Exoplanet atmospheric characterization using amortized simulation-based inference

Malavika Vasist Savoyant

Exoplanet atmospheric characterization using amortized simulation based inference

Malavika Vasist Savoyant

A thesis submitted in partial fulfillment
for the degree of PhD in Sciences

University of Liège

2025

STAR Institute
Planetary & Stellar system Imaging Laboratory (PSILab)

Montefiore Institute
Department of Electrical Engineering and Computer Science

Supervisors

Dr. Olivier Absil
Prof. Gilles Louppe

Jury Members

Dr. Quentin Changeat
Dr. Valentin Christiaens
Prof. Maxime Fays
Dr. Paul Molliere

Cover image credits:

front : IAU/L.Calçada
back : ESA/Hubble

To my younger self

Abstract

Around 30 years after the first exoplanet detection and over 5000 detections later, we have come a long way in characterizing a huge diversity of exoplanets to understand their formation, evolution and habitability. Thanks to modern instrumentation providing high-quality spectra (emission and transmission), it is possible to study their structure and composition by atmospheric retrieval. Although conventional retrieval algorithms such as MCMC and nested sampling are reliable, they are limited in terms of time efficiency, scalability and testability. These limitations become more pronounced with the anticipated influx of spectral observations from JWST and future missions such as ARIEL, particularly in the context of population studies. This leads us to explore an alternative family of algorithms called simulation based inference, specifically using a variational deep-learning-based retrieval algorithm called neural posterior estimation (NPE), to estimate the posterior distribution directly by sidestepping likelihood computations. This algorithm improves over the traditional algorithms in speed, scalability and testability. Particularly, it offers amortization, which involves training a posterior estimator once that can then be used to perform quasi-instantaneous retrievals on subsequent observations of a similar kind. While this is useful for the rapid characterization of thousands of spectral sources in only a few hours, its other key advantage lies in enabling statistical tests (such as coverage tests and L-C2ST tests) to assess the validity of the retrieved posteriors, something which is otherwise not possible using conventional algorithms. In this thesis, we use NPE to perform spectral retrievals of six brown dwarfs (aka exoplanet analogs) ranging from L to Y spectral types, using various spectral wavelength regions and resolutions, in order to characterize them. We conduct a detailed atmospheric retrieval of two Y-type brown dwarfs using their mid-IR spectrum obtained with JWST/MIRI, together with their archival near-IR spectra. We additionally perform systematic retrievals on five brown dwarfs (including the previous two) using an amortized approach, and compare the results to identify trends, thus setting a precedent for future population studies using SBI. Lastly, we perform a pilot NPE retrieval study on a high-resolution near-infrared spectrum of an L type dwarf. We repeatedly perform validation tests across all retrievals in this thesis, and in various cases, compare these retrievals with that of nested sampling. We find that the NPE posteriors are valid and consistently broader than those obtained with nested sampling. We dissect the sources of Bayesian model uncertainty through a single retrieval, and suggest that nested sampling may produce overconfident posterior estimates. With these results, we also identify the challenges and opportunities for SBI in exoplanet retrievals going forward.

Acknowledgements

PhD is a challenging intellectual (and mental/emotional/physical) endeavor and I could not have attempted it without help. First and foremost, I would like to thank my supervisors Olivier Absil and Gilles Louppe for their immense support. I feel truly privileged to have worked under your supervision. Interdepartmental work is always challenging and you made it easier. I am very grateful to Olivier for providing me a safe space to make mistakes and learn from it. For empowering me to figure things out on my own, at my own pace, even during disagreements and moments of doubt. For the patience, the many insights, countless feedback and constructive criticisms that propelled me forward each time. I want to also thank Gilles for constantly challenging me and holding me to a better standard of conceptual clarity and understanding. Not having a background in statistics and a non-rigorous training in machine learning set me up for a steep learning curve in this PhD, and I want to thank you for letting me pick your brain, sometimes even outside of office hours, to match-up to an academic rigor that I am grateful to have experienced. I hope to always strive for it. I also want to thank the president of the jury Maxime Fays, and the jury members Quentin Changeat, Valentin Christiaens and Paul Molliere for accepting to read my thesis and to provide feedback.

