Participant / Team Code: Comp-059

FlexARE: A modular framework for exoplanet atmospheric retrievals using generative preprocessing and simulation-based inference

Derrick Wong Longhin

Project Supervisor: Mr. Shearman, Head of Physics (German Swiss International School)

August 14, 2025

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Abstract

The recent launch of the James Webb Space Telescope (JWST) and the forthcoming Atmospheric Remote-sensing Infrared Exoplanet Large-survey (ARIEL) mission have positioned exoplanets at the forefront of astrophysical research. As the exploration of exoplanets advances, the challenge of efficient analysis of their atmospheres has become increasingly critical. Traditional methods, primarily based on Bayesian inference techniques, are often slow and resource-intensive, limiting their effectiveness.

In this context, this paper pioneers a **novel, flexible, and modular framework** that integrates deep learning concepts into classical retrieval methodologies, not only enhancing computational efficiency by an estimated minimum speed up of 360% compared to MultiNest but also increasing the flexibility of these retrievals significantly, allowing for a more streamlined retrieval process. In the following sections, we will demonstrate and exploit the effectiveness of this modular pipeline, utilising denoising autoencoder (DAE) [1] + generative adversarial network (GAN) [2] system for data processing along with the simulation based inference model TSNPE [3] for parameter prediction, achieving a final reconstruction accuracy of 99.05% and a parameter prediction accuracy of 89.2%. We will thoroughly explore the endeavours undertaken to enhance the model, notably through the incorporation of novel truncated and score-based methods for parameter inference which enhances sampling efficiency and optimisation. By addressing the limitations of current methods, this framework seeks to streamline the atmospheric retrieval process, facilitating more efficient analyses of exoplanet atmospheres and allowing for the rapid discovery of new insights.

Keywords: Exoplanet, Atmospheric Retrieval, Deep Learning, Modular Framework, Simulation-Based Inference, Computational Astrophysics

1 Introduction

The exploration of exoplanets has surged in recent years, driven by the discovery of over 5000 of these distant worlds and the tantalising possibility of uncovering atmospheres that may harbour life. The James Webb Space Telescope (JWST) and the upcoming ARIEL mission are set to further inundate exoplanet science with high-fidelity spectral data, transforming atmospheric characterization. This large sum of data, however, exposes a critical computational bottleneck: today's gold-standard retrieval frameworks, such as MultiNest and other nested sampling variants [4], scale poorly with data complexity, often demanding days of processing time for a single planetary spectrum. Consequently, our ability to analyse these rich datasets, not the acquisition of them, is rapidly becoming the rate-limiting step in exoplanet science.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Recent machine learning approaches have attempted to address this, but they present a stark trade-off. Amortized inference models offer remarkable speed after an intensive training phase, but their rigid nature falters when faced with new instrument noise profiles or incomplete data. Conversely, flexible simulation-based inference (SBI) methods can adapt to varied data quality but reintroduce a significant computational cost for each new target. The community is thus forced to choose between speed and adaptability.

This paper introduces a two-stage framework that breaks this dilemma by decoupling data conditioning from Bayesian inference. We first employ a generative pipeline - integrating an autoencoder with a Generative Adversarial Network (GAN) - to denoise, complete, and standardize raw spectra into a clean representation. This uniform output is then processed by a state-of-the-art SBI model, such as a truncated or score-based posterior estimator, to retrieve atmospheric parameters. Because the inference engine always operates on the standardized latent space, it remains agnostic to the noise and observational idiosyncrasies of the original input, eliminating the need for per-target retraining.

Our primary contributions are:

- A flexible, decoupled framework that separates generative data conditioning from simulation-based inference, resolving the prevailing speed-versus-adaptability conflict in machine learning-based retrievals.
- A robust generative pre-processor that leverages autoencoders and GANs to effectively handle noisy and incomplete spectral data, achieving a 99.05% reconstruction accuracy.
- The first pairing of this generative conditioner with advanced posterior estimation models, which achieves an 89.2% parameter retrieval accuracy while demonstrating a significant computational speed-up over traditional methods.
- A comprehensive evaluation of model architectures and data-tuning strategies to optimize the pipeline for the challenges posed by next-generation spectral datasets.

The remainder of this paper is organized as follows: Section 2 surveys the landscape of current retrieval methods. Section 3 provides a detailed architectural overview of our proposed solution. Section 4

2025 S-T Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

benchmarks its performance and accuracy against established techniques. Finally, Section 5 discusses the broader implications of our findings, acknowledges limitations, and outlines future research directions.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

2 Related works

The challenge of atmospheric retrieval has spurred the development of diverse computational strategies, each navigating a fundamental trade-off between mathematical rigor, computational speed, and observational flexibility. Early deep learning approaches showed promise but often had limitations; for instance, GAN-based retrievals struggled with parameter estimation [22], while Random Forest methods were often confined to low-dimensional problems [23-25], and early CNNs like ExoCNN could not produce complex, multi-modal posteriors.

This has led the field to converge on Simulation-Based Inference (SBI) as the most promising path forward. This section surveys the three dominant paradigms within this modern context - classic sampling, amortized inference, and sequential inference - to identify the architectural gap that our work addresses.

2.1 Classic Bayesian Sampling: The Rigorous but Slow Gold Standard

Traditional atmospheric retrieval relies on robust Bayesian sampling techniques. Nested sampling algorithms, as implemented in tools like MultiNest [4], and MCMC variants like Hamiltonian Monte Carlo (HMC) [19] and the No-U-Turn Sampler (NUTS) [20], are considered the gold standard for their mathematical rigor. These methods are the most reliable as they explore the full posterior distribution, providing parameter uncertainties and direct log-evidence values for Bayesian model comparison.

However, their strength is also their primary weakness. The computational cost scales super-linearly with the number of model parameters, as each of the thousands of required likelihood evaluations demands a full forward model simulation. This leads to prohibitive runtimes, often lasting up to days for a single JWST spectrum, rendering them impractical for the large-scale surveys that will define the next era of exoplanet science.

2.2 Amortized Deep Inference: Fast but Inflexible

To break the computational scaling barrier, amortized inference methods leverage deep learning to learn a direct mapping from an observation to its posterior distribution. After a significant, one-time, up-front training cost, inference becomes nearly instantaneous.

The power of this approach has been demonstrated by multiple groups. The work of Vasist et al. [7] used Neural Posterior Estimation (NPE) to achieve a 4000x speedup over MultiNest, but this required a

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

staggering 17,000 CPU hours for training data generation. More recently, the FASTER framework [Lueber et al., 2025] has emerged as the state-of-the-art in high-throughput retrieval. It uses Neural Ratio Estimation (NRE) to achieve inference in milliseconds, is Bayesian-complete (providing both posteriors and model probabilities), and has been validated on real JWST data.

However, while providing unparalleled speed, these powerful amortized models are fundamentally brittle. They are trained for a specific instrument and a pre-defined statistical noise model. Their performance degrades significantly when applied to data with different characteristics—for example, spectra from a different instrument like ARIEL, observations with higher-than-expected noise, or data with unexpected artifacts or missing wavelength coverage. For each new observational setup, the entire computationally expensive training pipeline must be repeated.

2.3 Sequential Simulation-Based Inference: Flexible but Costly

Positioned between the previous two extremes, sequential (or non-amortized) SBI methods adapt the inference process to a single, specific observation. The core method of Sequential Neural Posterior Estimation (SNPE) [15], as adapted for exoplanets in frameworks like Floppity [6], iteratively refines a proposal distribution to focus the simulation budget on regions of high posterior probability. This achieves high posterior fidelity with fewer forward model calls than classic samplers and offers crucial flexibility to handle unique, individual observations that amortized models cannot.

However, the drawback is that this process must be repeated from scratch for every new target, making it an inefficient use of computational resources when analyzing large datasets. While recent algorithmic advances like Truncated SNPE (TSNPE) [16] and score-based methods like SNPSE [3] improve the sample efficiency of this process, they do not eliminate the fundamental problem: the computational effort is spent on a per-target basis and is not amortized across a population.

2.4 The Architectural Gap and Our Contribution

This analysis reveals that the bottleneck is not the inference engines themselves, but the lack of a standardized interface between raw, heterogeneous observational data and these powerful tools. The components needed to bridge this gap—generative models for data cleaning—exist in isolation but have not been integrated into a unified retrieval pipeline. For instance, denoising autoencoders are a standard technique for noise reduction in spectroscopy [9], while Generative Adversarial Networks (GANs) have been explored for data imputation in other fields [8].

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

To date, no framework has systematically coupled a generative data-conditioning front-end with a flexible SBI back-end. This leaves the community choosing between the fast-but-brittle or flexible-but-slow paradigms. We resolve this dilemma by proposing a new paradigm of architectural decoupling. Our framework, FlexARE, introduces a generative pre-processor that standardizes raw, noisy, and potentially incomplete spectra into a clean, physically plausible representation. This conditioned spectrum then serves as a high-quality input for a downstream inference engine. This two-stage architecture synthesizes the strengths of prior approaches:

- It grants flexibility to brittle amortized models by transforming novel data into the format they were trained on, saving thousands of hours of retraining.
- It grants robustness to sequential models by ensuring their computational effort is focused on the true physical signal, not instrumental artifacts, leading to more accurate and reliable posteriors.

The following section provides a detailed overview of this architecture, explaining how the generative and inferential modules are designed and integrated to create a truly and flexible retrieval solution.

3 Methodology

3.1 Overview

To address the objective comprehensively, this research will employ a modular approach focused on two primary objectives: data processing and atmospheric retrieval. Each task utilizes separate architectures with dedicated training and testing phases, allowing for tailored optimization strategies and reduced complexities through specialized datasets. The experimental methodology comprises three main components: dataset creation and preprocessing (3.2), the generative preprocessing pipeline (3.3) and the parameter retrieval pipeline (3.4).

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

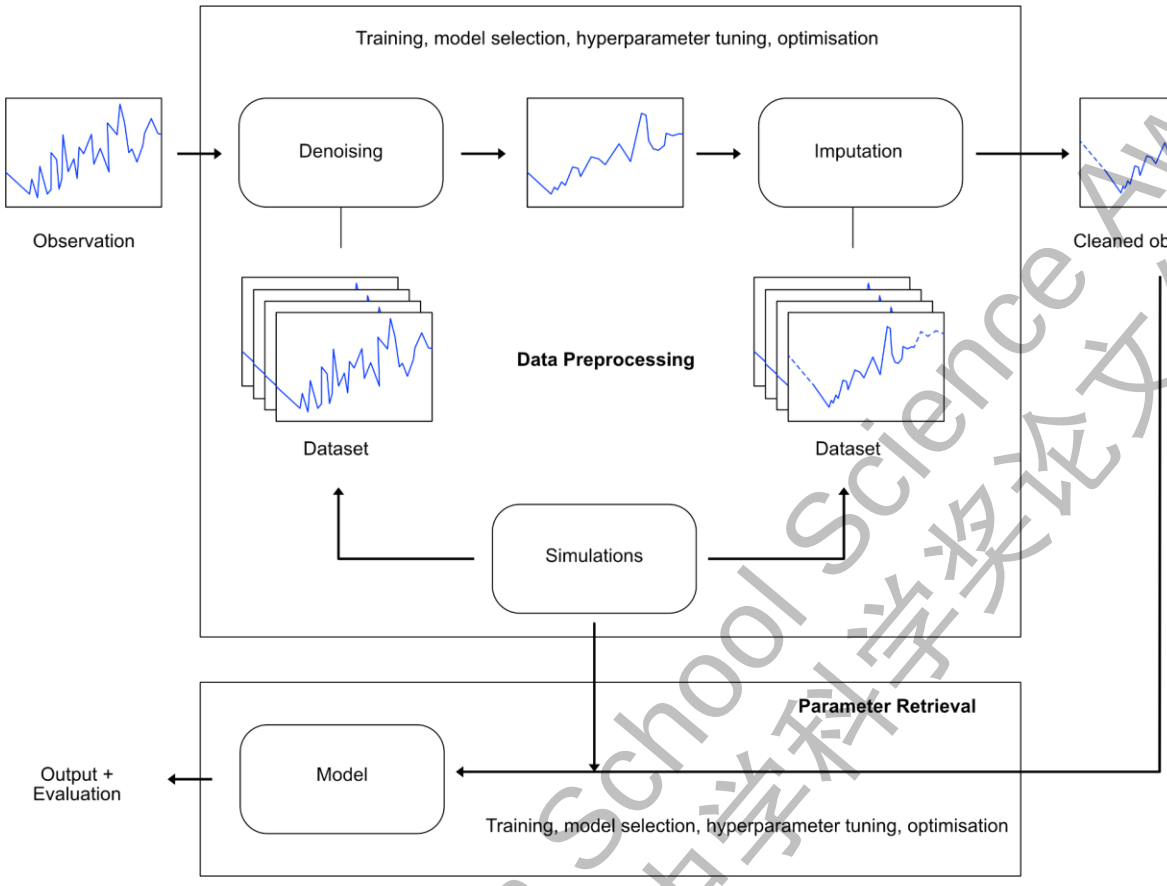

Figure 1 Flow chart showing tasks mentioned in this paper

The following sections shall delve into each objective separately and in more detail.

