Life, Machine Learning, and the Search for Habitability: Predicting Biosignature Fluxes for the Habitable Worlds Observatory

Mark Moussa[1], Amber V. Young[1], Brianna Isola[1,3], Vasuda Trehan[1,4], Michael D. Himes[1,5], Nicholas Wogan[2], Giada Arney[1]

[1]NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA
[2]NASA Ames Research Center, Moffett Field, CA 94035, USA
[3]University of New Hampshire, Durham, NH 03824, USA
[4]University at Albany (SUNY), Albany, NY 12222, USA
[5]Morgan State University, Baltimore, MD 21251, USA

Abstract

Future direct-imaging flagship missions, such as NASA's Habitable Worlds Observatory (HWO), face critical decisions in prioritizing observations due to extremely stringent time and resource constraints. In this paper, we introduce two advanced machine-learning architectures tailored for predicting biosignature species fluxes from exoplanetary reflected-light spectra: a Bayesian Convolutional Neural Network (BCNN) and our novel model architecture, the Spectral Query Adaptive Transformer (SQuAT). The BCNN robustly quantifies both epistemic and aleatoric uncertainties, offering reliable predictions under diverse observational conditions, whereas SQuAT employs query-driven attention mechanisms to enhance interpretability by explicitly associating spectral features with specific biosignature species. We demonstrate that both models achieve comparably high predictive accuracy on an augmented dataset spanning a wide range of exoplanetary conditions, while highlighting their distinct advantages in uncertainty quantification and spectral interpretability. These capabilities position our methods as promising tools for accelerating target triage, optimizing observation schedules, and maximizing scientific return for upcoming flagship missions such as HWO.

1 Introduction

The field of exoplanet science is entering a new era of planetary characterization, building on the legacy of *Kepler*, which showed that small, sub-Neptune-size ($R_p \lesssim 4\,R_\oplus$) planets are common (Borucki et al. 2010). This suggests that Earth-sized, potentially habitable worlds may be frequent, making their detection and study a central priority for future missions. NASA's **Habitable Worlds Observatory** (HWO) aims to directly image and spectrally characterize many potentially habitable exoplanets in reflected light across the ultraviolet (UV), visible (VIS), and near-infrared (NIR). Because coronagraph observations capture only selected bandpasses and telescope time is scarce, efficient triage and prioritization are essential: a precursor study projected up to two years of observations merely to identify high-priority detailed observation targets (Roberge and Moustakas 2018). With the multi-billion-dollar costs of a flagship like HWO and limited observing time, avoiding even a single 10-hour observation that would not yield useful information represents a near million-dollar cost saving. Advancing these capabilities now will directly inform mission planning and architecture, positioning artificial intelligence (AI) and machine learning (ML) as an operational capability from the outset, initially in ground-based planning and triage, and with a potential path to selective onboard use as flight computing and verification mature.

Direct-imaging pipelines must translate partial spectra into physically meaningful inferences about atmospheres and surfaces ("retrieval"). Classical atmospheric retrieval analysis techniques have emphasized chemical species abundances and surface properties (Madhusudhan 2018), yet abiotic photochemical and geological processes can produce biosignature species leading to false-positive interpretations (e.g., volcanic and hydrothermal activity producing abiotic CH_4). In contrast, biosignature *fluxes*, or the rates at which biosignatures (e.g., O_2 and CH_4) enter the environment, tie observations to underlying sources and sinks (i.e., processes that add or remove gases from a planet's atmosphere, respectively) (Schwieterman et al. 2018). Fluxes that exceed plausible abiotic production offer stronger evidence for life than abundances alone, but retrieving fluxes with traditional approaches is computationally prohibitive: 10^4–10^6 forward-model evaluations combined with minutes-per-simulation photochemistry render end-to-end flux retrieval impractical. Therefore, there is a practical need to accurately infer biosignature fluxes from physically realistic reflected-light spectra in a computationally efficient manner.

Robust uncertainty quantification (UQ) is central to decision-making under low signal-to-noise ratio (SNR), incomplete spectral coverage, and model misspecification. Calibrated posteriors that capture epistemic and aleatoric uncertainty give planners a quantitative basis to balance observing time against expected scientific return, allocating resources only when the probability that a candidate's gas flux exceeds abiotic limits surpasses a risk-adjusted threshold.

We address these challenges with two complementary models: **(1) a Bayesian convolutional neural network (BCNN)**, which places UQ at the core of prediction, and **(2) our novel query-based Transformer, the Spectral Query Adaptive Transformer (SQuAT)**, which employs a com-

pact set of learnable, biosignature-specific query tokens that cross-attend to the wavelength-encoded spectrum using absorption-region priors from the literature and incorporates physics-guided attention priors. Deterministic convolutional neural network (CNN) and vision transformer (ViT) baselines serve as references, and their Bayesian variants propagate epistemic uncertainty via weight distributions and Monte Carlo (MC) dropout. **SQuAT's query-driven attention yields interpretable heatmaps and ViT-level accuracy, while the BCNN delivers competitive accuracy with strong uncertainty calibration.** Predictions from both models offer credible intervals for analyzing exoplanets, ranking targets for biosignature detection probability and allocating scarce observation time.

