable of handling nonlinear spectral-abundance relationships, especially under conditions where spectra exhibit continuity but sharp local gradients due to isolated molecular features.

### 4.3.3 XGB

XGB is a gradient-boosted decision tree algorithm optimized for speed, regularization, and handling nonlinear feature interactions. It builds an ensemble of trees sequentially, where each tree corrects the errors of the previous ones to improve predictive accuracy. Gradient-boosted decision trees were selected owing to their success in handling structured nonlinear relationships and residual patterns that may emerge from imperfect normalization or interpolation. Prior spectroscopy ML studies report strong performance of tree-based methods where feature interactions are nontrivial. These three models together represent a spectrum from parametric to non-parametric inference and enable sensitivity analysis with respect to architectural assumptions .

## 4.4 Hyperparameter Optimization and Convergence Behaviour

For atmospheric composition prediction to ensure comparability across models, hyperparameters for the MLP and XGBoost models were tuned using Optuna's Bayesian optimization strategy . However, initial attempts to tune the SVR model with Optuna resulted in repeated timeouts due to non-convergence. As a practical alternative, a randomized search strategy was adopted for SVR to maintain feasible training time while still exploring the relevant parameter space.

Convergence behaviour differed considerably across models. The MLP exhibited gradual and stable improvements over successive trials, whereas the SVR showed high sensitivity to preprocessing and feature scaling. This behaviour is consistent with trends reported in prior atmospheric retrieval studies.

The hyperparameters of the best-performing configurations are reported below: XGBoost trained with $N$ feature scaled data for gaseous abundance prediction and $NMM$ feature scaled data for temperature estimation

| Hyperparameter | $N$ Feature Scaled Data |
|---|---|
| n_estimators | 1499 |
| learning_rate | 0.11269776375447682 |
| max_depth | 4 |
| min_child_weight | 14 |
| gamma | 0.00789553811012314 |
| subsample | 0.8071501278758849 |
| colsample_bytree | 0.9608473552741683 |
| reg_alpha | 6.1504575935622885 |
| reg_lambda | 2.6134050947658922 |
| random_state | 42 |

**Table 4** Hyperparameter configuration of tuned XGBoost regression model using $N$ feature scaled data for the estimation of Gaseous concentrations

| Hyperparameter | $NMM$ feature scaled data |
|---|---|
| n_estimators | 731 |
| learning_rate | 0.06405 |
| max_depth | 8 |
| min_child_weight | 3 |
| gamma | $8.09 \times 10^{-6}$ |
| subsample | 0.73580 |
| colsample_bytree | 0.76200 |
| reg_alpha | 1.12903 |
| reg_lambda | 4.26768 |
| objective | reg:squarederror |
| verbosity | 0 |
| n_jobs | -1 |

**Table 5** Hyperparameter configuration of the tuned XGBoost model using $NMM$ feature scaled data for estimation of Planet Temperature

## 4.5 Training and Validation Framework

The dataset was split into training and testing set in the 80-20 (16000 - 4000 instances) ratio. The training set was further split into training and validations sets in 80-20

(9600 - 2400 instances) ratio. Training of models was done up until convergence by early stopping through loss from the validation set. Various pre-scaling techniques of the features were applied in order to ascertain how they affected the consistency of the predictions. The setup allowed the researchers to establish whether differences in performance were brought about by pre-processing or the architecture itself, which was unexplored before.

### 4.6 Evaluation Metrics

The performance of the trained models was evaluated using standard error-based metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the Coefficient of Determination ($R^2$).

### 4.7 Graphical Analysis

To further assess model behavior, **scatter plots** (Figures 6,7,8) of predicted versus true values were employed. A tight clustering of points around the diagonal line $y = \hat{y}$ indicates good calibration and minimal systematic bias.

**Acknowledgements.** The authors gratefully acknowledge the support of the Department of Information Science and Engineering, BMS College of Engineering, Bengaluru, for providing laboratory facilities and computational resources used in this study. The authors also thank the faculty mentors for their guidance during the development and evaluation of machine learning models presented in this study.

## Declarations

### Author Contributions

Rashmi R - Survey and Literature Review
Shobha T - Survey and Literature Review
Hamshith A M - Worked on Development
Harimedha S C - Worked on Development
Gagandeep Y - Worked on Development
Dhruv H - Worked on Development

### Competing Interests

The authors declare no competing interests.

### Funding Details

No funding provided.

## Data Availability

The synthetic atmospheric spectra used in this study were derived from the publicly released ABC dataset, which is accessible through the official Zenodo repository at https://zenodo.org/records/6770103.

The observational dataset used for the WASP 39b case study is publicly available through the Mikulski Archive for Space Telescopes (MAST) https://mast.stsci.edu/portal/Mashup/Clients/Mast/Portal.html as part of the JWST ERS program.