3.2 Dataset Creation and Preprocessing

3.2.1 Synthetic Dataset Generation

To conduct this experiment, it was imperative to obtain a sufficiently large dataset for training the generative networks. Due to the scarcity of existing real data, I opted to create a synthetic dataset consisting of clean, noisy, and complete spectra within the wavelength range of 0.3 to 15 μm. This range is optimal for identifying key spectral features and aligns with the capabilities of both current and future instruments. For this purpose, I utilized the TauREx 3 (Tau Retrieval for Exoplanets) [17] framework, which is recognized for its user-friendly interface and computational efficiency. However, to develop a dataset of spectra that more accurately reflects real-world conditions, I selected a more complex model. This decision aimed to enhance the fidelity of the generated spectra while minimizing overall computation time.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Instead of employing a simple isothermal temperature profile, I implemented a modified two-stream approximation based on the model proposed by Guillot (2010) [10][11]. This approach allows for temperature variations with altitude, thereby capturing the thermal structure of exoplanetary atmospheres more accurately. Additionally, I incorporated a free chemistry framework, treating the abundances of key molecules such as H_2O, CO, CO_2, CH_4, and NH_3 as variable parameters rather than fixed values based on chemical equilibrium. This choice was made to enhance the realism of the spectra, despite the added computational complexity. Furthermore, I introduced a pressure-dependent infinitely opaque cloud layer and included a Mie scattering contribution as described in Lee et al. (2013) [12]. The prior ranges and opacity citations are detailed in Table 1.

Parameter	Prior Range	Reference
$R_P\ (R_J)$	$\mathcal{U}(0.1, 2.0)$	-
$\log g$	$\mathcal{U}(2.3, 4.3)$	-
$T_{irr}\ (K)$	$\mathcal{U}(200, 4000)$	-
$\log \kappa_1$	$\mathcal{U}(-4, 0)$	-
$\log \kappa_2$	$\mathcal{U}(-4, 0)$	-
$\log \kappa_{irr}$	$\mathcal{U}(-4, 0)$	-
α	$\mathcal{U}(0, 1)$	-
$\log H_2O$	$\mathcal{U}(-12, -2)$	Polyansky et al. (2018)
$\log CH_4$	$\mathcal{U}(-12, -2)$	Yurchenko et al. (2020)
$\log CO$	$\mathcal{U}(-12, -2)$	Li et al. (2015)
$\log CO_2$	$\mathcal{U}(-12, -2)$	Coles et al. (2019)
$\log NH_3$	$\mathcal{U}(-12, -2)$	Yurchenko et al. (2024)
$\log p_{cloud}$	$\mathcal{U}(2, 6)$	-
$\log \alpha_{lee}$	$\mathcal{U}(-1, 1)$	-
$\log Q_{ext}$	$\mathcal{U}(-1, 1)$	-
$\log \chi$	$\mathcal{U}(-40, -4)$	-

Table 1 *Prior values for synthetic dataset*

I generated 100,000 clean spectra, which required about 250 minutes on a cluster featuring 36.6 vCPUs with an AMD EPYC 7763 processor, totaling approximately 150 CPU hours. However, many simulations produced NaN or infinite values due to overflow errors, likely caused by the broad parameter ranges resulting in numerous invalid combinations.

Despite extensive parameter tuning, I could not resolve these issues. I decided to prune the dataset to remove NaN and infinite values, reducing the count from 100,000 to roughly 81,000 spectra. Additionally, I identified and eliminated spectra with excessively high transit depths and applied criteria to filter out

2025 S-T Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

negative values and ensure normalized values did not exceed one. This process yielded a final dataset of approximately 63,000 spectra.

3.2.2 Noise and Binning

After simulating the spectra, I utilized the public package PandExo [18] to add JWST instrument noise to the full spectrum, using the star parameters of HD 209458. I selected the NIRSpec PRISM, as prior observations of the exoplanet Wasp 39b with this instrument would allow for comparison with actual data and the best-fit model. Consequently, I applied a noise floor of 15 ppm, representing the instrument's upper limit. Following the addition of noise, I obtained the wavelength binning from the noisy spectrum and rebinned the clean spectra to match this wavelength range, producing a corresponding clean version of the noisy spectra. A comparison is provided below.

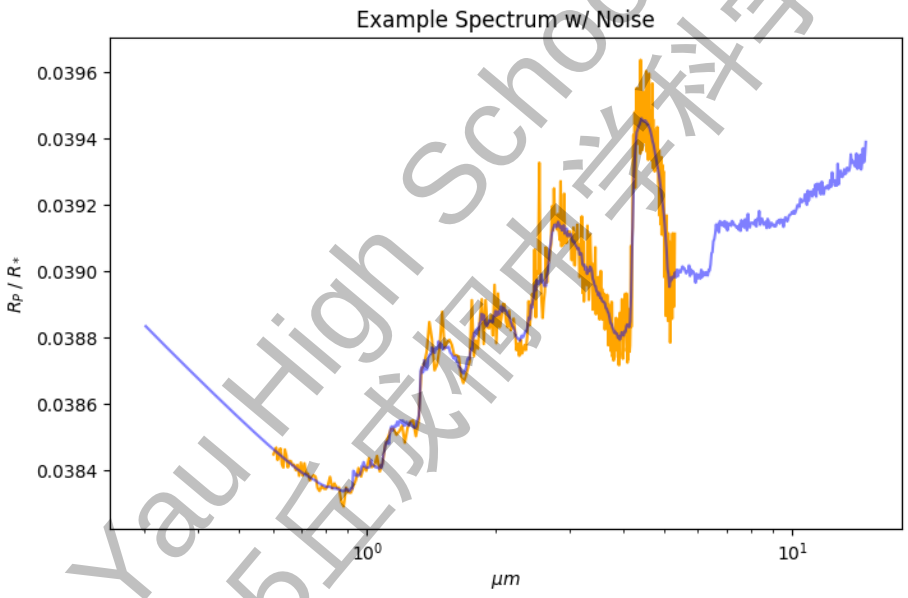

Figure 2 *Example spectrum with noise*

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

3.3 Generative Preprocessing Pipeline

Data processing plays a crucial role in the analysis of exoplanetary spectra, offering significant advantages both in general and specifically for this field. Previous works have already identified the problem of the inflexibility resulting in being limited to a singular wavelength grid and the need to retrain on different observational noise. As a result, data processing can be utilised as a way to generalise the input and can be divided into two primary processes: denoising and generation.

3.3.1 Denoising

A common trend in the field is to simply rely on the advancements in the field of machine learning when developing a solution to a problem. In this case, a particularly apt method commonly used for denoising is known as an **autoencoder**.

Autoencoders are a type of neural network architecture and consist of two functions: an encoder and a decoder. The encoder, typically a multi-layer perceptron or network, passes the original input data $x \in \mathcal{X}$ through hidden layers, reducing its dimensionality into its latent representation $E(x)$ also known as z. The denoiser denoted D then decodes z, increasing its dimensionality and creating a reconstruction of the input denoted $\hat{x}$.

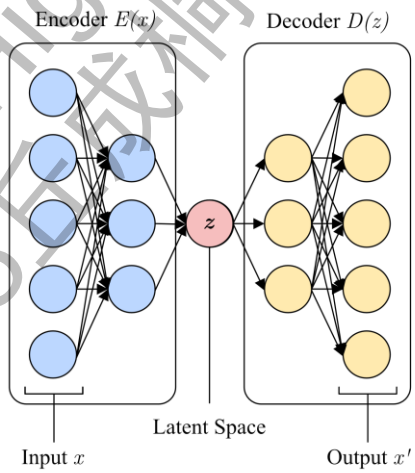

Figure 3 *Diagram of a simple autoencoder*

The dimensionality bottleneck created by the latent space z discourages perfect duplication or memorization of the signal, instead encouraging the model to extract salient features and learn a more efficient representation of the information.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Central to the accurate reproduction of inputs within autoencoders is the concept of the reconstruction error, quantifying the difference between the original input x and its reconstruction $\hat{x} = D(E(x))$. This is typically measured with a loss function such as the mean squared error or cross entropy.

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(x_i - D(E(x_i)))^2$$

Formula 1 *Typical autoencoder loss function*

However, to measure the accuracy of the reconstructions, I opted to utilise RMSE instead, as it could reliably represent accuracy as a percentage.

$$1 - \text{RMSE} = 1 - \left(\frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\bar{y}} \times 100\right)$$

Formula 2 *Evaluation Criteria / Metric*

Overall, the properties of this type of generative model can be exploited for denoising purposes, allowing it to learn key features present in the data separate from the noise itself. To test this, I implemented a simple dense autoencoder with the following architecture:

Component	Layer	Output Shape	Parameters
Encoder	Input	403	0
	Dense + ReLU	256	103,424
	Dense + ReLU	128	32,896
	Dense + ReLU	64	8,256
Decoder	Dense + ReLU	128	8,320
	Dense + ReLU	256	33,024
	Dense	403	103,571
		Total Parameters	**289,491**

Table 2 *Simple dense autoencoder architecture*

I implemented normalization of the clean spectra to align with the highs and lows of the noisy spectra. This approach ensures that in the final forward pass with real data, which will only include the noisy data, there are no dependencies on the clean data ranges. Normalizing to the noisy spectra is effective, as it typically exhibits more extreme highs and lows than the clean spectra. After training for 100 epochs using mean squared error loss (as shown in Formula 1) on a limited dataset comprising approximately 6,000 spectra (about 10% of the original size), the results were promising.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

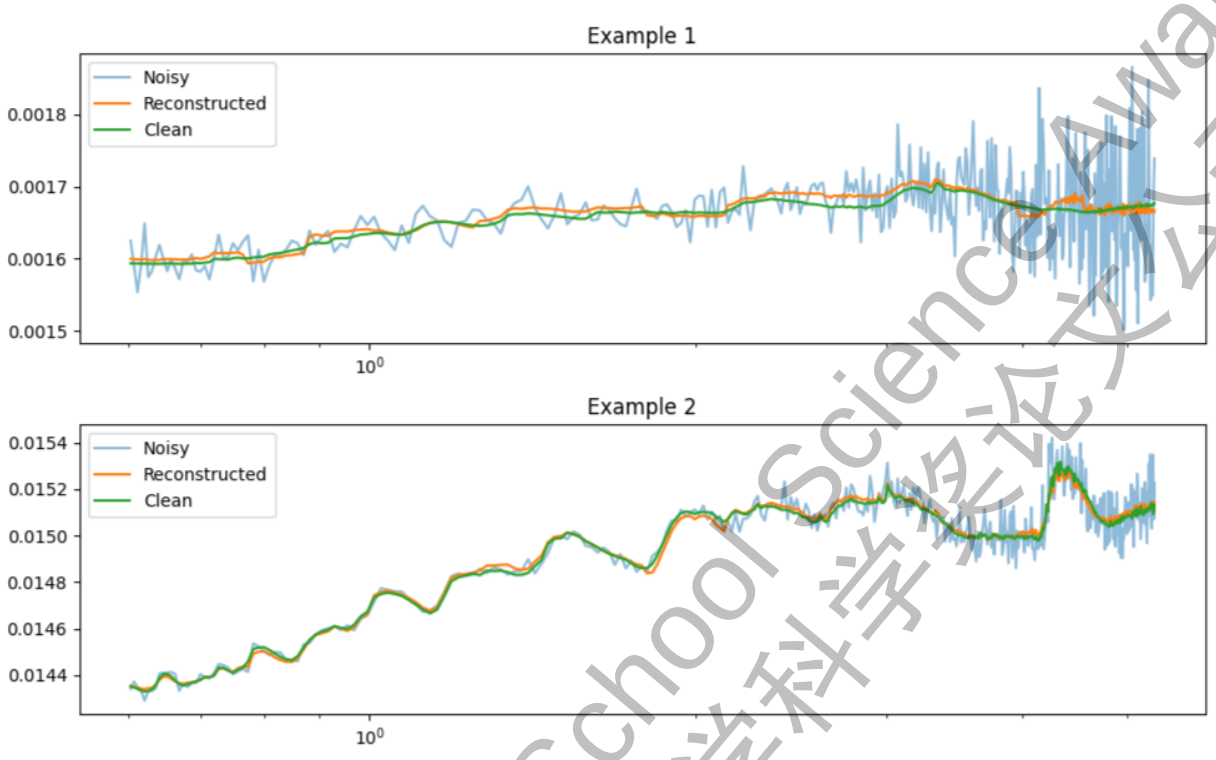

Figure 4 *Dense Autoencoder: Initial results*

I was pleasantly surprised by the model's performance despite the significant noise, as demonstrated in Example 1. This supports my hypothesis that the autoencoder could effectively denoise spectra. However, noticeable mismatches in certain spectral regions, even with minimal noise (see Example 2), indicated that the model still had limitations. Consequently, I decided to explore alternative autoencoder implementations, specifically a **convolutional** autoencoder and a **long short-term memory** (LSTM) autoencoder.

A **convolutional** autoencoder integrates convolutional layers to enhance image processing capabilities. In this architecture, convolutional layers apply filters—small, learnable matrices—to the input images, enabling the model to extract specific features such as edges and textures. The process of convolution involves sliding these filters across the image and computing dot products to create feature maps, which represent the presence of identified features. Following this, pooling layers reduce the spatial dimensions of the feature maps, summarizing the information while maintaining the most salient features. This dimensionality reduction not only decreases computational complexity but also enhances the model's robustness to variations in the input, such as translation and distortion. The encoder compresses these extracted features into a lower-dimensional bottleneck representation, while the decoder reconstructs the original image using transposed convolutional layers. Overall, CNN autoencoders effectively learn

2025 S-T Yau High School Science Award
仅用于2025正成国中学科学论文公示

complex representations for tasks like denoising and dimensionality reduction, leveraging the strengths of convolutional and pooling operations.