To rigorously train and evaluate these methods, we expand the Frontier Development Lab's (FDL) PyAtmos corpus (Chopra et al. 2023) with non-modern-Earth-like atmospheric grids and pair each self-consistent climate-photochemistry simulation with a reflected-light spectrum generated by the Planetary Spectrum Generator (PSG; Villanueva et al. 2018). The resulting dataset spans a wider range of stellar types and surface–atmosphere states.

2 Related Works

We find prior literature establishes **(i)** ML-based retrieval feasibility, **(ii)** adequate methods for robust uncertainty quantification and interpretability, and **(iii)** query-specific attention in other domains enabling structured, disentangled outputs (Madhusudhan 2018; Soboczenski et al. 2018; Yip et al. 2020; Carion et al. 2020). **However, none integrate all three for biosignature *flux* regression, motivating our unified SQuAT framework.**

Bayesian deep learning methods, including ensembles and variational or dropout-based approximations, have improved uncertainty calibration and parameter correlations in retrieval tasks (Gal and Ghahramani 2016; Soboczenski et al. 2018; Cobb et al. 2019). Studies show that embedding domain structure can enhance both accuracy and interpretability, but focus on static abundance or profile inference, not biosignature flux prediction.

For direct imaging, ML models have enabled rapid estimation of bulk atmospheric properties and triage of terrestrial analogs using albedo or photometric data (Johnsen and Marley 2019; Pham and Kaltenegger 2022). These approaches focus on classification or coarse regression, not multi-output flux prediction.

Interpretability methods reveal spectral models often attend to molecular absorption bands (Yip et al. 2020; Fisher et al. 2019). This supports architectures with built-in spectral traceability, such as attention mechanisms aligned with species-specific diagnostics.

Transformers use global self-attention for efficient sequence modeling (Vaswani et al. 2017). Query-based decoders, such as Detection Transformers (DETR), localize structured outputs via fixed semantic tokens (Carion et al. 2020). Such designs remain unexplored in exoplanet retrieval, where prior work used shared output heads without task-specific queries.

3 Dataset

We build on **PyATMOS**, an open grid of ∼125,000 1D coupled photochemistry–climate simulations for an Earth analog around the Sun (Chopra et al. 2023). Atmospheres with surface temperatures exceeding ∼320 K were excluded to conservatively avoid regimes near the inner edge of the habitable zone limit, where runaway greenhouse conditions are expected (Kopparapu et al. 2014). As a result, we retain N_{atm} = 77,882 conservative habitable states.

We extended the dataset to include Proterozoic Earth-like scenarios that capture a broader atmospheric diversity of biologically active planets from an earlier period in Earth's inhabited history. We vary O_2, CH_4, and H_2O on a logarithmic grid from 10^{-15} to 10^{-2}, producing ∼1,500 self-consistent equilibrated cases after removing unphysical mixtures (total gas fraction > 1). The stellar spectrum is scaled for reduced luminosity following Claire et al. (2012). Additionally, planetary parameters (gravity, radius, orbital distance) are kept fixed at Earth-like values. The surface pressure is 1 bar, CO_2 and CO are fixed at 10^{-2} and 10^{-4}, and N_2 serves as the main background gas.

For non-habitable baselines, we systematically vary the effective stellar flux received by modern Earth by reducing incident stellar flux from 1.0 to 0.3 in steps of 0.02, yielding 36 simulations with subfreezing surfaces.

Each simulation provides altitude-dependent temperature, pressure, and abundances, plus the *net lower-boundary flux* ($z = 0$) for eight biosignature-relevant gases: O_2, O_3, CH_4, N_2O, CO_2, H_2O, CO, SO_2. These signed fluxes (source >0, sink <0) are the predictive targets.

We generate reflected-light spectra with PSG using each case's temperature–pressure and vertical abundance profiles. Spectra span 0.2–2.5 μm (355 points within the UV-VIS-NIR wavelength regions) at resolving power $R = 140$, reported as planet–star contrast for an Earth-twin star–orbit configuration. Instrumental noise is added downstream via SNR augmentation (5–100). These spectra serve as inputs to the learning pipeline.

We use 59,202/19,735/19,735 samples for train/val/test and hold out 1,033 TRAPPIST-1e samples, generated with the system's spectral energy distribution and planetary parameters, for out-of-distribution evaluation.

We apply z-score normalization independently per wavelength using training statistics. Targets y undergo a per-species asinh transform, $y' = \operatorname{asinh}(y/\beta)$, with β set to the training 90th percentile of $|y|$ for that species, stabilizing heavy-tailed flux distributions.

We intend to open source this augmented dataset in the near future.