I would like to thank my collaborators Paul, and Francois, and the entire MIRI/GTO team for providing me profound insights during our (some ongoing) collaborations. I would like to also thank all my (ex-) colleagues in Montefiore and PSILab (and other groups) namely Antoine, Arjun, Arnaud, Augustina, Carles, Christians, Djordje, Elena, Gilles, Guillaume, Omer, Iremsu, Jahanvi, Jyotirmay, Lorenzo, Mariam, Matt, Maxime, Nikolay, Niyati, Prashant, Rakesh, Sandrine, Sébastien, Valentin, and many more, for making my workplace experience cordial and fun-filled. As an international student who lived alone working long hours, colleagues become friends, and some friends become family. I am deeply grateful for the friendships I made and memories I can cherish. It was a great professional and personal experience alike. I am thankful for all the birthdays, the Christmas, Eid and Diwali celebrations throughout the years.

I would like to thank Nicole for answering all my administrative questions, organizing fun events, and most importantly processing my reimbursements. I would also like to thank Sacha, Joeri and Gilles for maintaining Alan, and David Collignon for nic5, and for the university of Liege.

Lastly, I would like to thank my family and closest friends for their support. Anaël, Mariam, Santhosh, Joelle and Josette. Anaël your support has been constant and unwavering throughout all the years I've known you in every way- emotional, mental, and practical. It is so silent, all encompassing and unconditional that sometimes it is hard to identify. I consider myself very privileged to have a partner like you who provides me such love, encouragement and strength.

Like with everything else in my life, thank you for making this arduous journey a little easier on me too. Also, Mariam, thank you for always being there for me. Your empathy has been a great source of healing and love, and I deeply cherish our friendship (and your food). Santhosh, thank you for all the fun and support these past few years. Merci à Josette et Joëlle pour votre gentillesse (et pour tout le délicieux vin), and my parents for always providing me the practical resources I needed to pursue my interests. My high-school physics teacher Mr George Pinto, for inspiring me and believing in me when I didn't.

This project has received funding from the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) through a FRIA doctoral grant, from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No 819155), and from the Wallonia-Brussels Federation (grant for Concerted Research Actions). Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by F.R.S.-FNRS under Grant No. 2.5020.11 and by the Walloon Region.

Acronyms

ABC	approximate Bayesian computation
ARIEL	Atmospheric Remote-sensing Infrared Exoplanet Large-survey
BI	Bayesian inference
CHARIS	Coronagraphic High Angular Resolution Imaging Spectrograph
CLT	central limit theorem
CNF	continuous normalizing flow
CNN	convolutional neural network
CRIRES	CRyogenic high-resolution InfraRed Echelle Spectrograph
DENIS	DEep Near-Infrared Southern sky survey
DNN	deep neural network
ELT	Extremely Large Telescope
ELU	exponential linear unit
ERIS	Enhanced Resolution Imager and Spectrograph
ESO	European Space Observatory
ESPRESSO	Echelle SPectrograph for Rocky Exoplanet and Stable Spectroscopic Observations
FMPE	flow matching posterior estimation
GAN	Generative adversarial neural network
GCM	general circulation models
GPI	Gemini Planet Imager
GPU	graphical processing units
GRAVITY	VLT interferometer for precision narrow-angle astrometry and interferometric imaging
GTO	guaranteed time observation
HARPS	High Accuracy Radial velocity Planet Searcher
HIRES	High Resolution Echelle Spectrometer
HMC	Hamiltonian Monte Carlo
HST	Hubble space telescope
HWO	Habitable Worlds Observatory
IS	importance sampling
JWST	James Webb Space Telescope
KL	Kullback-Leibler
KPIC	Keck Planet Imager and Characterizer
LBT	Large Binocular Telescope
L-C2ST	local classifier two-sample test

LIFE	Large Interferometer For Exoplanets
LMIRCam	LBT Mid-Infrared Camera
LTE	local thermodynamic equilibrium
MAF	masked auto-regressive flow
MAP	maximum-a-posteriori
MCD	Monte Carlo dropout
MCMC	Markov chain Monte Carlo
MH	Metropolis-Hastings
MLE	maximum likelihood estimation
MLP	multilayer perceptron
MIRI	Mid-Infrared Instrument
NAF	neural auto-regressive flow
NCSF	neural circular spline flow
NGRST	Nancy Grace Roman Space Telescope
NF	normalizing flow
NIRC2	Near-Infrared Camera 2
NIRSpec	Near-Infrared Spectrograph
NLE	neural likelihood estimation
NPE	neural posterior estimation
NS	nested sampling
NSF	neural spline flow
NRE	neural ratio estimation
OE	Optimal estimation
ODE	ordinary differential equation
PPD	posterior predictive distribution
ReLU	rectified linear unit
ResMLP	residual MLP
RF	Random forest
RV	radial velocity
SED	spectral energy distribution
SBI	simulation-based inference
SCExAO	Subaru Coronagraphic Extreme Adaptive Optics
SDSS	Sloan digital sky survey
SNPE	sequential neural posterior estimation
SPHERE	Spectro-Polarimetric High-contrast Exoplanet Research
VLT	Very Large Telescope
WISE	Wide-field infrared survey explorer
2MASS	Two micron all sky survey