On the other hand, **Long Short-Term Memory (LSTM)** networks are a specialized type of recurrent neural network (RNN) designed to effectively handle sequential data, particularly one-dimensional (1D) data like time series. LSTMs excel at capturing long-range dependencies due to their unique architecture, which features memory cells and gating mechanisms—including the input gate, forget gate, and output gate. These gates regulate the flow of information, allowing the network to retain relevant data over extended periods while discarding irrelevant information. This structure alleviates the vanishing gradient problem commonly faced by standard RNNs. Overall, LSTMs provide a robust framework for managing sequential information, making them particularly effective for understanding temporal patterns and contextual relationships in 1D data.

To ensure a fair comparison, I maintained a similar number of parameters across all models through adding extra layers and adjusting output dimensionality.

Component	Layer	Output Shape	Parameters
Conv Encoder	Input	(1, 403)	0
	Conv1d + ReLU	(16, 403)	64
	MaxPool1d	(16, 202)	0
	Conv1d + ReLU	(8, 202)	392
	MaxPool1d	(8, 101)	0
Dense Encoder	Flatten	808	0
	Linear + ReLU	128	103,552
	Linear + ReLU	64	8,256
Dense Decoder	Linear + ReLU	128	8,320
	Linear + ReLU	808	104,232
Conv Decoder	Reshape	(8, 101)	0
	Conv1d + ReLU	(8, 101)	200
	Upsample	(8, 202)	0
	Conv1d + ReLU	(16, 202)	400
	Upsample	(16, 403)	0
	Conv1d	(1, 403)	49
Total Parameters			**285,465**

Table 3 *Convolutional autoencoder architecture*

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

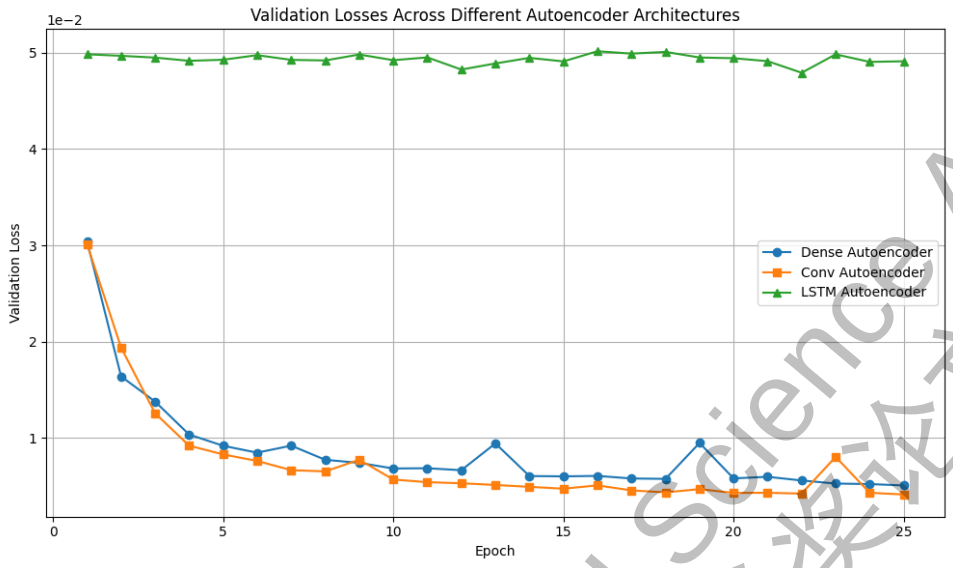

Figure 5 *Validation losses across autoencoder architectures*

Ultimately, both the dense and convolutional autoencoders showed promise, while the LSTM autoencoder seemed to perform poorly, with its loss plateauing within the first few epochs, indicating minimal learning. This may be attributed to the LSTM's complexity, its limitations in effectively capturing local patterns, and potential issues with gradient descent during training. While LSTMs are adept at handling sequential data, they may not be optimal for tasks that require a focus on the local features inherent in spectral data. For those more familiar with LSTM architectures, further exploration might uncover optimizations to enhance performance. However, I chose to proceed with the convolutional and dense autoencoders, as they demonstrated greater effectiveness for denoising in this context.

The loss graph clearly indicated that the convolutional autoencoder outperformed the dense autoencoder, prompting me to fine-tune and optimize this model. I conducted a sequential hyperparameter sweep, testing learning rates from 0.01 to 0.00001, activation functions (ELU, ReLU, GELU, and LeakyReLU), normalization functions (batchNorm, instanceNorm, groupNorm), dropout rates from 0 to 0.5, kernel sizes from 3 to 7, and hidden dimension configurations of [64, 32], [128, 64], [256, 128], and [512, 256].

I ultimately identified the optimal hyperparameters as a learning rate of 0.001, an ELU activation function, batch normalization, no dropout, a kernel size of 7, and hidden dimensions of [64, 32]. With these settings, I began training the optimal model on the full dataset. However, after early stopping at 30 epochs, the validation losses were unexpectedly poor. Previously, while testing the dense autoencoder on the full dataset for 100 epochs, I had halted training early at around 5 epochs due to excessive duration. I preserved the loss measurements for potential future reference, but the convolutional autoencoder I was currently

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

training yielded worse loss metrics than the initial dense autoencoder. This may be due to the convolutional structure being advantageous for smaller datasets, facilitating more effective backpropagation hence creating deceiving results. Consequently, I decided to switch my focus back to the dense autoencoder and conduct another hyperparameter sweep.

This time, I identified an optimal learning rate of 0.001, a GELU activation function, instance normalization, and a dropout rate of 0. However, this optimization appeared ineffective since the parameters were largely unchanged, apart from the activation function. The results of these optimisations can be found below:

Method	**1-RMSE (%)**	**Standard Deviation (%)**
After Hyperparameter Tuning	**99.36**	**2.17**
Before Tuning	99.10	2.49
Convolutional Network (ConvNet)	97.40	4.38

Table 4 *1 - RMSE before and after optimisation*

Despite the modest gains, it is clear that the optimisations did indeed benefit the model, confirming their potential. With the denoising process refined, I turn to the next phase of the generative pipeline: data imputation.

3.3.2 Generation

In the realm of data generation, Generative Adversarial Networks (GANs) emerge as a prominent choice. The key advantage of this model lies in its unique adversarial training mechanism, which involves two competing neural networks: the generator and the discriminator. The generator's objective is to produce samples that can deceive the discriminator, which is tasked with distinguishing between the real and generated spectra.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

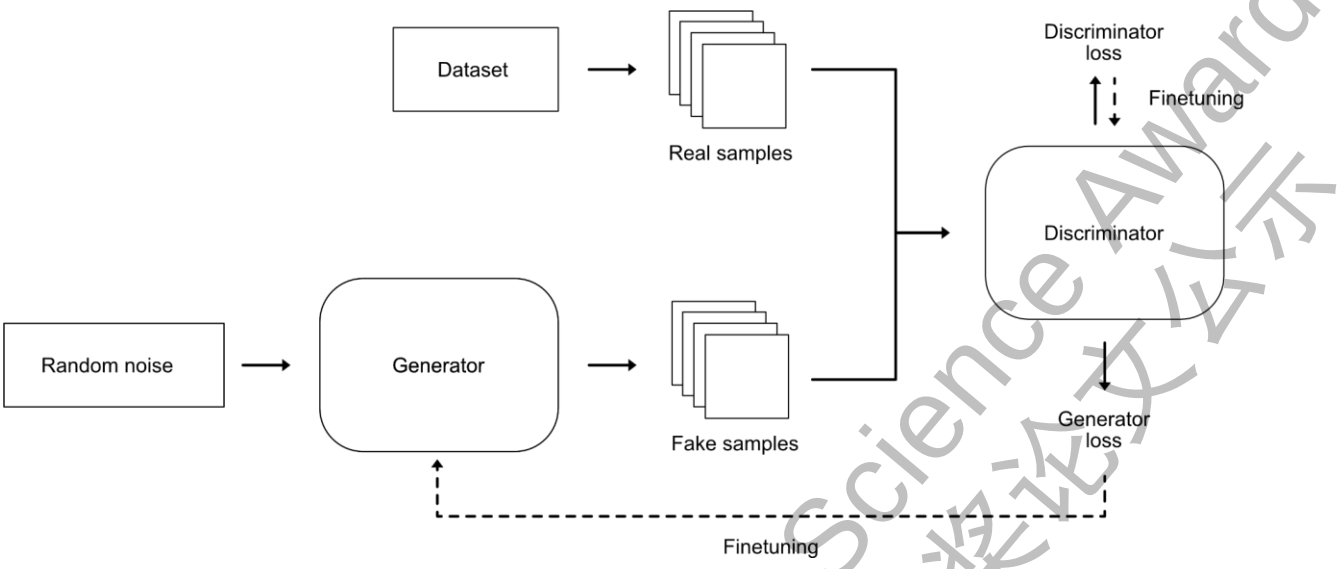

Figure 6 *Diagram of a Generative Adversarial Network*

This competitive interaction seeks to converge and reach the **Nash equilibrium**, a theoretical concept in game theory in which both players play perfectly and reach a state where changes in strategy would lead to inefficiencies. This results in the generation of high-quality, realistic samples as both networks progressively improve their respective abilities.

$$\mathbb{E}_{x\sim p_{\text{data}}}[\log D^*(G(x))] + \mathbb{E}_{x\sim p_g}[\log(1 - D^*(G(x)))]$$

Formula 3 *Minimax objective function for optimal discriminator*

As for our main purpose of data imputation, we can make a reference to the framework utilised within Generative Adversarial Imputation Nets (GAIN) [8], which employs a random mask to simulate missing data. However, instead of utilising a random mask, we propose using a static mask that specifically represents the wavelength range of the instrument used for the spectra to be retrieved, in our case the James Webb Space Telescope (JWST). This approach allows for a more focused application of our model. Additionally, it is theoretically possible to implement a different mask for retrieving data from other instruments, such as the Hubble Space Telescope's Wide Field Camera 3 (HST WFC3).

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

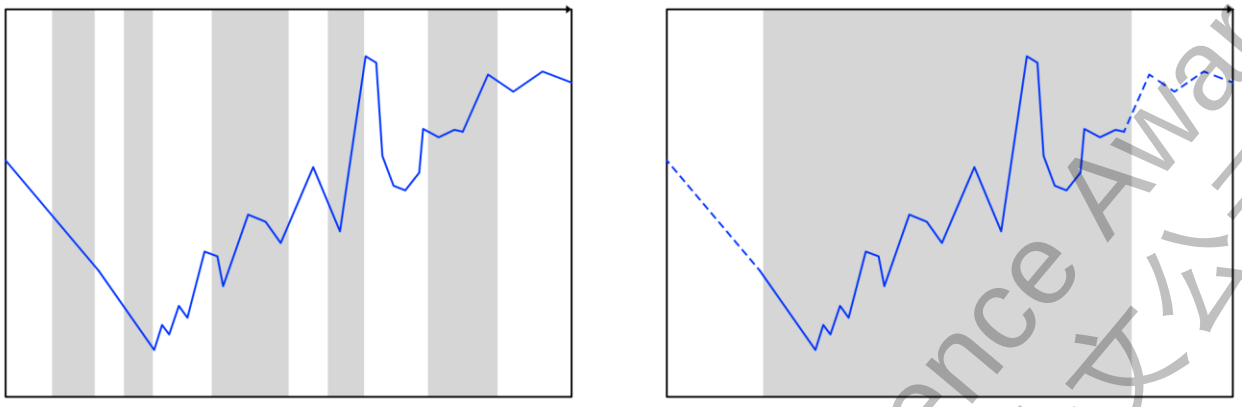

Figure 7 *Mask for spectral data (left: random mask, right: specialised mask)*

I first tried out this model with a **DCGAN,** one of the most commonly used GAN architecture-types, standing out as a simple yet effective choice. DCGAN is a **convolutional neural network**, which differs from a vanilla GAN in that it specialises more in visuals. Specifically, a convolutional neural network makes use of a convolution filter or a multi-dimensional vector representation to extract specific features such as edges or textures from the input data. Although the spectral data is 1 dimensional, it can be encoded using a 2D image such that we can fully exploit the feature extracting prowess of the DCGAN.

To preprocess the data, I first normalised the data and converted it into a pixel grid. The difference can be observed below.