4 Model Architectures

In this section, we explain the different model architectures experimented with in our work. Our models take as input raw reflected–light spectra, represented as a single-channel sequence of 355 wavelength points, and the outputs are eight biosignature fluxes, described in detail in Section 3.

For all models, we report point-estimation metrics including the coefficient of determination (R^2), root-mean-squared

error (RMSE), and mean absolute error (MAE); mean squared error (MSE) is used as the loss. For Bayesian or MC-dropout models, we additionally assess predictive uncertainty using coverage-based calibration (fractions within $\pm 1\sigma$ and $\pm 2\sigma$ of the predictive mean), sharpness (mean predictive standard deviation), the mean coefficient of variation (std/|mean|), and the correlation between predicted uncertainty and absolute error. The final checkpoint is selected via early stopping on validation MSE with a ReduceLROnPlateau scheduler. Model hyperparameters are optimized using a Bayesian algorithm to minimize validation loss, where models presented herein use the best-performing configuration. Development is conducted on a Linux system equipped with a single NVIDIA Tesla V100 GPU (32 GB VRAM) and Intel Xeon Platinum 8270 CPU (2.70GHz). Code was developed using Python 3.11 and PyTorch 2.5, with relevant packages detailed within the repository, which will be open-sourced. Global random seed is set to 42 for reproducibility.

4.1 Convolutional and Bayesian Convolutional Neural Networks

Our baseline is a lightweight 1D CNN. The backbone consists of five Conv1D–ReLU–MaxPool blocks (filters: 32, 64, 128, 256, 512; kernel sizes: 13, 11, 9, 7, 5), followed by two fully connected layers (256 and 128 units, dropout $p_D = 0.5$) and a final linear output layer. We train for $\sim$130 epochs using MSE loss with Adam ($\eta = 10^{-5}$, batch size 128, dropout $p_D = 0.5$). The CNN serves as a lower-bound reference for model performance.

The BCNN retains this architecture but replaces all convolutional and linear layers with Bayesian counterparts that maintain Gaussian weight posteriors (Blundell et al. 2015), and uses a heteroscedastic output head to jointly predict the mean and variance for each biosignature (Kendall and Gal 2017). Training runs for $\sim$140 epochs and minimizes an MSE term plus a Kullback–Leibler (KL) regularizer on the weight posteriors:

$$\mathcal{L}_{\text{train}} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{2}\|y_n - \hat{y}_n\|_2^2 \;+\; \beta\,\mathrm{KL}\big(q_{\boldsymbol{\theta}}(\mathbf{W})\,\|\,p(\mathbf{W})\big), \tag{1}$$

where β scales the regularization strength. At inference, $T = 50$ stochastic forward passes are used to estimate predictive means and decompose total uncertainty into epistemic and aleatoric components.

4.2 Spectral Query Adaptive Transformer

The SQuAT model extends a ViT backbone (Dosovitskiy et al. 2020) to better capture species-specific features in 1D exoplanet spectra. Our ViT configuration uses an embedding dimension $D = 256$, $L_{\text{enc}} = 6$ Transformer layers, $h = 8$ attention heads, MLP ratio $= 4$, and dropout $p_D = 0.2$. The spectrum is divided into overlapping patches (stride $= 1$), embedded with positional encodings, and processed by this stack of multi-head self-attention and feed-forward blocks.

SQuAT augments this design in three key ways. First, it employs a *multi-scale patch encoder* with three parallel convolutional branches (patch sizes 3, 5, and 10) to capture both narrow molecular absorption lines and broader spectral structures. The resulting features are concatenated and projected to a $D = 256$-dimensional embedding before entering a Transformer encoder ($L_{\text{enc}} = 6$ layers, $h = 8$ heads, dropout $p_D = 0.2$). Second, instead of a single global output token, SQuAT introduces a set of $K = 8$ learnable *biosignature queries*, one per target species. These queries attend to the encoded spectrum via multi-head cross-attention, producing per-species embeddings. This use of a fixed set of biosignature queries is conceptually related to DETR's "object queries" (Carion et al. 2020), but SQuAT performs fixed-K per-species flux regression on 1D spectra with physics-guided attention priors and omits Hungarian matching and detection losses. To inject spectroscopic domain knowledge, each attention map is softly biased toward known absorption bands for that species using

$$\mathbf{A}' = (1-\alpha)\,\mathbf{A} + \alpha\,\mathbf{P} \tag{2}$$

where $\mathbf{A}$ are the learned attention weights, $\mathbf{P}$ is a biosignature-specific prior mask, and $\alpha \in [0, 1]$ is a learnable mixing coefficient controlling the strength of prior injection during training. Finally, a *species interaction module* applies self-attention between the species embeddings to capture interdependencies, followed by a lightweight MLP head that produces a scalar flux prediction for each species. A diagram of this architecture is presented in Figure 1.