Contents

Figures and Tables

Figures

Tables

Chapter 1

Introduction

1.1 Advent of exoplanet characterization

The very first exoplanet was detected around a pulsar by Wolszczan and Frail (1992), using the *pulsar timing method*. This method identifies variations in the pulsar's radio signals received due to exoplanets orbiting around it. It provided the first glimpse into an exoplanetary system, allowing measurements of planetary masses, orbital parameters such as their period and eccentricity, and distances from the host. However, not many planets could be detected this way since pulsars are rare.

A few years later, the first detection of an exoplanet around a sun-like (G-type) star was carried out by Mayor and Queloz (1995) using the *radial velocity method*. This method observes the Doppler shift in the star's spectrum due to the gravitational pull from an orbiting planet, allowing us to measure the planet's orbital parameters, and a lower limit on mass ($M \sin i$). However, it was the combination of radial velocity (RV) and *transit photometry method* that truly enhanced characterization. Charbonneau et al. (1999) used the transit method, which detects slight dips in a star's brightness when a (transiting) planet passes in front of it, along with the radial velocity method to measure the planet's radius and inclination for the first time. This allowed for the calculation of the true mass (not just a lower limit) and, in turn, its density, providing insights into its composition. Other notable methods for detection have been *microlensing* and *astrometry*, which also provide constraints on the mass and orbital parameters of some exoplanets.

Following advancements in instrumentation, high-precision *transmission spectra*, which measure the wavelength-dependent stellar light filtered through a planet's atmosphere during transit, enabled the first detection of a strong sodium absorption feature in a transiting exoplanet by Charbonneau et al. (2002). This characterization was followed by Charbonneau et al. (2005), who detected water vapor and carbon monoxide in another transiting exoplanet using its *emission spectrum*, which captures the planet's thermal radiation. Both kinds of spectra, transmission

and emission, complement each other, with the emission spectrum generally capturing the atmosphere at a lower altitude (2 to -2 in $\log_{10}$ bar), and the transmission spectrum probing higher (-2 to -3 in $\log_{10}$ bar). Currently there is also the avenue of the reflected light spectroscopy that includes the light reflected from a planet's atmosphere (Vaughan and Birkby, 2024).

5867 detections* and over 30 years later, we have come a long way in studying exoplanets. We can now characterize a huge diversity of exoplanetary sizes, temperatures, compositions and atmospheric processes. But why is this important?

1.2 Why study exoplanet atmospheres?

Two main motivations for studying exoplanet atmospheres are to understand how planets are formed, how they evolve, and if they have the potential to harbor life.

1.2.1 Planetary formation

Exoplanet atmospheres can be classified into primary and secondary atmospheres based on the nature of their formation. Primary atmospheres are those that have been retained since their formation, and they hold clues to how and where the planet was formed in the proto-planetary disk. In contrast, secondary atmospheres are those that were formed after their primary atmospheres were lost due to lower retention because of their planet's smaller size. These atmospheres are usually formed either by out-gassing from the planet's interior or by external factors such as comets (Tian and Heng, 2024).

In case of the primary atmospheres, signatures of an exoplanet's formation are imprinted in their atmospheres (Blanc et al., 2018). For instance, a higher metallicity implies that a higher amount of solid material was accreted during formation since solids are rich in heavy metals (Madhusudhan, 2019). The specific composition of the solids varies with the planet's relative position to the various molecular ice-lines and its migration history along with the evolution and temperature of the proto-planetary disk. Further, these variations are also reflected in the C/O ratio in their atmospheres (Öberg et al., 2011). It is also hypothesized that we can ascertain the formation pathway taken by an exoplanet to be either via core accretion (Pollack et al., 1996) or gravitational instability (Boss, 1998), using its isotopologue ratios in the atmosphere such as $^{12}C/^{13}C$, $^{14}N/^{15}N$, D/H ratios, etc.