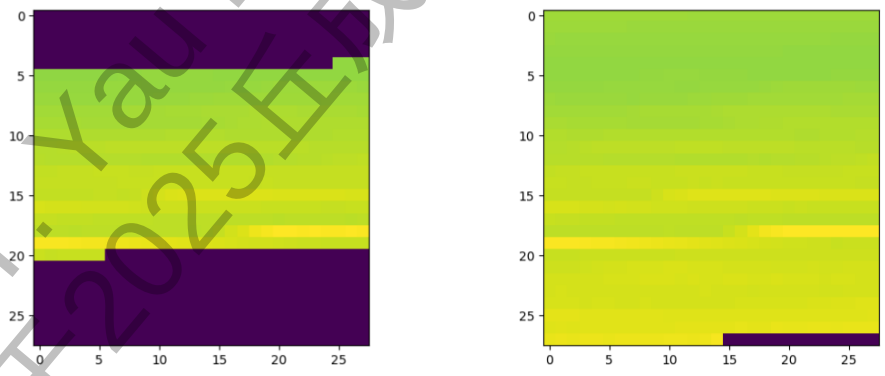

Figure 8 *Spectrum converted into image for GAN (left: masked, right: unmasked)*

I then trained the DCGAN to recreate the image on the right given the image on the left by adding the initial input into the output, masking the generated spectrum for 20 epochs. I obtained the following terrible looking result on a benchmark spectrum:

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

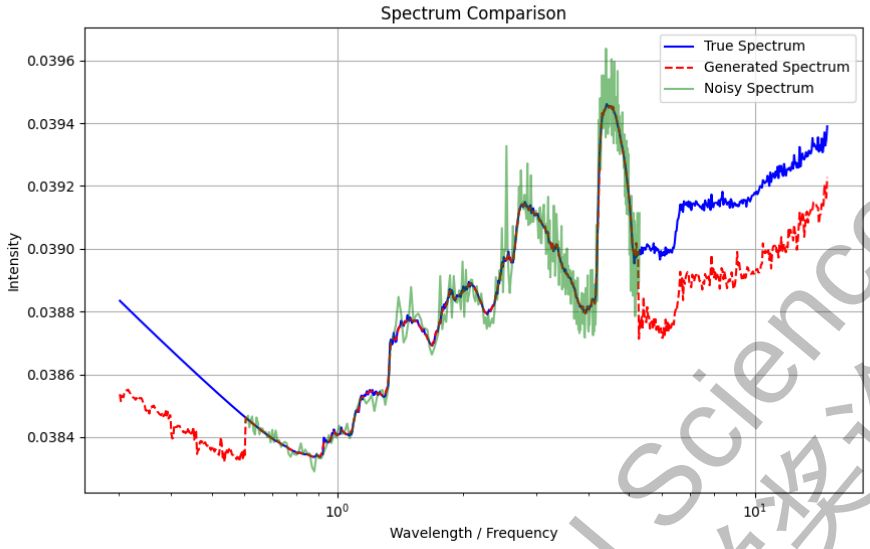

Figure 9 *GAN result after 20 epochs*

From this result, it was apparent that the GAN reconstruction was slightly offset from that of the real spectra, leading to this unfortunate result. To fix this, I attempted to add a baseline calibration unit to the generator, which has trainable offset and scale parameters to remedy this issue. However, the end result got worse, as it appeared the GAN had likely experienced **mode collapse**, where it found a certain number of suitable modes of data distribution rather than capturing the full diversity of the target distribution for the baseline calibration.

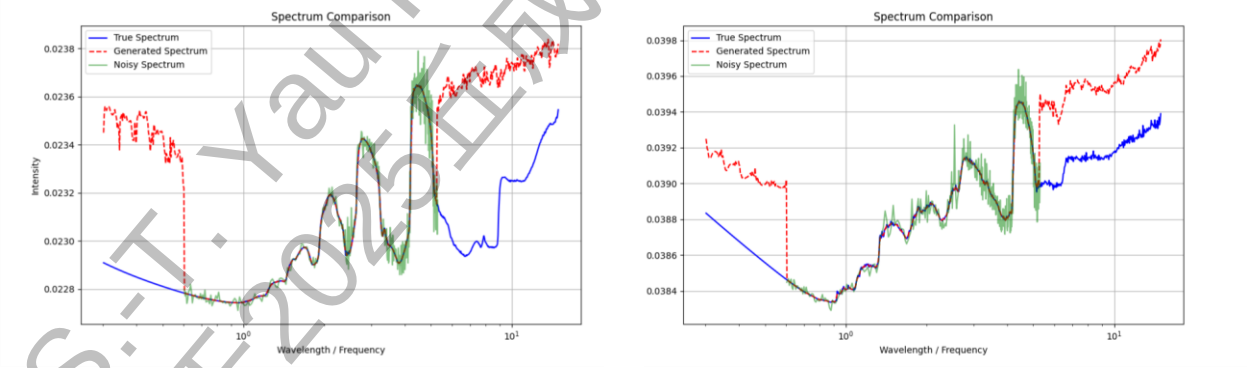

Figure 10 *GAN results with baseline calibration*

As a result, I decided that it was likely a problem with the architecture itself, and decided to experiment with other architectures.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

U-Net is another particularly promising architecture due to its unique design and capabilities, which can be effectively leveraged in the context of generative tasks. Its distinctive feature is the U-shaped structure, which consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. This architecture is particularly advantageous for tasks involving image-to-image translation, as it effectively combines low-level features from earlier layers with high-level features from later layers. In the context of spectra generation, U-Net's ability to retain spatial information while generating outputs makes it a strong candidate for producing high-fidelity spectral data. Its skip connections facilitate the transfer of detailed information, resulting in more accurate reconstructions of spectra.

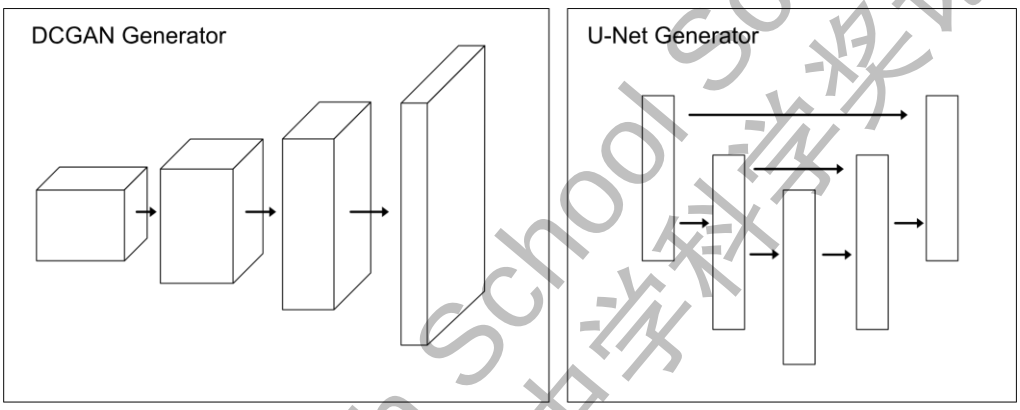

Figure 8 *DCGAN (left) vs U-Net (right) architecture*

I also decided to try out an **Attention**-based generator, which utilises an attention block, allowing the model to focus on specific parts of the input data when making predictions. This means that the neural network may potentially be able to focus on the specific positions where the mask transition occurs (i.e. the edges of the input), potentially allowing it to gain an edge over the other models. I decided to utilise comparison metrics of MSE, MAE and a custom metric of spectral continuity to prevent the sharp dropoff that occurs at the mask boundaries.

$$L_{\text{cont}} = \frac{1}{N-1} \sum_{i=1}^{N-1} |g_{i+1} - g_i|$$

Formula 4 *Spectral continuity metric*

However, this ended up being obsolete as shown below.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

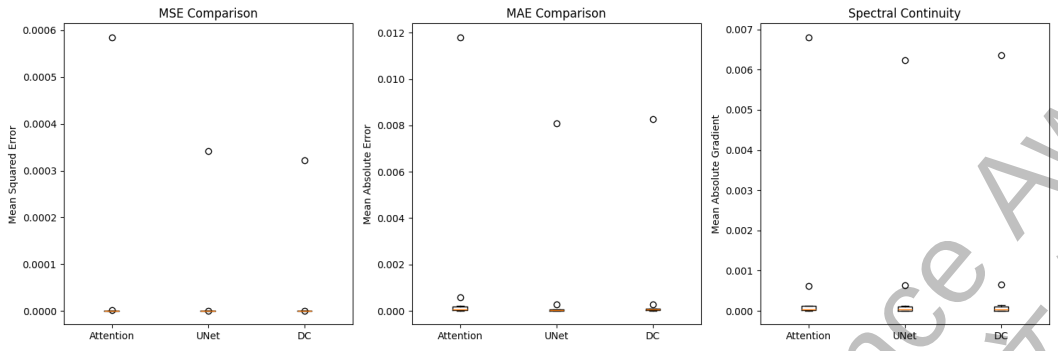

Figure 9 *Model comparisons*

Although not easily observed due to the incredibly high outliers, UNet comes out on top in terms of both the median, where it narrowly beats the Attention and DCGAN in MAE and spectral continuity and loses to Attention in MSE by a small margin.

As a result, I decided to conduct a hyperparameter search on the UNet, obtaining an optimal learning rate of 0.00005, ReLU activation, 2 residual blocks and instance normalisation. Additionally, in an attempt to fix the issue of the misaligned reconstructions, rather than simply using a combination of the F1 loss and adversarial loss, I chose to include the spectral continuity metric in addition to a new physics-based loss:

$$g_{\text{loss}} = g_{\text{loss}}^{\text{adv}} + g_{\text{loss}}^{\text{L1}} + g_{\text{loss}}^{\text{physics}} + g_{\text{loss}}^{\text{grad}}$$

where:

$$g_{\text{loss}}^{\text{adv}} = \mathcal{L}_{\text{adv}}(f_{\text{pred}}, r_{\text{label}})$$

$$g_{\text{loss}}^{\text{L1}} = \lambda_{\text{L1}} \cdot \| f_{\text{spectra}} \cdot -r \cdot \|_1$$

$$g_{\text{loss}}^{\text{physics}} = \lambda_{\text{physics}} \cdot (\text{shape\_loss} + \text{amp\_loss})$$

$$\text{shape\_loss} = \frac{1}{N} \sum_{i=1}^{N} (\nabla g_i - \nabla r_i)^2$$

$$\text{amp\_loss} = \frac{1}{N} \sum_{i=1}^{N} (g_i \cdot -r_i \cdot)^2$$

$$g_{\text{loss}}^{\text{grad}} = \lambda_{\text{grad}} \cdot \| \nabla f_{\text{spectra}} - \nabla r \|_1$$

Formula 5 *GAN Loss*

This led me to obtain the following results:

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

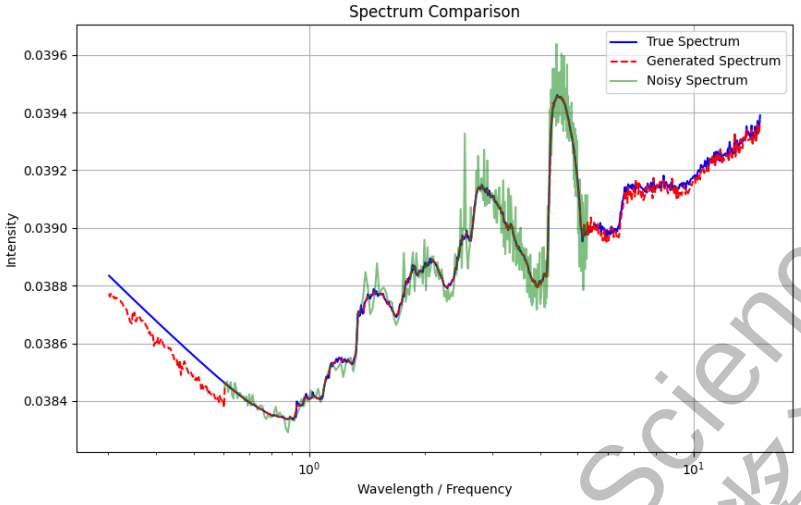

Figure 10 GAN w/ loss + modifications

To enhance the reconstruction quality, I opted to implement pretraining for the GAN. Specifically, I pre-trained the model for 10 epochs to ensure that the discriminator does not become overly dominant, which could hinder the generation of accurate losses.

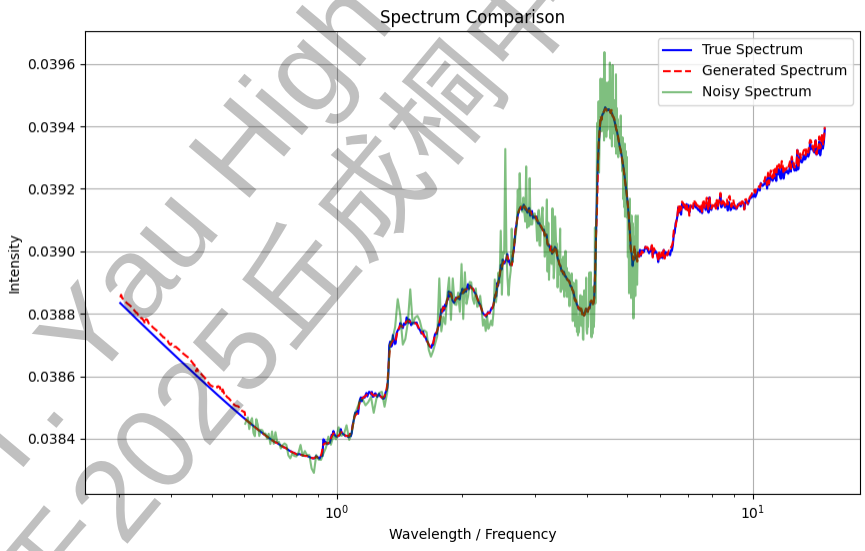

Figure 11 *Final GAN output*

With these new modifications, I was able to achieve a **98.5%** reconstruction accuracy compared to an initial 85.4% from the start of the experimentation.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

3.3.3 Validation

To benchmark this entire process, I decided to take the trained GAN and VAEs and combine them into the pipeline, before using 1000 spectra to test their reconstruction loss. I was able to obtain a spectacular **99.05%** average reconstruction accuracy according to Formula 2 (1 - RMSE). An example of the retrieved spectra can be found below.