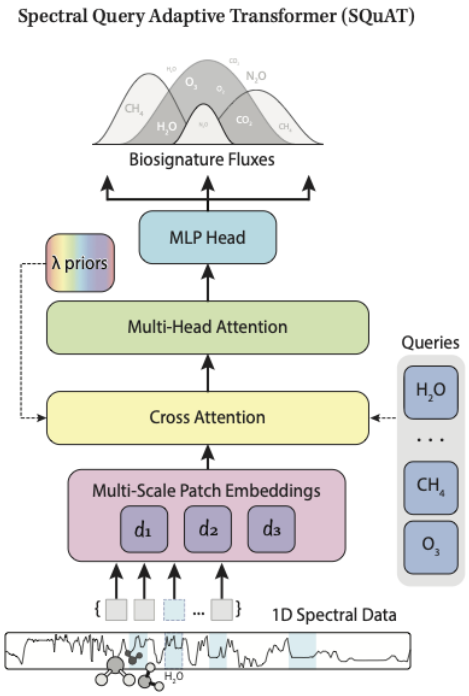

Figure 1: SQuAT model architecture.

Predictive uncertainty is estimated using MC dropout (Gal and Ghahramani 2016): dropout remains active at inference and $T = 30$ stochastic forward passes yield the predictive mean and epistemic variance. SQuAT is trained for $\sim$35 epochs, while the ViT baseline is trained for $\sim$50 epochs, and both use Adam ($\eta = 10^{-4}$, batch size 64).

5 Results

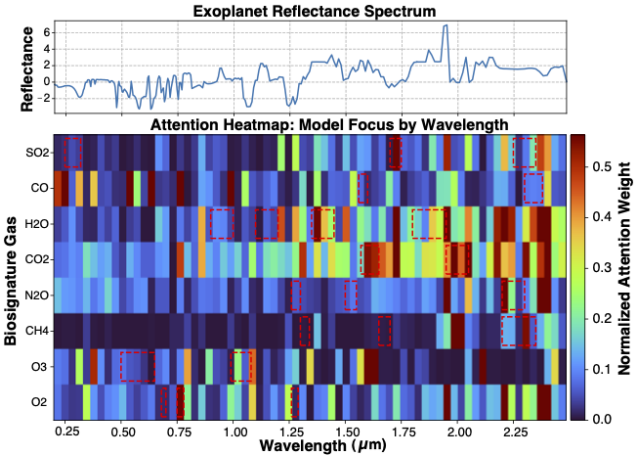

Figure 2: SQuAT attention map. Top: reflected-light exoplanet spectrum sample. Bottom: per-species attention across wavelength, highlighting prominent regions for each biosignature's flux prediction. Red dashed boxes mark canonical absorption bands. Attention is normalized per species.

Figure 2 visualizes the attention weights of one exoplanet sample learned by the SQuAT model across the input wavelength range. Each row corresponds to a target biosignature, with color intensity reflecting the relative importance assigned to each spectral region. Normalizing attention weights independently per biosignature enables comparison of intra-species focus without being dominated by stronger species. Attention weight mapping enhances diagnostic analysis and transparency by revealing which parts of the spectrum the model attends to when making predictions.

Figure 3 describes the Pearson correlation coefficients between prediction errors across all modeled species. Each matrix entry quantifies the degree to which errors in one species' predicted flux are linearly associated with errors in another. Notably, errors in predicting O_2 and CO_2 are strongly anti-correlated ($r = -0.79$), suggesting shared spectral or model confusion. This analysis reveals potential dependencies or disentanglements in the model's learned representations across atmospheric species.

Figure 4a presents scatter plots of predicted versus true biosignature flux values for four selected species—O_2, O_3, CH_4, and H_2O—as well as for all species aggregated. The top row depicts results from the BCNN, while the bottom row shows results from the SQuAT. For each model, we compute R^2 as a measure of fit quality. Across all targets, SQuAT achieves higher R^2 scores, indicating stronger agreement with ground truth. The dotted diagonal line represents perfect prediction.

Figure 4b displays predicted flux values with associated 95% credible intervals for each of four molecular species: O_2, O_3, CH_4, and H_2O. Each subplot shows the true values sorted in ascending order on the x-axis, allowing for clearer visualization of prediction fidelity and uncertainty

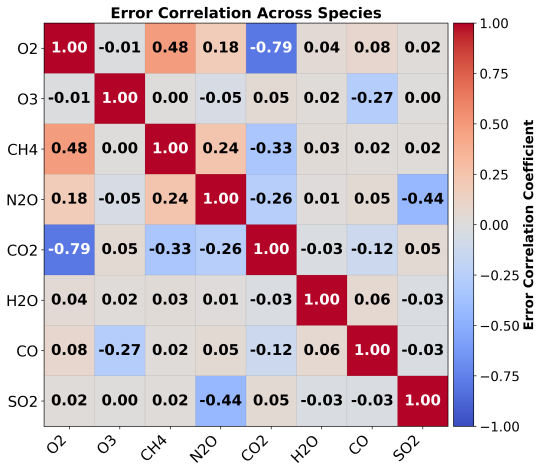

Figure 3: Error correlation matrix across predicted biosignature species. Each cell shows the Pearson correlation coefficient between the model prediction errors for a pair of species. High positive values (red) indicate that errors tend to co-occur in the same direction, while strong negative values (blue) suggest opposing error trends.