## Code Availability

The source code for this study is available at: https://github.com/Hamshith/Exoplanet-Atmospheric-Retrieval. The repository includes all scripts, model configurations, and instructions necessary to reproduce the experiments.

## Ethics approval and consent to participate

Not Applicable

## Consent for publication

Not Applicable

# References


[1] Tsiaras, A., Waldmann, I.P., Zingales, T., et al.: A population study of gaseous exoplanets. The Astronomical Journal **155** (2018) https://doi.org/10.3847/1538-3881/aaaf75

[2] Ahrer, E.-M., Stevenson, K.B., Mansfield, M., *et al.*: Early release science of the exoplanet wasp-39b with jwst nircam. Nature **614**, 653–658 (2023) https://doi.org/10.1038/s41586-022-05590-4

[3] Alderson, L., Wakeford, H.R., Alam, M.K., *et al.*: Early release science of the exoplanet wasp-39b with jwst nirspec g395h. Nature **614**, 664–669 (2023) https://doi.org/10.1038/s41586-022-05591-3

[4] Rustamkulov, Z., Sing, D.K., Mukherjee, S., *et al.*: Early release science of the exoplanet wasp-39b with jwst nirspec prism. Nature **614**, 659–664 (2023) https://doi.org/10.1038/s41586-022-05677-y

[5] Nixon, M.C., Madhusudhan, N.: Assessment of supervised machine learning for atmospheric retrieval of exoplanets. Monthly Notices of the Royal Astronomical Society **496**, 269–281 (2020) https://doi.org/10.1093/mnras/staa1150

[6] Cobb, A.D., Himes, M.D., Soboczenski, F., *et al.*: An ensemble of bayesian neural networks for exoplanetary atmospheric retrieval. The Astronomical Journal **158**, 33 (2019) https://doi.org/10.3847/1538-3881/ab2390

[7] Márquez-Neila, P., Fisher, C., Sznitman, R., Heng, K.: Supervised machine learning for analysing spectra of exoplanetary atmospheres. Nature Astronomy **2**, 719–724 (2018) https://doi.org/10.1038/s41550-018-0504-2

[8] Waldmann, I.P.: Dreaming of atmospheres. The Astrophysical Journal **820**, 107 (2016) https://doi.org/10.3847/0004-637X/820/2/107

[9] Yip, K.H., Changeat, Q., Al-Refaie, A., Waldmann, I.P.: To sample or not to sample: Retrieving exoplanetary spectra with variational inference and normalizing flows. The Astronomical Journal **162**, 195 (2021) https://doi.org/10.3847/1538-4357/ad063f

[10] Forestano, R.T., Matchev, K.T., Matcheva, K., Unlu, E.B.: Supervised machine learning methods with uncertainty quantification for exoplanet atmospheric retrievals. The Astrophysical Journal (2025) 10.48550/arXiv.2508.04982. In review

[11] Constantinou, S., Madhusudhan, N., Gandhi, S.: Early insights for atmospheric retrievals of exoplanets using jwst transit spectroscopy. The Astrophysical Journal Letters **943**, 10 (2023) https://doi.org/10.3847/2041-8213/acaead

[12] Espinoza, N., Steinrueck, M.E., Kirk, J., *et al.*: Inhomogeneous terminators on the exoplanet wasp-39b. Nature **632**, 1017–1024 (2024) https://doi.org/10.1038/s41586-024-07768-4

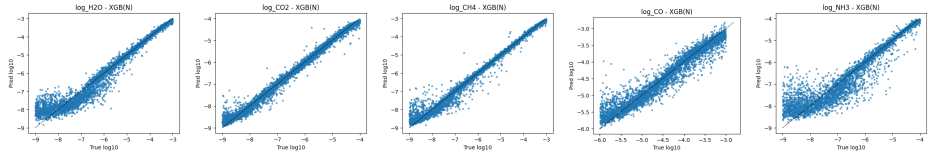
[13] Rukdee, S.: Instrumentation prospects for rocky exoplanet atmospheres studies with high resolution spectroscopy. Scientific Reports **14**, 27356 (2024) https://doi.org/10.1038/s41598-024-78071-5

[14] Callie E. Hood, M.R.L. Jonathan J. Fortney, *et al.*: Prospects for characterizing the haziest sub-neptune exoplanets with high-resolution spectroscopy. Astronomical Journal **160**, 198 (2020) https://doi.org/10.3847/1538-3881/abb46b