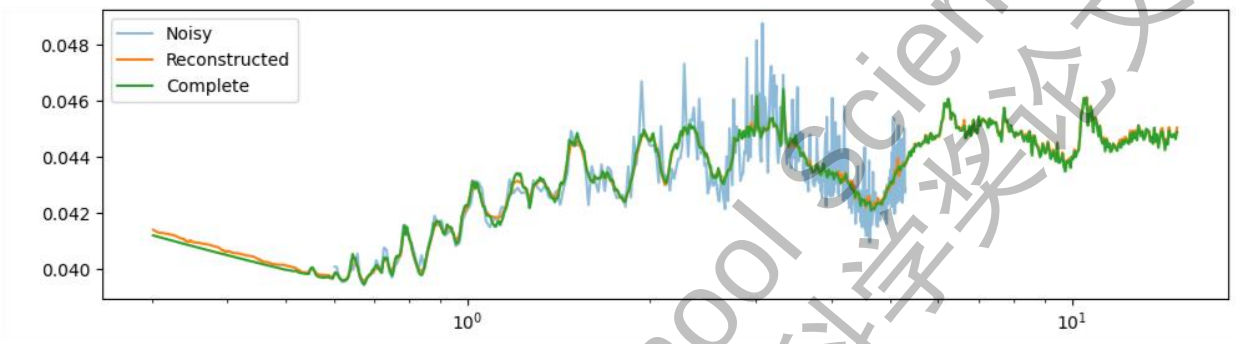

Figure 12 *Example of generative pipeline output*

3.4 Parameter Retrieval

The key difficulty in performing parameter retrievals for spectral data lies in the **high dimensionality** of parameter spaces and **complex distributions** that can arise from the complex models used to generate them. The ultimate goal of a parameter retrieval is to identify the most likely parameter values given some data. However, in the case of complex distributions, multiple modes may exist, making a single prediction insufficient. Such an oversimplification fails to capture the full complexity of the underlying data.

As a result, the underlying distribution must be obtained, and this can be solved using the Bayesian paradigm. Bayesian inference is a statistical approach that incorporates prior knowledge or beliefs about parameters along with observed data to update our beliefs in light of new evidence. It is grounded in Bayes' theorem, which provides a mathematical framework for this updating process.

$$P(\theta|\mathbf{D}) = \frac{P(\theta)P(\mathbf{D}|\theta)}{P(\mathbf{D})}$$

Formula 6 *Bayes' Theorem*

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

Bayes' theorem states that the posterior probability $P(\theta|D)$ of the parameters given data D can be represented in the form of likelihood $P(D|\theta)$, the probability of observing the data given the parameters, multiplied by the prior probability $P(\theta)$, reflecting our prior beliefs about the parameters all divided by the marginal likelihood or evidence $P(D)$ which normalises the posterior. This provides a concrete mathematical framework that can be used to obtain the underlying distribution given that $P(D|\theta)$ can be calculated.

However, in the case of complex forward models such as those present in atmospheric spectra simulations, the likelihood $P(D|\theta)$ is intractable, meaning it cannot be computed analytically. The introduction of **simulation-based inference** methods address these challenges fully, allowing for the approximation of the posterior distribution without needing to directly compute the likelihood, hence bypassing the main issue.

3.4.1 Neural Posterior Estimation

Whereas traditional simulation-based inference methods such as Monte Carlo Markov Chain estimation utilise nested sampling, where they sample the posterior repeatedly to obtain the approximate underlying distribution, recent developments in simulation-based inference methods have allowed for deep learning to generate large computational efficiencies.

An important concept in simulation based inference is the idea of a **normalising flow** [13], which can be used to generate the complex posteriors required. The concept can be formalised as follows: Let z be a random variable with a simple base distribution $p(z)$. A normalising flow transforms z into a more complex random variable x through a series of K invertible and differentiable transformations f_K.

$$x = f_K \circ f_{K-1} \circ \cdots \circ f_1(z)$$

Formula 7 *Normalising Flow*

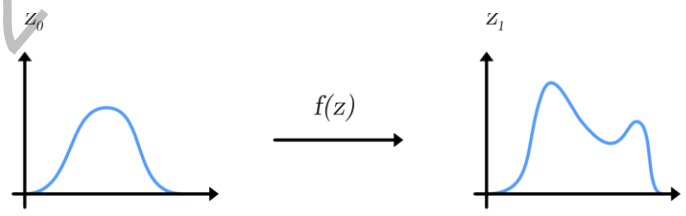

Figure 13 *Example of normalising flow*

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Each transformation f_K gradually increases the complexity of the distribution, and the probability density of the transformed variable $p_x(x)$ can be computed using the change of variables formula.

$$p_x(x) = p_z(f_1^{-1} \circ \cdots \circ f_K^{-1}(x)) \left| \det \frac{\partial f_1^{-1} \circ \cdots \circ f_K^{-1}(x)}{\partial x} \right|$$

Formula 8 *Change of variables formula*

The crucial properties of the normalising flow are invertibility, which allows for efficient sampling from $p_x(x)$ by first sampling from the base distribution $p_z(z)$ and then applying the inverse transformations f_k^{-1}. It also allows for efficient evaluation of the density $p_x(x)$ by computing the Jacobian determinant of the inverse transformations.

However, normalising flows can be computationally expensive. An alternative, known as neural posterior estimation [14], learns a conditional normalising flow that directly maps from observed data to posterior distributions. This amortisation allows for rapid inference on new observations once the model is trained, significantly reducing computational costs for multiple retrievals.

Neural posterior estimation is also particularly well-suited for simulator based models, where the likelihood is intractable but sampling from the forward model is possible. This makes it highly applicable to atmospheric retrieval problems where complex physical models are involved. As for training, the KL divergence between the true posterior $p(x|z)$ and estimation $q(x|z)$ produced by the normalising flow is minimised, or

$$KL(p||q) = \int p(x|z) \log \frac{q(x|z)}{p(x|z)} d\theta$$

Formula 9 *KL Divergence formula*

To efficiently optimise the KL divergence during training, the loss function can be rewritten using Bayes' theorem as

$$L = \mathbb{E}_{x \sim p(x), z \sim p(z|x)} [\log q(x|z) - \log p(x)]$$

Formula 10 *Optimised formula for training*

Here, the training data is generated first by sampling parameters x from the prior p(x), then simulating the data z from the likelihood p(x|z). This allows the neural network to learn a flexible, normalising flow-

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

based representation of the posterior without requiring explicit likelihood evaluations. However, neural posterior estimation is an **amortised** estimator, meaning it is written to be generalised to **many observations** and requires a **large dataset**. This problem can be solved using an alternative approach.

3.4.2 Sequential and Truncated Methods

Whereas neural posterior estimation models are trained with a fixed set of simulations drawn from the prior and likelihood and can generate posterior samples for any observation x after training, **sequential neural posterior estimation** [15] proceeds in multiple rounds, with each round drawing parameters from the proposal distribution, generating simulations and updating the estimator using the data collected at each point. The proposal distribution is also refined to focus on more relevant regions of the parameter space, meaning it is more tailored to retrieve posterior samples for the observation of interest rather than any new observation outside of the relevant parameter space. The differences can be observed in the diagram below.

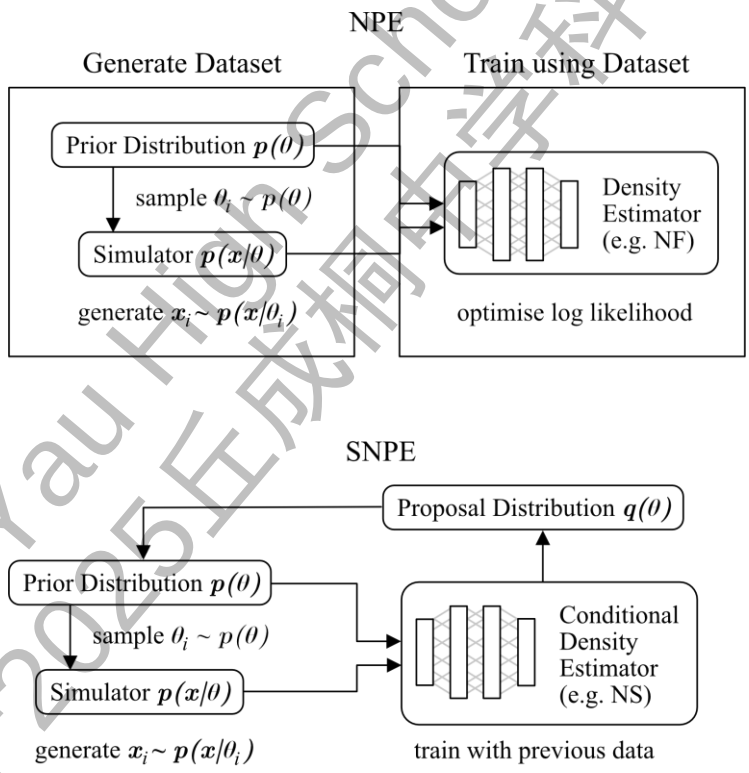

Figure 14 *NPE vs SNPE training process*

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

This singular change means that a large and computationally costly dataset is no longer needed, and this **non-amortised** approach also allows for greater accuracy around the particular observation and less samples needed.

A new truncated alteration of the SNPE algorithm has also proven to be an attractive alternative. Known as Truncated Sequential Neural Posterior Estimation or TSNPE [16], it aims to overcome some of the limitations of previous SNPE methods while maintaining their efficiency.

The key modification lies within the use of truncated proposal distributions. Rather than sampling from the full prior or previous posterior estimate, TSNPE samples from a truncated version of the prior. This is carried out through the definition of a proposal distribution proportional to the prior, but only within a certain region called the Highest Probability Region (HPR), defined as the smallest region containing a certain fraction $(1-\varepsilon)$ of the mass of the current posterior estimate. As a result, the parameters sampled from the prior are only accepted if they lie within the HPR of the current posterior estimate, creating a truncated proposal distribution proportional to the prior within the HPR. This means that unlike other SNPE methods requiring complex loss modifications, TSNPE allows training the neural density estimator with a simple maximum likelihood objective in all rounds.

Figure 15 *Illustration of truncated prior*

$$L = \mathbb{E}_{\theta \sim p(\theta)} \mathbb{E}_{x \sim p(x|\theta)} [-\log q(\theta|x)]$$

Formula 11 *Log-likelihood loss function for TSNPE*

Another advantage of TSNPE is that it speeds up computation time, as it leads to prior samples outside of the approximated posterior being rejected, leading to decreased sample sizes and greater efficiency.

In addition to TSNPE, I decided to explore a score-based method pioneered by Sharrock et al. [3], which leverages score-based diffusion models to estimate the posterior's score rather than sampling directly from

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

the posterior estimate. This innovative approach transforming samples from the target posterior into a tractable reference distribution has demonstrated significant promise, particularly in high-dimensional retrieval tasks, suggesting that it may offer enhanced performance compared to traditional sampling methods. To investigate the efficacy of these advanced methods, I will conduct a series of experiments comparing the performance of TSNPE and the score-based approach against traditional SNPE techniques. The results of these experiments will provide insights into their relative strengths and applicability in various scenarios.

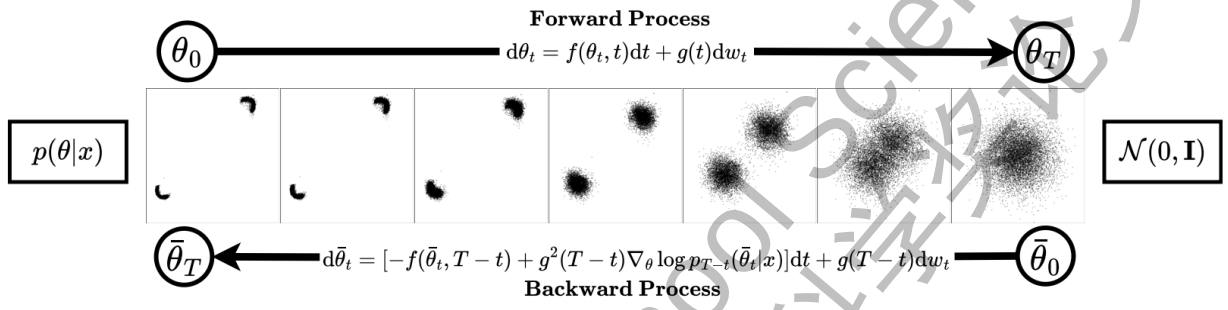

Figure 16 *Diagram representing SNPSE process (obtained from L. Sharrock et al [3])*

$$\mathcal{J}_{\text{post}}^{\text{TSNPSE-SM}}(\psi) = \frac{1}{2}\int_0^T \lambda_t \mathbb{E}_{\tilde{p}_t^r(\theta_t, x)} \qquad (10)$$
$$[||s_\psi(\theta_t, x, t) - \nabla_\theta \log p_t(\theta_t|x)||^2]\mathrm{d}t,$$

Figure 17 *SNPSE Loss function (obtained from L. Sharrock et al [3])*

3.4.3 Model Selection

To test these models, I utilised the **sbi** package [26] to implement the algorithms of TSNPE, SNPE and TSNPSE. In order to ensure that the neural posterior estimation models were sufficiently apt in their predictions, I utilised a custom neural spline flow based density estimator, with the implementation architecture detailed below detailed below:

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Layer	Type	Output Shape	Parameters
CNN Embedding Network			
Input	–	$(B, 1, 771)$	0
Conv1D-1	Conv1D	$(B, 6, 767)$	30
MaxPool1D-1	MaxPool1D	$(B, 6, 383)$	0
Conv1D-2	Conv1D	$(B, 12, 379)$	360
MaxPool1D-2	MaxPool1D	$(B, 12, 189)$	0
Flatten	Reshape	$(B, 2, 268)$	0
Linear-1	Linear	$(B, 64)$	$145,216$
Linear-2	Linear	$(B, 64)$	$4,160$
Neural Spline Flow			
Input	–	(B, D)	0
NSF Block-1	NSF	(B, D)	$2D \cdot 50 + 50^2 \cdot 10$
NSF Block-2	NSF	(B, D)	$2D \cdot 50 + 50^2 \cdot 10$
15 transforms per block, 10 bins per transform			

where:

- B is the batch size
- D is the dimensionality of the target distribution

Table 5 *Density estimator with neural spline flows: architecture*

To carry out the trials, I first generated a simple spectrum with a limited number of parameters compared to the simulation dataset. The ground truth can be found below.