	CNN		BCNN		ViT		SQuAT	
SNR	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
5	0.155	0.917	0.812	0.432	0.827	0.415	0.747	0.502
10	0.806	0.439	0.933	0.258	0.939	0.247	0.924	0.274
20	0.939	0.246	0.962	0.195	0.968	0.179	0.970	0.174
40	0.966	0.184	0.970	0.174	0.982	0.136	0.979	0.144
50	0.969	0.175	0.971	0.171	0.983	0.131	0.981	0.139
100	0.973	0.164	0.972	0.167	0.985	0.122	0.982	0.134

Table 1: R^2 and RMSE for models at selected target SNRs.

across the value range. The BCNN (top row) and the SQuAT (bottom row) both achieve high R^2 scores, indicating accurate performance. Notably, both models capture structured uncertainty that varies with flux magnitude, and visually track epistemic variability, particularly for more challenging species like O_3. Credible intervals are narrowest for well-predicted regions and broaden with greater flux variability.

Table 1 reports R^2 and RMSE for CNN, BCNN, ViT, and SQuAT across target SNRs (5–100). Scores improve with SNR for all models. For SNR $\geq$ 20, ViT and SQuAT differ by at most $\Delta R^2 \leq 0.003$ and ΔRMSE ≤ 0.013 (e.g., at SNR$=$ 20: $R^2 = 0.970$ vs. 0.968, RMSE$=$ 0.174 vs. 0.179), whereas at SNR$=$ 5 the spread is larger (e.g., ViT $0.827/0.415$ vs. SQuAT $0.747/0.502$; CNN $0.155/0.917$).

6 Discussion

To assess whether the model has learned scientifically meaningful associations between spectral features and biosignature species, we visualize the learned attention weights across the input spectrum for a single representative sample (Figure 2). Encouragingly, the model attends strongly to wavelength regions that correspond to known molecular absorption features for most species. For instance, H_2O atten-

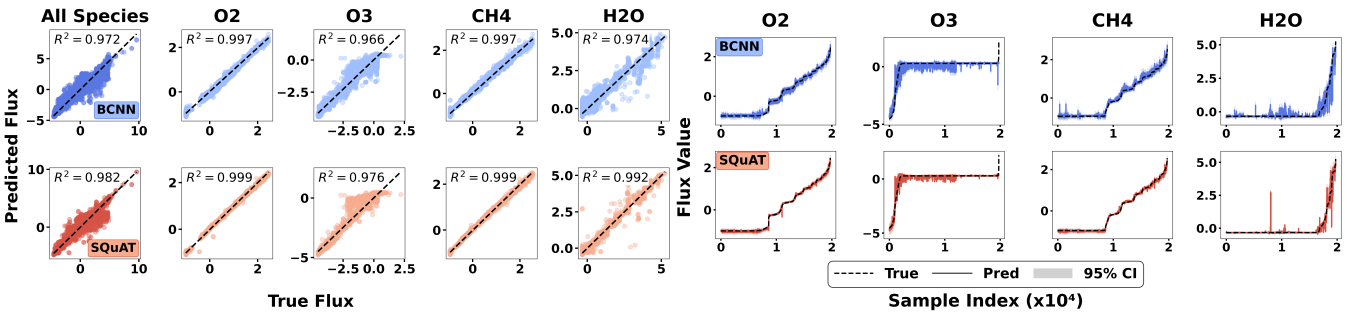

Figure 4: (a) Predicted versus true biosignature flux for four molecular species (O_2, O_3, CH_4, H_2O) and for all species combined. (b) Corresponding model predictions with 95% credible intervals, with samples sorted by increasing true flux.

tion peaks align with established bands at 1.1 µm, 1.4 µm, and ∼1.9 µm (Rothman et al. 1998). CH_4 exhibits clear attention near its strong absorption window at 2.3 µm, while CO_2 focuses on features near 2.0 µm. Importantly, the model successfully captures the O_2 0.76 µm A-band, which is one of the strongest, most accessible oxygen features for direct imaging; detection of a pronounced A-band can indicate substantial atmospheric O_2, which on rocky planets is often associated with oxygenic photosynthesis and potential biological activity (Schwieterman et al. 2018). However, the O_2 1.27 µm band receives only modest attention, while that same region shows relatively strong attention for CO_2, which may contribute to their observed error correlation (Figure 3). Interestingly, this is consistent with how retrieval methods like that used for the Total Carbon Column Observatory Network use the 1.27 µm O_2 band as a proxy for total air column when inferring CO_2 columns (Bertaux et al. 2020), suggesting that while the model did not successfully predict O_2 flux, it learned other scientifically relevant relationships embedded in this spectral region, such as its role as a proxy for total atmospheric column, which can indirectly inform CO_2 predictions. We also note partially overlapping short-wave infrared attention for CH_4 and CO_2 in the 2.0–2.3 µm range, which may also contribute to correlated prediction errors (Figure 3), though CO_2's strongest sensitivity remains closer to 2.06 µm. Finally, we observe attention patterns outside the conventionally highlighted absorption bands, suggesting either error in the model, or identification of additional subtle or blended spectral features that contribute to its predictions. It is important to note that Figure 2 reflects the attention pattern for a single spectrum; if the model pays little attention to a canonical absorption region for a given molecule, it may simply indicate that the corresponding species is absent or weak in that specific spectrum.