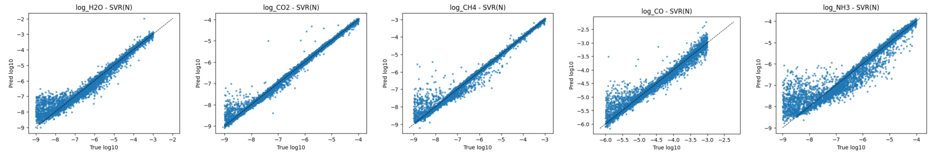
[15] Laura Kreidberg, J.-M.D. Jacob L. Bean, *et al.*: Clouds in the atmosphere of the super-earth exoplanet gj1214b. Nature **505**, 69–72 (2014) https://doi.org/10.1038/nature12888

[16] Changeat, Q., Yip, K.H., *et al.*: Esa-ariel data challenge neurips 2022: Inferring physical properties of exoplanets from next-generation telescopes. RAS Techniques and Instruments **2**, 45–61 (2023) https://doi.org/10.1093/rasti/rzad001

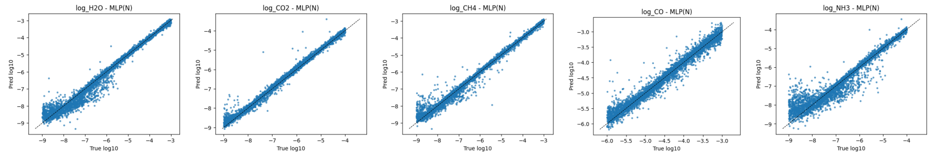
[17] Datta, R., Ahmed, F., Hare, J.D.: Machine learning assisted analysis of visible spectroscopy in pulsed-power-driven plasmas. Transactions on Plasma Science **52**, 4755–4763 (2024) https://doi.org/10.1109/TPS.2024.3364975




(a)

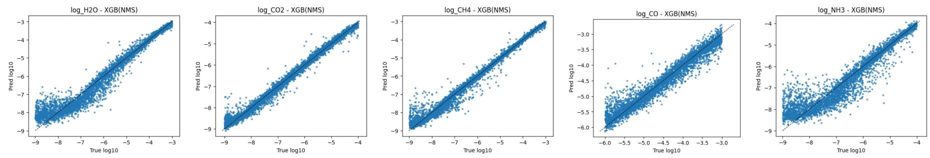



(b)

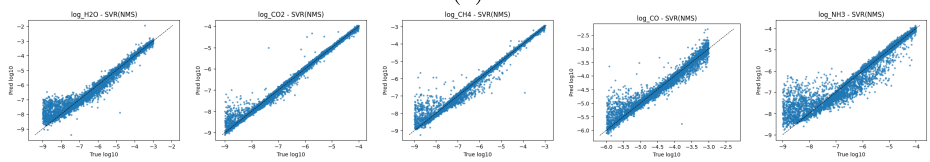



(c)

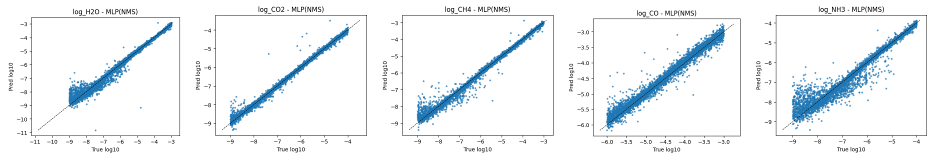
**Fig. 6** Scatter plots of true versus predicted log-abundance for five atmospheric gases ($H_2O$, $CO_2$, CO, $CH_4$, $NH_3$) using (a) XGB, (b) SVR, and (c) MLP models trained on $N$ preprocessed data. Diagonal alignment indicates accurate predictions with minimal systematic bias.




(a)




(b)




(c)

**Fig. 7** Scatter plots of true versus predicted log-abundance for five atmospheric gases using (a) XGB, (b) SVR, and (c) MLP models trained on $NMS$ preprocessed data. Comparison with 6 isolates the effect of feature augmentation on prediction accuracy.