Parameter	**Ground Truth**
Planet Radius (R_J)	0.742
Planet Log g	4.028
Temperature (K)	1778
log H_2O	-4.305
log CO_2	-7.894
log CH_4	-3.274
log NH_3	-8.103
log CO	-5.115

Table 6 *Ground truths for simulated observation*

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

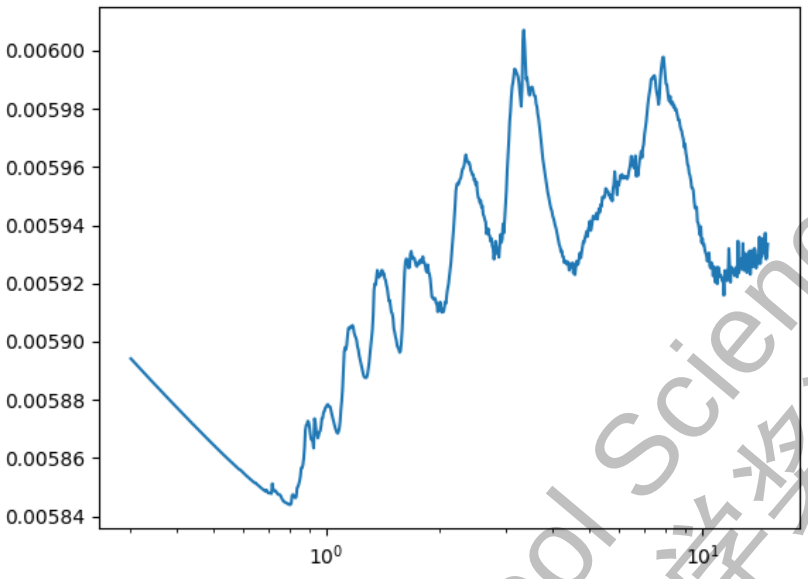

Figure 18 *Simulated spectrum with above parameters*

Ultimately, I obtained incredibly complex posteriors for both TSNPE and SNPE. However, the results obtained for TSNPSE were completely different, as it returned comparatively Gaussian-like posteriors, likely a result of its score-based diffusion model. A comparison between the results generated by corner [27] obtained by each model can be found below, where the predictions are located above the histograms with the numbers beside them representing +1σ and -1σ values. The true values are also shown by the blue lines.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

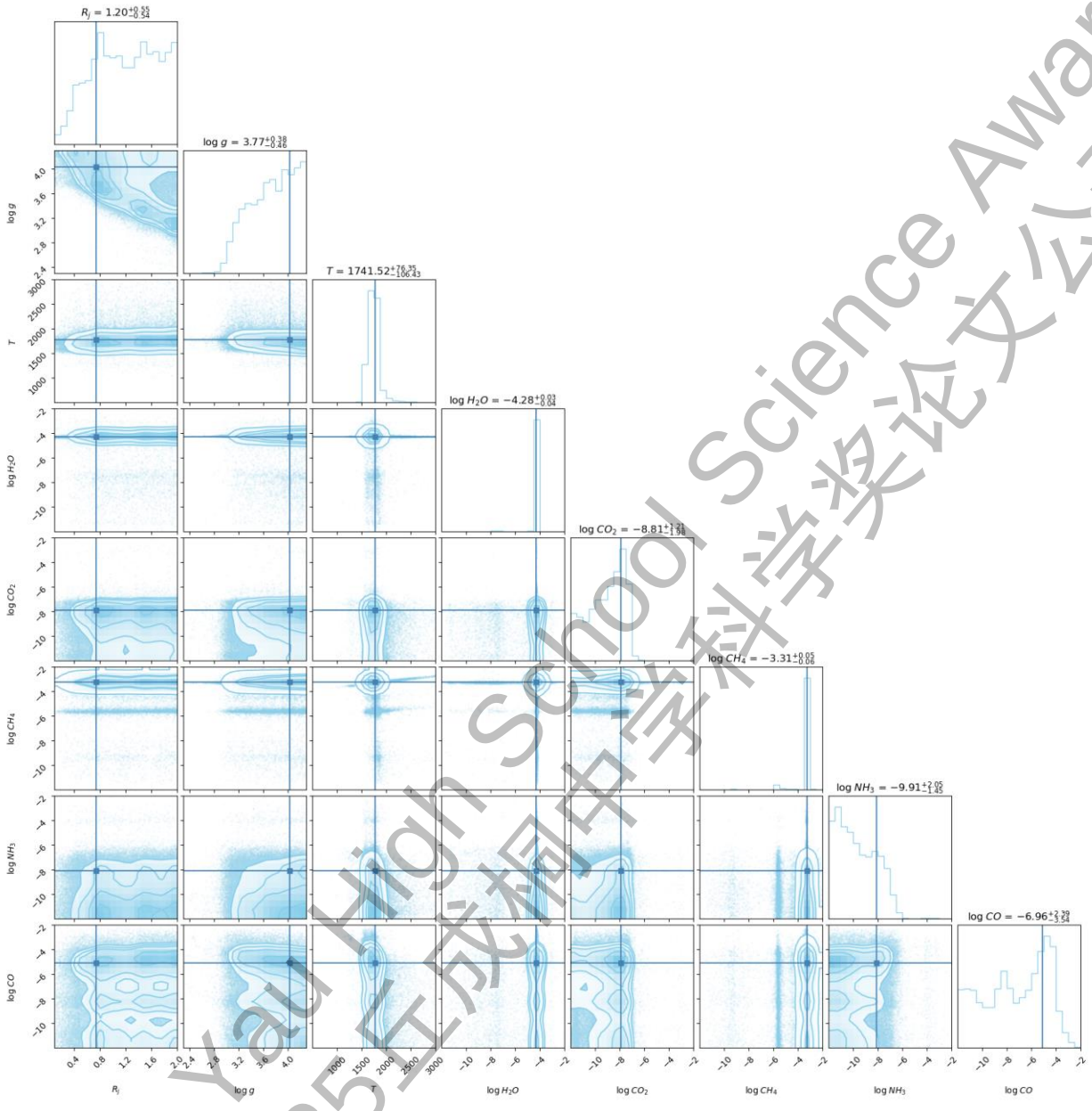

Figure 18 *Posteriors obtained using SNPE*

These posteriors above from SNPE demonstrate its capability to produce complex, multimodal posteriors, as can be observed in the contour plots in the lower section. The simulation-based inference methods also allow for uncertainty estimation, as can be seen by the +1σ and -1σ values demonstrating greater confidence in the parameters which are predicted with greater accuracy, and lower confidence in the parameters predicted with lower accuracy.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

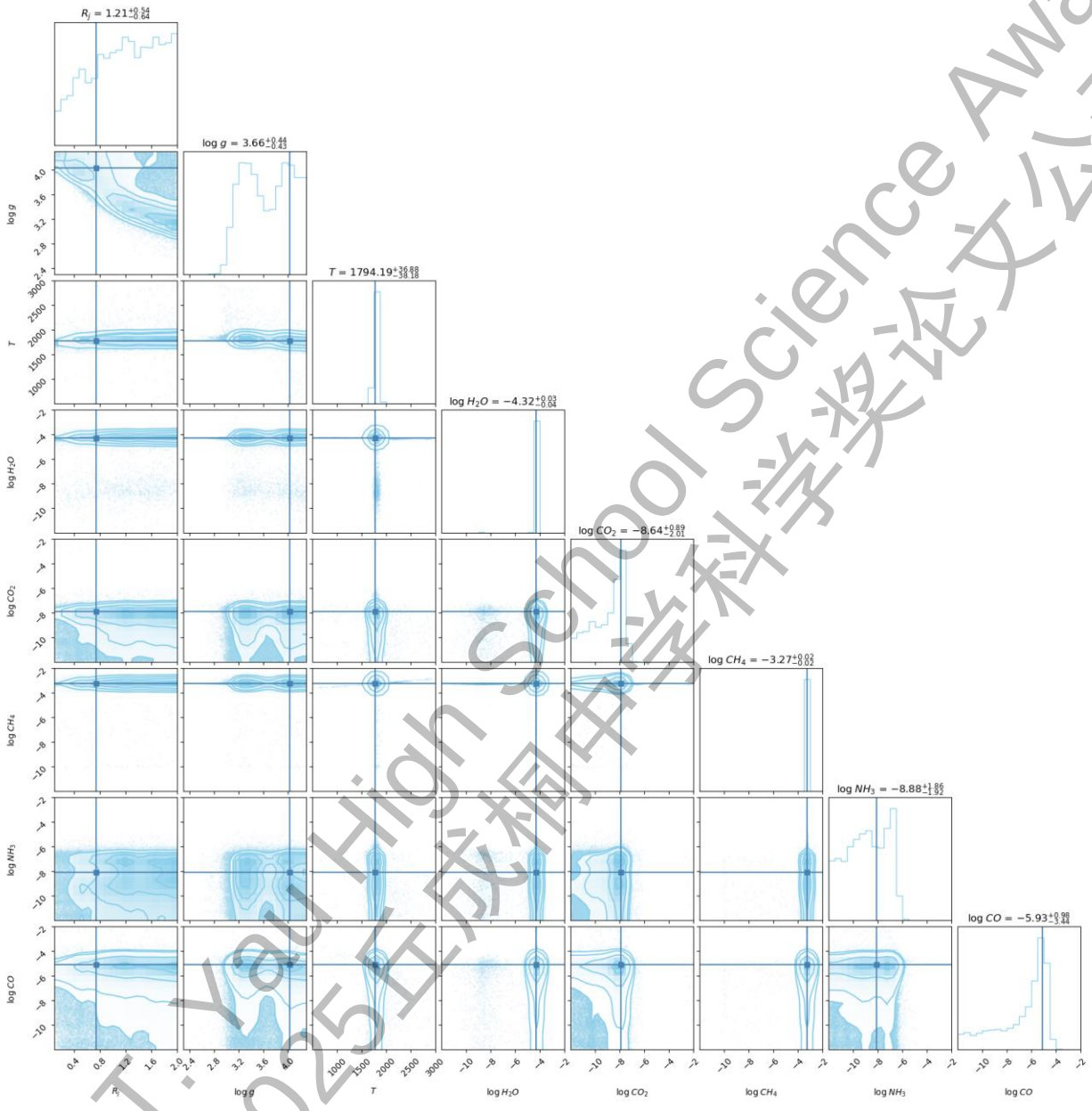

Figure 19 *Posteriors obtained using TSNPE*

The TSNPE process also demonstrates the same capabilities, due to the fact that it utilises the same base algorithm to predict the spectra. However, its use of truncated proposals means it had almost half the training time, which makes it impressive that it is able to produce these posteriors as well. It also appears to have higher confidence in general compared to the SNPE results from earlier.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