Figure 4a and Table 1 show high R^2 across species for both BCNN and SQuAT. This close performance ($R^2 = 0.972$ for BCNN and $R^2 = 0.985$ for SQuAT) on all-species regression suggests that both models are near the representational ceiling given the SNR ratio and data complexity. On all-species regression SQuAT is only marginally above BCNN (0.985 vs. 0.972), and gaps are small for SNR$\geq$ 20. By design, these variants were not intended to raise point accuracy but to add capabilities. BCNN replaces deterministic layers with variational ones to quantify predictive uncertainty, and SQuAT augments a ViT with gas-query cross-attention and spectroscopic priors to expose species-aligned attributions. Consistent with this intent, their point accuracy largely tracks the CNN/ViT baselines; the added value is complementary to accuracy. BCNN provides predictive distributions with credible intervals for risk-aware use, and SQuAT yields attention maps that localize the wavelengths driving each prediction. Thus, even when R^2 and RMSE are nearly indistinguishable, the models supply the quantified uncertainty and interpretable attributions needed for traceable flux retrievals and observation planning.

The shape and magnitude of the 95% credible intervals in Figure 4b provide insight into each model's internal uncertainty calibration. While both models show narrow bands for well-constrained species like O_2 and CH_4, the SQuAT appears to express sharper and more localized uncertainty, potentially reflecting its attention-based structure and discrete focus on relevant wavelengths. Conversely, the BCNN shows smoother uncertainty profiles that may stem from the global nature of convolutional filtering and weight-level variational inference. These patterns suggest transformer-based models might be better suited for uncertainty disentanglement in spectrally entangled regimes, though further calibration metrics would be needed to quantify this.

Our results highlight distinct strengths of the BCNN and the SQuAT, each making them particularly suitable for different operational and scientific priorities in exoplanet biosignature flux prediction. The BCNN exhibits notable performance in terms of robust uncertainty quantification, which is crucial in scenarios demanding reliability under varied observational conditions. Its probabilistic nature provides inherent flexibility to capture both aleatoric and epistemic uncertainties. Thus, the BCNN is particularly well-suited to mission planning phases where conservative decision-making is necessary. For instance, it can significantly contribute to preliminary prioritization of observational targets by efficiently identifying planetary atmospheres with the greatest scientific potential, even under conditions of limited data quality or incomplete spectral coverage. The smoother uncertainty profile generated by the

BCNN (as observed in Figure 4b) also suggests potential advantages for broader generalization across observational regimes with varying noise levels.

In contrast, the SQuAT leverages a query-driven attention mechanism to explicitly disentangle molecule-specific spectral features. This architectural choice enhances interpretability, allowing scientists to directly associate predictions with known molecular absorption bands. SQuAT's structured attention mechanism enables focused extraction of spectral signals, making it particularly suitable for detailed scientific analysis where understanding the specific spectral drivers behind predictions is essential. The ability to visualize per-molecule attention maps (as demonstrated in Figure 2) provides valuable insight into the underlying physics captured by the model, making SQuAT an excellent candidate for exploratory data analysis and hypothesis generation. Furthermore, SQuAT's multi-scale processing and explicit incorporation of spectroscopic priors enhance its robustness to spectral overlap between molecules, increasing confidence in disentangled flux estimations.

Interestingly, despite these differences in architecture and interpretability, both models achieve comparable and exceptionally high predictive performance (Figure 4a and Table 1). This observation underscores an important point: while architectural sophistication (as exemplified by SQuAT) provides significant interpretative and analytical advantages, the simpler BCNN remains highly effective and may represent a preferable choice when computational resources are constrained, or when interpretability is not the priority. **Ultimately, the complementary strengths of SQuAT and BCNN suggest an ensemble approach may address the dual requirements of future flagship missions seeking signs of habitability**.

7 Limitations and Future Work

One main limitation for our models is the lack of real-world observations. Presently, the vast majority of exoplanet spectra come from the transit method, which is mainly successful on planets significantly hotter than Earth and unlikely to contain life as we know it (Stark et al. 2020). Additionally, simulated data have their own limitations and biases. For example, the Atmos simulation described in Section 3 does not properly converge at higher temperature regimes (Virtual Planetary Laboratory 2025), and clouds and hazes remain challenging to model realistically.