*as of 25 March 2025 from `https://exoplanetarchive.ipac.caltech.edu/`

By conducting population studies of exoplanet atmospheres, we can empirically test and validate various planet formation hypotheses (Booth et al., 2017; Khorshid et al., 2022; Crossfield, 2023; Mordasini et al., 2016; Cridland et al., 2016). In this thesis, we study the isotopologue ratios $^{12}C/^{13}C$ (see Chapter 8) and $^{14}N/^{15}N$ (see Chapter 7) of some exoplanet analogues (brown dwarfs) to understand their formation pathways. Additionally, we also map C/O ratios of several brown dwarfs close in spectral type, to find any existing patterns linking them to their formation.

1.2.2 Habitability and the search for life

Another reason exoplanet atmospheres are studied is to find suitability for life. Telescopes like the Extremely Large Telescope (ELT) aim to search for bio-signatures that could indicate the plausibility of life (Zhang et al., 2024). Although this is beyond the scope of this thesis, the methods introduced here are applicable to future data analysis of bio-signatures.

1.3 Studying exoplanet atmospheres

The research of exoplanet atmospheres has broadly branched out into three areas of study, namely: *instrumentation/observational methods*, *modeling*, and *data analysis*. *Instrumentation and observational methods* includes the development of such technology as coronagraphs, advanced optics, and pre/post-processing techniques for high-contrast imaging. Observational strategies include various spectroscopic avenues such as transmission spectroscopy, direct imaging, and high-resolution Doppler spectroscopy, that enable the detection and characterization of atmospheric features. *Modeling* involves simulating the physical and chemical processes occurring within exoplanet atmospheres. This includes formalizing atmospheric dynamics, thermal structures, and chemical compositions to provide a theoretical framework for interpreting observational data. *Data analysis* refers to the application of statistical and computational techniques to extract meaningful information from observational data using the above mentioned theoretical models. Traditional methods include algorithms such as Markov chain Monte Carlo (MCMC) and nested sampling (NS) for parameter estimation and model comparison. This thesis explores the use of machine learning-based data analysis to study exoplanet atmospheres through spectroscopy, also with the long-term goal of conducting population studies. The following sections provide a description of the background of the observational methods and modeling framework within which this thesis contributes.

Chapter 2

Background

2.1 Observational methods

Although the first batch of exoplanets was studied through transits, there are now a few other avenues to study exoplanet atmospheres. They mainly include direct imaging spectroscopy and high-resolution Doppler spectroscopy. In this thesis, we study exoplanet atmospheres in both of these avenues.

2.1.1 Direct imaging spectroscopy

Direct imaging spectroscopy observes the light emitted/reflected by exoplanets by directly imaging them (Marois et al., 2008). In order to single out the planetary light, a coronagraph is generally used to block the light from the host star. Since this is a challenging task, this method is limited to planets that have large orbital separations from their host star, and/or are very bright or isolated. Although this limits the population of planets that can be studied, the final spectrum obtained can reach a sufficiently better S/N ratio than the transit method. This method has been used to obtain several spectra of young giant exoplanets in the near and mid-infrared wavelengths.

Direct imaging spectroscopy obtains the emission and reflection spectrum of an exoplanet. The emission spectrum provides insights into the temperature along the planet's atmosphere. Additionally, it also sheds light on the specific opacity sources such as molecular gases, that leave distinct signatures in the emitted radiation, hence revealing the atmospheric composition. Thus, with an emission spectrum, we can probe both the chemistry and the thermal profile of a planet's atmosphere. However, unlike in transits, it is not possible to constrain the true radius, mass, and hence gravity via this method. As the opacity sources heavily depend on gravity, this can lead to degeneracies between these parameters (see Lew et al. 2024). In contrast, the reflection spectrum provides information about the clouds, hazes, surface scattering, and composition of the upper atmosphere or surface.

Ground-based telescopes such as the Very Large Telescope (VLT) with Spectro-Polarimetric High-contrast Exoplanet Research (SPHERE) and Enhanced Resolution Imager and Spectrograph (ERIS), Keck with Near-Infrared Camera 2 (NIRC2) and Keck Planet Imager and Characterizer (KPIC), Subaru Telescope with Subaru Coronagraphic Extreme Adaptive Optics (SCExAO) + Coronagraphic High Angular Resolution Imaging Spectrograph (CHARIS), Gemini Observatory with Gemini Planet Imager (GPI), and the Large Binocular Telescope (LBT) with LBT Mid-Infrared Camera (LMIRCam), along with space-based instruments like the James Webb Space Telescope (JWST) with Near-Infrared Spectrograph (NIRSpec) and Mid-Infrared Instrument (MIRI), currently have the capabilities to provide direct imaging and spectroscopy for exoplanet research. In this thesis, we analyze some emission spectral data of isolated objects using data from Gemini, Hubble space telescope (HST) and JWST in Chapters 5 and 7.