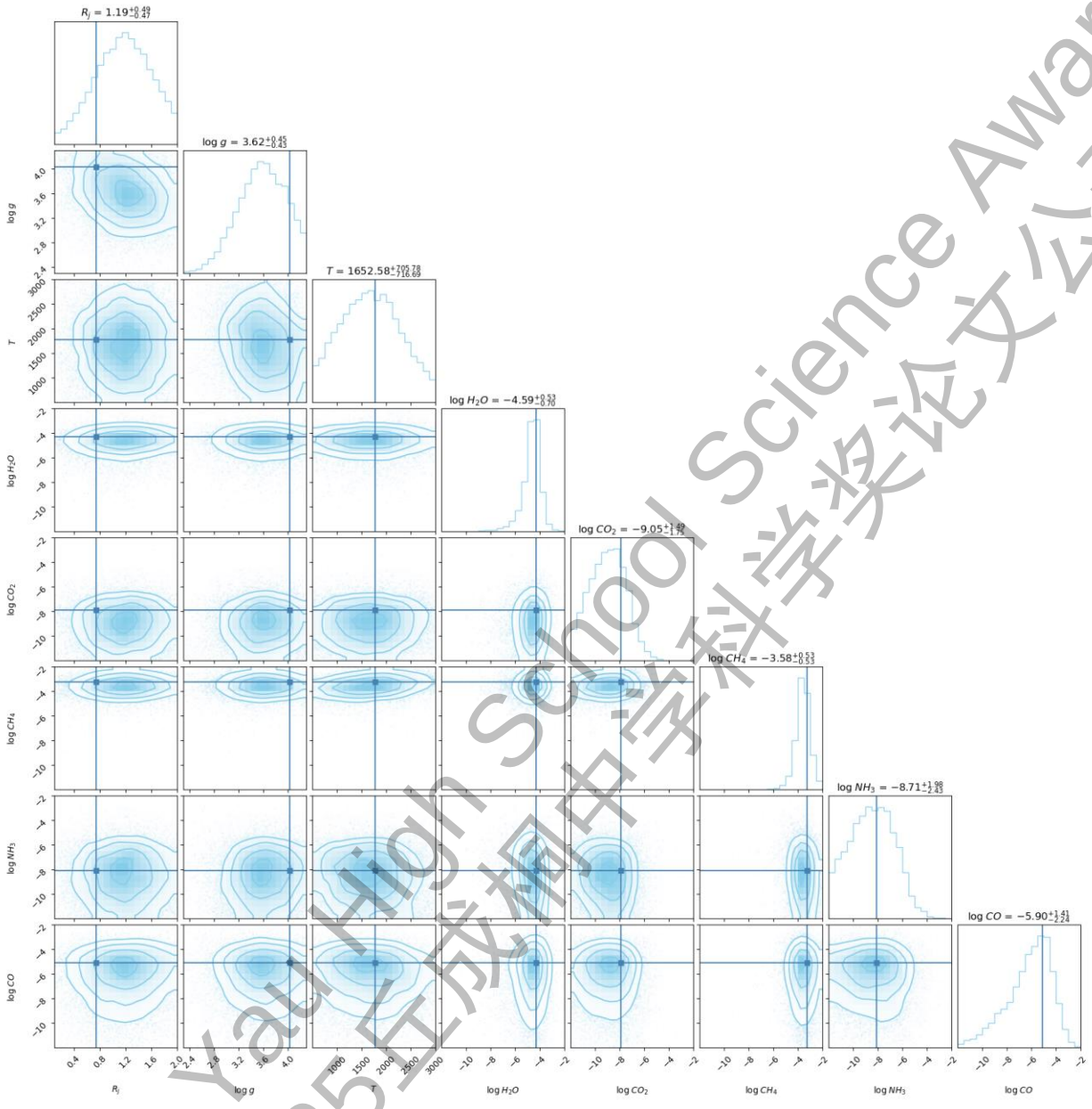

Figure 20 *Posteriors obtained using TSNPSE*

Finally, the TSNPSE returned incredibly clean posteriors that lack any sign of multimodality. This is likely due to its score-based nature and transformation of samples from the target posterior to a tractable reference distribution, leading to a cleaner, simpler distribution. Compared to the other models, this model exhibits a significant capability in predicting the parameter values with greater accuracy, although its wide uncertainties are a large downside.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

	TSNPE	SNPE	TSNPSE
Total Training Time	**5641**	8180	5704
Training time	3147	6213	**2678**
Simulation time	2494	**1967**	3026
1 - MAPE	**88.9%**	87.0%	86.1%

MAPE Formula:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \times 100$$

where A_i is the actual value and F_i is the forecasted value.

Table 7 *Accuracy and Training Time across Models + Mean Absolute Percentage Error Formula*

As can be observed in the table above, TSNPE generally performed the best with an accuracy of 88.7%, whilst SNPE obtained an accuracy of 87.0% and TSNPSE was worse by a decent margin at an accuracy of about 86.1%. TSNPE not only has more accurate parameter predictions, but also has a lower training time, beating SNPE by 31% in computation time.

However, it appeared that SNPE performed the best during simulations, as it doesn't require any truncated proposals which take quite long to sample from in order to generate the spectra. TSNPSE also likely has a more straightforward training approach, leading to a lower training time.

Despite its lack of confidence in its posteriors though, TSNPSE does indeed demonstrate the most accurate posteriors for the mix ratios of the parameters, as can be observed by the clear fact that the modes of the histogram occur at the specific bins that contain the parameter values.

This demonstrates that TSNPSE is indeed a promising candidate for potential future improvements, when we are able to make more modifications to the model and perform a hyperparameter sweep. However, in the meantime, it is clear that TSNPE is the optimal model, and we will perform evaluations with this model in the proceeding sections.

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

4 Evaluation

To demonstrate the performance and viability of the end-to-end FlexARE framework, we conducted a comprehensive case study on a complex, 15-parameter simulated observation. The raw, noisy spectrum was first processed by our generative module to produce a clean, high-fidelity input for the inference engine, which took approximately 2 seconds. The ground truth parameters and final reconstruction can be seen below:

Parameter	Ground Truth
$R_P(R_J)$	0.645
$\log g$	3.510
$T_{irr}(K)$	2305
κ_1	-1.534
κ_2	-2.088
κ_{ir}	-3.111
α	0.101
$\log H_2O$	-5.106
$\log CO_2$	-8.936
$\log CH_4$	-3.945
$\log NH_3$	-8.359
$\log CO$	-2.925
$\log P_{cloud}$	3.992
$\log mie_R^{lee}$	0.816
$\log mie_Q^{lee}$	-2.214
$\log mie_X^{lee}$	-34.04

Table 8 *Ground truth for final spectrum*

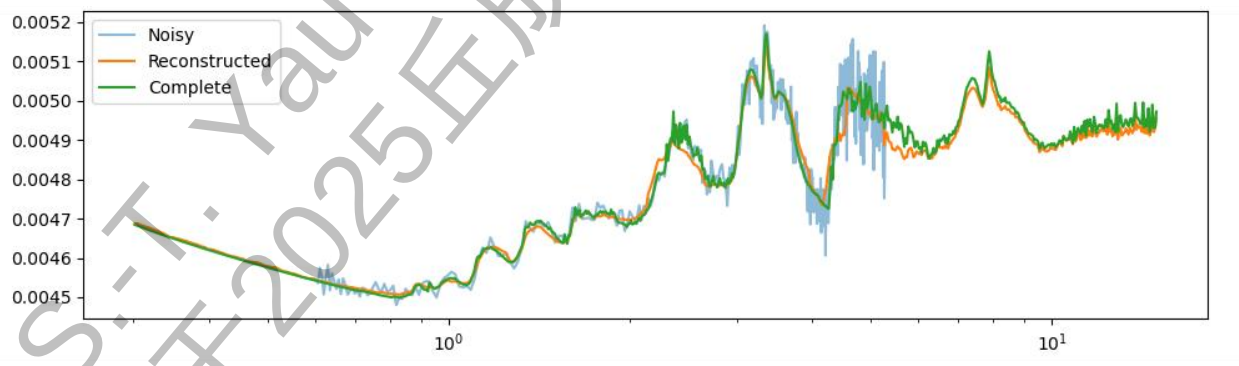

Figure 21 *Final reconstructed spectrum*

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

4.1 Retrieval Accuracy and Posterior Quality

To retrieve the spectra, the preprocessed spectra was put into the SBI backend, where the TSNPE model was then run for 10 rounds with 1500 simulations per round to retrieve the atmospheric parameters. The entire process, from noisy spectrum to final posterior, serves as a proof-of-concept for the full pipeline.

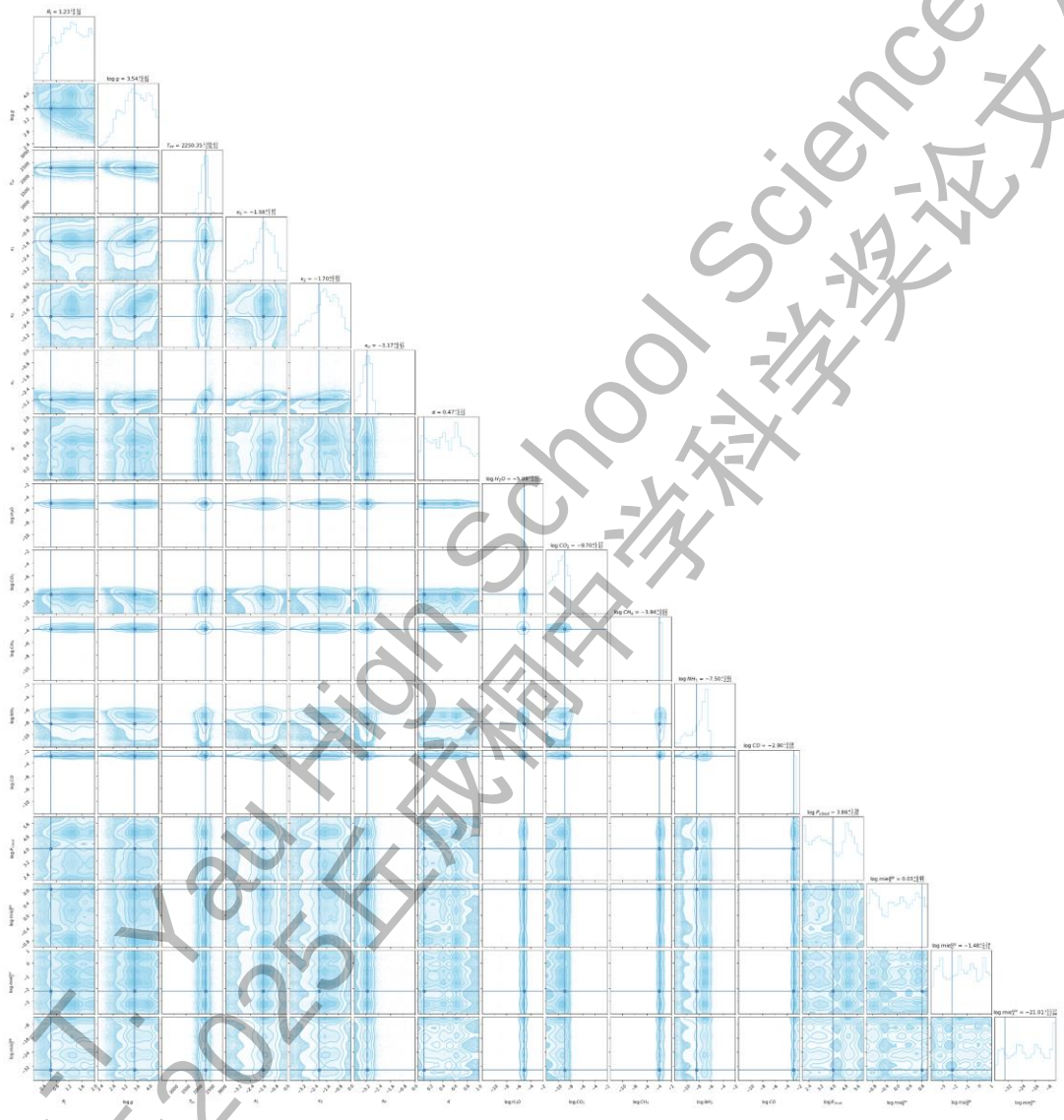

Figure 22 *Final TSNPE posteriors for pre-processed spectrum*

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Ground Truth	Prediction	Accuracy (%)
0.645	1.23	30.79
3.510	3.54	1.50
2305	2250.35	1.44
-1.534	-1.58	1.15
-2.088	-1.70	9.70
-3.111	-3.17	1.48
0.101	0.47	36.90
-5.106	-5.09	0.16
-8.936	-9.70	7.64
-3.945	-3.94	0.05
-8.359	-7.50	8.59
-2.925	-2.90	0.25
3.992	3.86	3.30
0.816	0.03	39.30
-2.214	-1.48	14.68
-34.04	-21.01	36.19

Table 9 *Final prediction results for preprocessed spectrum*

We obtain incredible results from the FlexARE retrieval for the key parameters (those determining the temperature and molecular composition). As shown in Table 9 and the posterior distributions in Figure 22, the framework successfully constrained the vast majority of the 15 atmospheric parameters. However, it is apparent that the prediction values for the lee-mie scattering and radius (which is quite unusual) is uncertain. However, during typical retrievals, this value is often already known. As a result, in the calculations for the accuracy, I chose not to include this value. The values which are not as accurate also have wider posteriors, indicating that the values are uncertain.

The retrieved values are well-centred on the ground-truth, leading to a final mean parameter accuracy of **89.2%**. This high accuracy on a complex, high-dimensional problem validates the design of both the generative pre-processor and the SBI back-end. The posteriors are well-behaved and provide meaningful uncertainty quantification for the retrieved parameters

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

5 Conclusion and Future Works

In this work, we have designed, implemented, and validated FlexARE, a novel, two-stage framework for exoplanet atmospheric retrieval that decouples data conditioning from inference. Through a comprehensive case study on a complex 15-parameter problem, we have demonstrated that our end-to-end pipeline is both accurate and computationally efficient. The framework achieved a high parameter retrieval accuracy of 89.2% while being several orders of magnitude faster than classical gold-standard methods like MultiNest. This result serves as a successful proof-of-concept, showing that our modular architecture is a viable and powerful solution for tackling the challenges of next-generation exoplanet spectroscopy and offers many advantages.

First, FlexARE is **accurate**. FlexARE's preprocessing module enables the removal of noise from obtained spectra and the imputation of missing data, allowing for re-correction of inaccurate data to an accuracy of **99.05%**. Additionally, the usage of posterior estimation methods leads to the ability to obtain uncertainty estimates, allowing for better understanding and evaluation of the parameters that have led to the creation of the spectra.