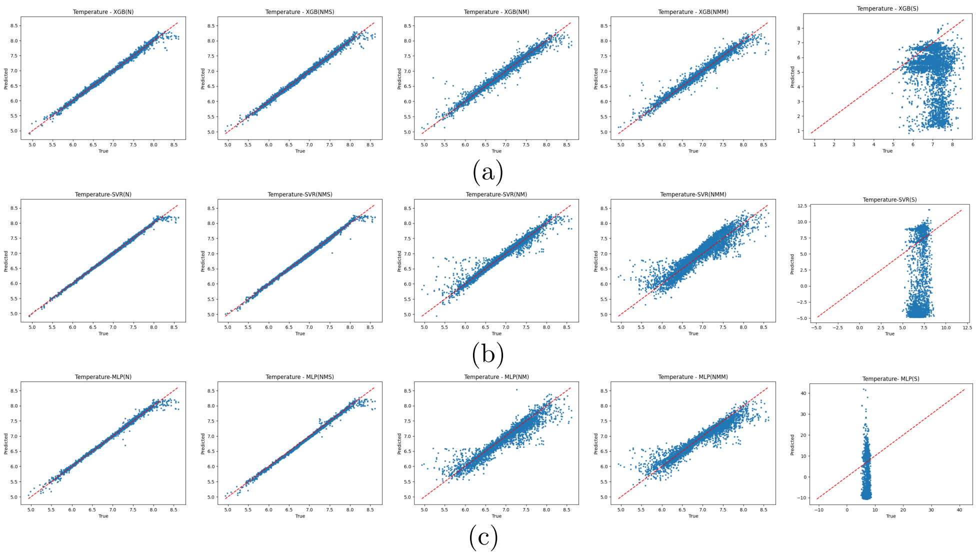



**Fig. 8** Scatter plots of true versus predicted planetary equilibrium temperature using (a) XGB, (b) SVR, and (c) MLP models across all five preprocessing strategies. The progressive degradation in scatter plot alignment from $N$ to $S$ preprocessing demonstrates the sensitivity of temperature prediction to feature scaling choice.

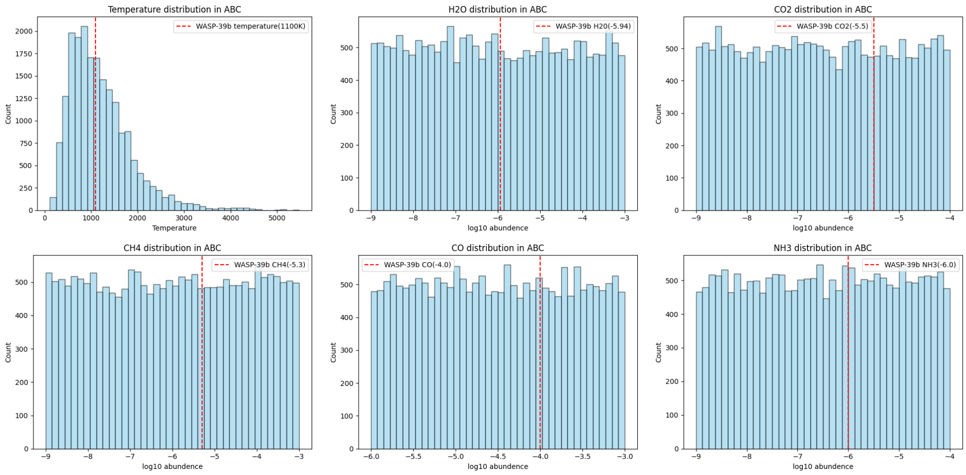



**Fig. 9** Distribution of target parameter values in the ABC training dataset compared against the known parameter values for WASP 39b (dashed red line). The overlap between training distribution and WASP 39b values confirms that the planet lies within the model's training domain for most species, supporting the validity of applying the trained models to this target.

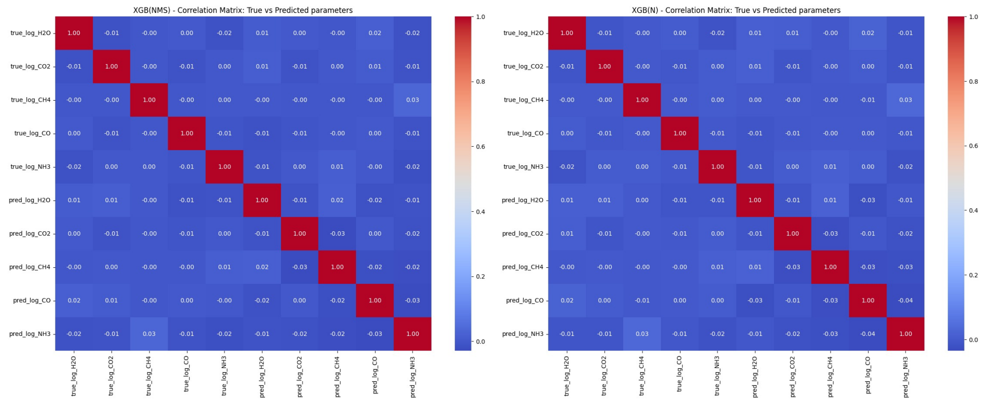



**Fig. 10** Correlation matrices comparing true and predicted log-abundance values for five atmospheric gases using the two best-performing model configurations, XGB($NMS$) and XGB($N$). Diagonal values of 1.0 reflect perfect self-correlation of true labels, while off-diagonal values near zero confirm that predictions for individual gases are retrieved independently without cross-contamination.