2.1.2 High resolution Doppler spectroscopy

High resolution Doppler spectroscopy (Birkby, 2018) observes the thermal emission or reflected spectrum at a typical spectral resolution of 100,000 or higher. This method observes giant exoplanets close to the host star, especially hot Jupiters and isolated objects. It is motivated by the fact that at high resolutions such as this, the atomic and molecular lines in the spectrum are Doppler-shifted due to the planet's orbital motion. This Doppler shift allows for the separation of these lines from the relatively stationary stellar and telluric lines, enabling their removal (Brogi et al., 2012; Birkby et al., 2013; Cabot et al., 2019). The resulting residuals are cross-correlated with template planetary spectra containing the expected molecules. This is used to identify the signatures of the chemical species embedded in the spectra at a higher accuracy from the ground.

This method was first demonstrated with the detection of CO in the hot Jupiter HD 209458 b (Snellen et al., 2010), using the CRyogenic high-resolution InfraRed Echelle Spectrograph (CRIRES), which was installed on the ESO VLT in Chile until 2014 (Kaeufl et al., 2004), and further repurposed as CRIRES+ later in the decade. It has been used to obtain high-significance molecular detections and constraints on (C/O) ratio and metallicity, infer atmospheric processes (Flowers et al., 2019; Ehrenreich et al., 2020; Kesseli and Snellen, 2021), rotation rate (Snellen et al., 2014; Brogi et al., 2016) and the temperature profile (Schwarz et al., 2015; Brogi and Line, 2019; Gandhi et al., 2019; Yan et al., 2020) of various exoplanets. Further, constraints obtained on their radial velocity semi-amplitude and stellar velocity lead to independent constraints on their mass and orbital inclination (Brogi et al., 2012).

Currently, ground-based telescopes such as Echelle SPectrograph for Rocky Exoplanet and Stable Spectroscopic Observations (ESPRESSO) at the VLT, High Accuracy Radial velocity Planet Searcher (HARPS) at the European Space Observatory (ESO) 3.6-meter Telescope, High Resolution Echelle Spectrometer (HIRES) at Keck Observatory, and CRIRES+ at the VLT have the capability to perform high-resolution Doppler spectroscopy to observe exoplanet atmospheres. In this thesis, we analyze some spectral data obtained through this method from CRIRES+ in Chapter 8.

2.2 Modeling exoplanet atmospheres

Exoplanet atmospheres are modeled based on first principles by incorporating important properties and physical processes occurring within them. To effectively model these atmospheres, it is essential to first understand their underlying structure and mechanisms.

2.2.1 Understanding the atmospheres of exoplanets

An exoplanet atmosphere is a layered gaseous envelope with various ongoing processes inside it. When probed in certain regions of the atmosphere by observing specific wavelength ranges, we can observe processes associated with their spectral signatures. The major aspects of an atmosphere are summarized below (see also Figure 2.1).

Thermal profile

The thermal profile represents how the temperature varies along the height of the atmosphere. This depends on the composition of the atmosphere, stellar irradiation, and surface cooling and advection.

High opacity regions (lower regions of the atmosphere) result in absorbing and scattering more radiation, leading to higher temperatures, whereas less opaque regions (upper regions of the atmosphere) result in lower temperatures. Similarly, weakly irradiated objects (such as brown dwarfs and distant exoplanets) have a negative temperature gradient along the height of the atmosphere, whereas hot Jupiters that are highly irradiated have a positive gradient. Both strong irradiation from the host star and the presence of molecules such as TiO and VO can lead to thermal inversions, which make the temperature gradient positive in the upper atmosphere.

We can construct the thermal profile of exoplanet atmospheres using their emission spectra, and to a lesser extent, transmission spectra.