The scarcity of observational data necessitates comprehensive synthetic datasets. We will augment the dataset to include a broader diversity of exoplanetary atmospheres and incorporate emerging simulation capabilities as they become available. A larger dataset also enables development of a pre-trained foundation model, with biosignature flux prediction as a downstream fine-tuning task.

Increasing model robustness at low SNR is a key priority, because practical applications must assume minimal signal quality. We will explore other Bayesian deep learning approaches, integrate additional physics priors into learned representations, and develop attention mechanisms that better capture localized spectral dependencies while disentangling overlapping molecular features.

8 Path to Deployment

The Habitable Worlds Observatory (HWO) is envisioned as NASA's next flagship direct-imaging mission, designed to detect and characterize Earth-sized exoplanets in the habitable zones of nearby stars (Zurlo 2024). With its enormous cost, finite mission lifetime, and extremely limited observing-time budget, HWO must maximize scientific return from every observation. An AI-assisted decision pipeline offers a compelling route to achieve this, and our work represents a critical step toward this by demonstrating high-accuracy, uncertainty-aware biosignature flux inference from realistic reflected-light spectra. Together, our models satisfy two mission-critical AI requirements: risk-aware decision-making and scientifically traceable predictions.

To reach deployment, milestones are structured according to the **Technology Readiness Level** (TRL) framework, a systematic metric used by NASA and other space agencies to assess status and operational readiness of mission technologies (NASA 2016). Progress to lower-mid TRL includes advancing our models with increasingly realistic synthetic spectra training data covering diverse planetary atmospheres, surface types, clouds, and hazes. These are fine-tuned on HWO-specific end-to-end simulations to incorporate realistic noise, coronagraph bandpasses, and throughput models, ensuring hardware-resilient algorithms. To reach higher TRL, simulations would quantify how AI-guided prioritization improves science yield under realistic operational constraints like slew times and visibility windows. Once validated, models can be deployed on the ground during commissioning and early science operations, serving as decision-support tools alongside human experts. Upon achieving highest TRL, models can operate within the Science Operations Center (SOC) for real-time observation triage.

9 Conclusion

In this study, we introduced and evaluated two deep-learning approaches for predicting exoplanet biosignature fluxes: a BCNN and SQuAT. Both address key requirements of future direct-imaging missions, including robust uncertainty quantification and interpretability. The BCNN excelled at capturing epistemic and aleatoric uncertainties, ideal for limited data scenarios. SQuAT, our novel model architecture, provided superior interpretability through explicit attention mechanisms aligned with biosignature absorption priors, facilitating exploratory analyses and hypothesis validation.

Both models exhibited comparably high predictive accuracy, reinforcing their potential as complementary components within a unified analytical pipeline. Identified limitations, such as spectral confusion between different biosignatures, highlight areas for improvement and establish a need for broader observational scenarios and ensemble methods that combine Bayesian inference with query-based transformers. Overall, our findings highlight the significant promise and flexibility of machine-learning approaches in the search for habitable worlds and their role in maximizing scientific yield from upcoming flagship missions.

References

Bertaux, J.-L.; et al. 2020. The use of the 1.27 μm O_2 absorption band for greenhouse gas retrievals in TCCON. *Atmospheric Measurement Techniques*, 13(6): 3329–3344.

Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; and Wierstra, D. 2015. Weight Uncertainty in Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, 1613–1622.

Borucki, W. J.; Koch, D.; Basri, G.; Batalha, N.; Brown, T.; Caldwell, D.; Caldwell, J.; Christensen-Dalsgaard, J.; Cochran, W. D.; DeVore, E.; Dunham, E. W.; Dupree, A. K.; Gautier, T. N.; Geary, J. C.; Gilliland, R.; Gould, A.; Howell, S. B.; Jenkins, J. M.; Kondo, Y.; Latham, D. W.; Marcy, G. W.; Meibom, S.; Kjeldsen, H.; Lissauer, J. J.; Monet, D. G.; Morrison, D.; Sasselov, D.; Tarter, J.; Boss, A.; Brownlee, D.; Owen, T.; Buzasi, D.; Charbonneau, D.; Doyle, L.; Fortney, J.; Ford, E. B.; Holman, M. J.; Seager, S.; Steffen, J. H.; Welsh, W. F.; Rowe, J.; Anderson, H.; Buchhave, L.; Ciardi, D.; Walkowicz, L.; Sherry, W.; Horch, E.; Isaacson, H.; Everett, M. E.; Fischer, D.; Torres, G.; Johnson, J. A.; Endl, M.; MacQueen, P.; Bryson, S. T.; Dotson, J.; Haas, M.; Kolodziejczak, J.; Van Cleve, J.; Chandrasekaran, H.; Twicken, J. D.; Quintana, E. V.; Clarke, B. D.; Allen, C.; Li, J.; Wu, H.; Tenenbaum, P.; Verner, E.; Bruhweiler, F.; Barnes, J.; and Prsa, A. 2010. Kepler Planet-Detection Mission: Introduction and First Results. *Science*, 327(5968): 977.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers.