Second, FlexARE is **fast.** FlexARE only took ~500 CPU hours in total, with ~150 used for dataset generation, ~130 for training the autoencoder, ~150 for training the GAN and ~70 for the training the sequential model. The generation of the dataset can be easily parallelised on CPU clusters (as I have done here) and model training can be easily optimised on GPUs. Although I was originally going to use MultiNest to obtain a proper benchmark for the spectra, it ultimately took too long, as even after 48 hours (or ~1800 CPU hours) of sampling (with a reduced number of parameters i.e. 12 parameters) the model still hadn't converged. As a result, (due to the impending deadline) I have chosen to make an estimation instead. As shown in M. Vasist *et al.* [7], when retrieving the same amount of parameters, MultiNest is said to have used about 60000 CPU hours to converge. However, it does indeed utilise a more complex forward model, taking ~5s on average compared to ~1s with our current forward model. Hence, we obtain a lower and upper bound of 1800 CPU hours and 12000 CPU hours. This means that this pipeline achieves a minimum speed up of 3.6 times or 360% over MultiNest given the lower bound, and possibly 24 times or 2400% given the upper bound. If the upfront training times for the VAE and GAN are removed, the speedup could be up to **~26-171 times faster**. This means that FlexARE may even be able to be performed on home electronics in reasonable time periods.

Third, FlexARE is **flexible**. Its **modular** structure and lack of dependencies between the units mean that one may easily be exchanged for another. For example, rather than TSNPSE, we could utilise an

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

amortised training method instead, using the noise processing module to preprocess the observation. This not only means a reduced dataset size due to not having to generate noisy copies of the data, but also leads to more accurate results. Additionally, multiple modules could be added, meaning the pipeline can allow for multiple noise profiles from multiple instruments simply by adding more autoencoders, or introducing a multi-view autoencoder that can handle all noise profiles without issue. This means that multiple observations from multiple instruments could potentially be combined into one. Additionally, each separate model can be trained and optimised individually to allow for greater efficiency.

Through these various benefits, FlexARE opens up the possibility of having and even combining more sources of data (such as from amateur astronomers, as the preprocessing module can allow for high signal to noise (SNR) ratios) in addition to allowing for more efficient exploration of exoplanetary atmospheres.

However, FlexARE does indeed still have its flaws. Most importantly, the method in which data, inputs and outputs are transferred between the different modules are important and may lead to extra hyperparameters involved and more complex modifications compared to single models. Additionally, the lack of dependencies between the modules means that loss cannot propagate through the entire model, meaning certain optimisations cannot be achieved. This issue could potentially be resolved through implementing learnable normalisation layers between the modules after each model is individually trained and optimised, allowing for the loss propagation and simplifying the different data normalisation processes that may be used within each model.

In terms of the explorations of different modules, it is clear that there are definitely ways to refine this model tremendously in the future. For example, we can experiment with adding in different autoencoder models for different noise profiles, and even incorporate a multi-view variational autoencoder [28] due to it allowing for learning joint representation of multiple modalities (in this case noise profiles) of data. As for the GAN, we can experiment with adding masks other than JWST such as for the Hubble Space Telescope spectra or even future ARIEL spectra wavelengths. However, a better option would be to simply implement the GAIN [8] algorithm, allowing for completion of even patchy spectra and the combination of multiple observations from multiple instruments. This may even allow amateur astronomers to submit their own data and have it be used for atmospheric retrievals if a web interface is created.

In the case of the retrieval module which utilises simulation-based inference methods, there also exist many methods that allow for non-amortised retrievals based on multiple observations, such as flow

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

matching posterior estimation (FMPE) [29] and Partially Factorised Neural Posterior Score Estimation (PF-NPSE) [30].They both stand out in their ability to perform inference on high-dimensionalities of data. Specifically, FMPE allows for unconstrained architectures, enabling exact density evaluation and fast training, having been shown to reduce training time and increase scalability. On the other hand, PF-NPSE [30] allows for the aggregation of an arbitrary number of observations at inference time whilst requiring a low number of simulator calls. This means it avoids the limitation of standard inference methods as well whilst remaining sample efficient. In addition to TSNPSE, the above models can be further explored and optimised.

Finally, a web interface in combination with a module can be developed to allow amateurs and scientists alike to utilise this framework conveniently, similar to the ones found at online in the exoplanet characterisation toolkit by the space telescope science institute (ExoCTK by STScI) [31].

Extra Notes

The code be uploaded onto the github here as soon as it has been tidied and written into a package. Further concrete results will also be found in an extra folder.

https://github.com/cookie-is-yummy/FlexARE

Bibliography

[1] Vincent, P. *et al.* (2008) 'Extracting and composing robust features with denoising autoencoders', *Proceedings of the 25th international conference on Machine learning - ICML '08*, pp. 1096–1103. doi:10.1145/1390156.1390294.

[2] Goodfellow, I.J. *et al.* (2014) *Generative Adversarial Networks*, *arXiv.org*. Available at: https://arxiv.org/abs/1406.2661 (Accessed: 04 December 2024).

[3] Sharrock, L. *et al.* (2024) *Sequential neural score estimation: Likelihood-free inference with conditional score based diffusion models*, *arXiv.org*. Available at: https://arxiv.org/abs/2210.04872 (Accessed: 04 December 2024).

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖公示

[4] Feroz, F., Hobson, M.P. and Bridges, M. (2008) *MultiNest: An efficient and robust Bayesian inference tool for cosmology and particle physics*, *arXiv.org*. Available at: https://arxiv.org/abs/0809.3437 (Accessed: 28 November 2024).

[5] Cranmer, K., Brehmer, J. and Louppe, G. (2020) 'The frontier of simulation-based inference', *Proceedings of the National Academy of Sciences*, 117(48), pp. 30055–30062. doi:10.1073/pnas.1912789117.

[6] Ardévol Martínez, F. *et al.* (2024) 'Floppity: Enabling self-consistent exoplanet atmospheric retrievals with machine learning', *Astronomy & Astrophysics*, 681. doi:10.1051/0004-6361/202348367.

[7] Vasist, M. *et al.* (2023) 'Neural posterior estimation for exoplanetary atmospheric retrieval', *Astronomy & Astrophysics*, 672. doi:10.1051/0004-6361/202245263.

[8] Yoon, J., Jordon, J. and van der Schaar, M. (2018) *Gain: Missing data imputation using generative adversarial nets*, *arXiv.org*. Available at: https://arxiv.org/abs/1806.02920 (Accessed: 28 November 2024).

[9] Brandt, J., Mattsson, K. and Hassellöv, M. (2021) 'Deep learning for reconstructing low-quality FTIR and Raman Spectra─a case study in microplastic analyses', *Analytical Chemistry*, 93(49), pp. 16360–16368. doi:10.1021/acs.analchem.1c02618.

[10] Guillot, T. (2010) 'On the radiative equilibrium of irradiated planetary atmospheres', *Astronomy and Astrophysics*, 520. doi:10.1051/0004-6361/200913396.

[11] Line, M.R. *et al.* (2012) 'Information content of Exoplanetary Transit Spectra: An initial look', *The Astrophysical Journal*, 749(1), p. 93. doi:10.1088/0004-637x/749/1/93.

[12] Lee, J.-M., Heng, K. and Irwin, P.G. (2013) 'Atmospheric retrieval analysis of the directly imaged exoplanet HR 8799B', *The Astrophysical Journal*, 778(2), p. 97. doi:10.1088/0004-637x/778/2/97.

[13] Rezende, D.J. and Mohamed, S. (2016) *Variational inference with normalizing flows*, *arXiv.org*. Available at: https://arxiv.org/abs/1505.05770 (Accessed: 30 November 2024).

[14] Papamakarios, G. and Murray, I. (2018) *Fast ε-free inference of simulation models with Bayesian conditional density estimation*, *arXiv.org*. Available at: https://arxiv.org/abs/1605.06376 (Accessed: 30 November 2024).

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

[15] Greenberg, D.S., Nonnenmacher, M. and Macke, J.H. (2019) *Automatic posterior transformation for likelihood-free inference, arXiv.org*. Available at: https://arxiv.org/abs/1905.07488 (Accessed: 30 November 2024).

[16] Deistler, M., Goncalves, P.J. and Macke, J.H. (2022) *Truncated proposals for scalable and hassle-free simulation-based inference, arXiv.org*. Available at: https://arxiv.org/abs/2210.04815 (Accessed: 30 November 2024).

[17] Al-Refaie, A.F. *et al.* (2021) *Taurex III: A fast, dynamic and extendable framework for retrievals, arXiv.org*. Available at: https://arxiv.org/abs/1912.07759 (Accessed: 30 November 2024).

[18] Batalha, N.E. *et al.* (2017) *Pandexo: A community tool for Transiting exoplanet science with jwst & hst, arXiv.org*. Available at: https://arxiv.org/abs/1702.01820 (Accessed: 30 November 2024).

[19] Neal, R.M. *et al.* (2012) *MCMC using Hamiltonian Dynamics*. Available at: https://arxiv.org/pdf/1206.1901 (Accessed: 04 December 2024).

[20] Hoffman, M.D. and Gelman, A. (2011) *The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, arXiv.org*. Available at: https://arxiv.org/abs/1111.4246 (Accessed: 04 December 2024).

[21] Blei, D.M., Kucukelbir, A. and McAuliffe, J.D. (2018) *Variational inference: A review for statisticians, arXiv.org*. Available at: https://arxiv.org/abs/1601.00670 (Accessed: 04 December 2024).

[22] Zingales, T. and Waldmann, I.P. (2018) *Exogan: Retrieving exoplanetary atmospheres using deep convolutional generative Adversarial Networks, arXiv.org*. Available at: https://arxiv.org/abs/1806.02906 (Accessed: 04 December 2024).

[23] Márquez-Neila, P. *et al.* (2018) *Supervised machine learning for analysing spectra of exoplanetary atmospheres, NASA/ADS*. Available at: https://ui.adsabs.harvard.edu/abs/2018NatAs...2..719M/abstract (Accessed: 04 December 2024).

[24] Fisher, C. *et al.* (2020) *Interpreting high-resolution spectroscopy of exoplanets using cross-correlations and supervised machine learning, arXiv.org*. Available at: https://arxiv.org/abs/1910.11627 (Accessed: 04 December 2024).

2025 S-T Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

[25] Nixon, M.C. and Madhusudhan, N. (2020) *Assessment of supervised machine learning for atmospheric retrieval of exoplanets*, *NASA/ADS*. Available at: http://ui.adsabs.harvard.edu/abs/2020MNRAS.496..269N/abstract (Accessed: 04 December 2024).

[26] Tejero-Cantero, A. *et al.* (2020) *SBI -- A toolkit for simulation-based inference*, *arXiv.org*. Available at: https://arxiv.org/abs/2007.09114 (Accessed: 04 December 2024).

[27] Foreman-Mackey, D. (2016) *Corner.py: Scatterplot matrices in Python*, *Journal of Open Source Software*. Available at: https://doi.org/10.21105/joss.00024 (Accessed: 04 December 2024).

[28] Aguila, A.L. (2024) *A tutorial on multi-view autoencoders using the multi-view-AE library*. Available at: https://arxiv.org/html/2403.07456v1 (Accessed: 04 December 2024).

[29] Dax, M. *et al.* (2023) *Flow matching for scalable simulation-based inference*, *arXiv.org*. Available at: https://arxiv.org/abs/2305.17161 (Accessed: 04 December 2024).

[30] Geffner, T., Papamakarios, G. and Mnih, A. (2023) *Compositional score modeling for simulation-based inference*, *Proceedings of the 40 th International Conference on Machine Learning*. Available at: https://proceedings.mlr.press/v202/geffner23a/geffner23a.pdf (Accessed: 04 December 2024).

[31] Bourque, M. *et al.* (2021) *The Exoplanet Characterization Toolkit (exoctk)*, *Zenodo*. Available at: https://zenodo.org/records/4556063 (Accessed: 04 December 2024).

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Acknowledgements

I would like to express my heartfelt gratitude to Mr. Shearman for his unwavering support throughout my research journey. His encouragement and motivation were invaluable, especially during challenging moments. His belief in my abilities helped me persevere and ultimately complete this work. Thank you for being an inspiring mentor and a constant source of guidance.

2025 S.-T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示

Commitments on Academic Honesty and Integrity

We hereby declare that we

1. are fully committed to the principle of honesty, integrity and fair play throughout the competition.

2. actually perform the research work ourselves and thus truly understand the content of the work.

3. observe the common standard of academic integrity adopted by most journals and degree theses.

4. have declared all the assistance and contribution we have received from any personnel, agency, institution, etc. for the research work.

5. undertake to avoid getting in touch with assessment panel members in a way that may lead to direct or indirect conflict of interest.

6. undertake to avoid any interaction with assessment panel members that would undermine the neutrality of the panel member and fairness of the assessment process.

7. observe the safety regulations of the laboratory(ies) where the we conduct the experiment(s), if applicable.

8. observe all rules and regulations of the competition.

9. agree that the decision of YHSA(Asia) is final in all matters related to the competition.

We understand and agree that failure to honour the above commitments may lead to disqualification from the competition and/or removal of reward, if applicable; that any unethical deeds, if found, will be disclosed to the school principal of team member(s) and relevant parties if deemed necessary; and that the decision of YHSA(Asia) is final and no appeal will be accepted.

Signed by Student:

Signed by Supervisor & School Principal:

Mr. Shearman / Supervisor

Mr Misso Veness / ESD Principal

2025 S.T. Yau High School Science Award
仅用于2025丘成桐中学科学奖论文公示