Chopra, A.; Bell, A. C.; Fawcett, W.; Talebi, R.; Angerhausen, D.; Baydin, A. G.; Berea, A.; Cabrol, N. A.; Kempes, C.; and Mascaro, M. 2023. PyATMOS: A Scalable Grid of Hypothetical Planetary Atmospheres. *arXiv preprint arXiv:2308.10624*.

Claire, M. W.; Sheets, J.; Cohen, M.; Ribas, I.; Meadows, V. S.; and Catling, D. C. 2012. The Evolution of Solar Flux from 0.1 nm to 160 μm: Quantitative Estimates for Planetary Studies. *The Astrophysical Journal*, 757(1): 95.

Cobb, A. D.; Himes, M. D.; Soboczenski, F.; Zorzan, S.; O'Beirne, M. D.; Baydin, A. G.; Gal, Y.; Domagal-Goldman, S. D.; Arney, G. N.; and Angerhausen, D. 2019. An Ensemble of Bayesian Neural Networks for Exoplanetary Atmospheric Retrieval.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fisher, C.; Hoeijmakers, H. J.; Kitzmann, D.; Márquez-Neila, P.; Grimm, S. L.; Sznitman, R.; and Heng, K. 2019. Interpreting High-Resolution Spectroscopy of Exoplanets Using Cross-Correlations and Supervised Machine Learning.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, 1050–1059.

Johnsen, T. K.; and Marley, M. S. 2019. Multilayer Perceptron and Geometric Albedo Spectra for Quick Parameter Estimations of Exoplanets.

Kendall, A.; and Gal, Y. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, 5574–5584.

Kopparapu, R. K.; Ramirez, R. M.; SchottelKotte, J.; Kasting, J. F.; Domagal-Goldman, S.; and Eymet, V. 2014. Habitable Zones around Main-sequence Stars: Dependence on Planetary Mass. *The Astrophysical Journal Letters*, 787(2): L29.

Madhusudhan, N. 2018. Atmospheric Retrieval of Exoplanets. *Springer Handbook of Exoplanets*.

NASA. 2016. NASA Systems Engineering Handbook. Technical Report NASA/SP-2016-6105 Rev2, NASA. See Appendix G.4 for Technology Readiness Levels.

Pham, D.; and Kaltenegger, L. 2022. Follow the Water: Finding Water, Snow and Clouds on Terrestrial Exoplanets with Photometry and Machine Learning.

Roberge, A.; and Moustakas, L. A. 2018. The Large Ultraviolet/Optical/Infrared Surveyor. *Nature Astronomy*, 2: 605–607.

Rothman, L. S.; Rinsland, C.; Goldman, A.; Massie, S.; Edwards, D.; Flaud, J.; Perrin, A.; Camy-Peyret, C.; Dana, V.; Mandin, J.-Y.; et al. 1998. The HITRAN molecular spectroscopic database and HAWKS (HITRAN atmospheric workstation): 1996 edition. *Journal of quantitative spectroscopy and radiative transfer*, 60(5): 665–710.

Schwieterman, E. W.; Kiang, N. Y.; Parenteau, M. N.; Harman, C. E.; DasSarma, S.; Fisher, T. M.; Arney, G. N.; Hartnett, H. E.; Reinhard, C. T.; Olson, S. L.; et al. 2018. Exoplanet biosignatures: a review of remotely detectable signs of life. *Astrobiology*, 18(6): 663–708.

Soboczenski, F.; Himes, M. D.; O'Beirne, M. D.; Zorzan, S.; Baydin, A. G.; Cobb, A. D.; Gal, Y.; Angerhausen, D.; Mascaro, M.; Arney, G. N.; and Domagal-Goldman, S. D. 2018. Bayesian Deep Learning for Exoplanet Atmospheric Retrieval.

Stark, C. C.; Dressing, C.; Dulz, S.; Lopez, E.; Marley, M. S.; Plavchan, P.; and Sahlmann, J. 2020. Toward complete characterization: prospects for directly imaging transiting exoplanets. *The Astronomical Journal*, 159(6): 286.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need.

Villanueva, G. L.; Smith, M. D.; Protopapa, S.; Faggi, S.; and Mandell, A. M. 2018. Planetary Spectrum Generator: An accurate online radiative transfer suite for atmospheres, comets, small bodies and exoplanets. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 217: 86–104.

Virtual Planetary Laboratory. 2025. VPL Climate Photochemistry Model (Atmos). https://github.com/VirtualPlanetaryLaboratory/atmos/tree/master.

Yip, K. H.; Changeat, Q.; Nikolaou, N.; Morvan, M.; Edwards, B.; Waldmann, I. P.; and Tinetti, G. 2020. Peeking inside the Black Box: Interpreting Deep Learning Models for Exoplanet Atmospheric Retrievals.

Zurlo, A. 2024. Direct imaging of exoplanets. *arXiv preprint arXiv:2404.05797*.