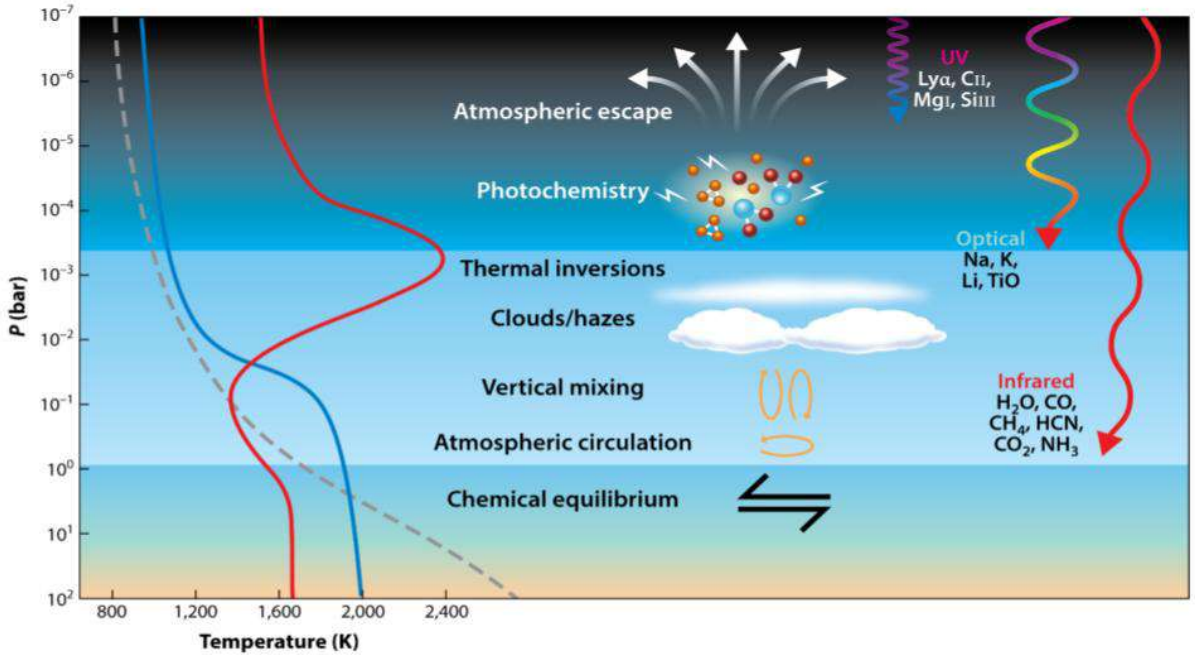


Figure 2.1: Illustrates the different processes occurring in an exoplanet atmosphere. On the left, the three types of thermal profiles generally seen in exoplanet atmospheres are represented. The red and blue lines on the left represent the thermal profile of an irradiated exoplanet with and without thermal inversion respectively, whereas, the gray dashed line represents a non-irradiated atmosphere. On the right, the different atmospheric depths and molecules accessible at different wavelengths are illustrated. This Figure was obtained from Madhusudhan (2019).

Chemistry

Exoplanet atmospheric chemistry refers to the composition, formation, and evolution of molecules and elements in the atmosphere, governed by processes like equilibrium reactions, photochemistry, and condensation.

The clues regarding a planet's chemistry lie mainly in the many absorption/emission features seen in different regions of its spectrum. Prominent molecular species such as H_2O, CO, CO_2 and CH_4 absorb in the infra-red due to rovibrational transitions. Heavy metals such as TiO, TiH and VO, and atomic species have strong absorption in the ultra-violet/optical light. Further, thermal inversions and clouds can exhibit emission features in the obtained spectrum. Lower regions of the atmosphere which are characterized by higher pressures (P>1 bar) and temperatures are considered to be in local thermodynamic equilibrium (LTE) and chemical equilibrium due to fast thermochemical reactions, whereas intermediate and upper regions can be in disequilibrium. This disequilibrium can be due to many external or internal atmospheric processes such as mixing, winds, or photo-chemistry due to host-star irradiation. This can be detected by comparing the observed abundances to equilibrium values calculated in theory.

Emission, reflection and transmission spectra of a planet provide deep insights into the chemistry of its atmosphere.

Clouds

If the atmosphere is cool enough, molecules can condense to form clouds. Alternatively, photochemical hazes are formed via stellar irradiation (Helling and Woitke, 2006; Helling et al., 2019). These particles absorb and scatter light coming from below and partly reflect external radiation (if any), which affects the spectra obtained from observations. By analyzing these spectral changes, we can constrain the microphysics of clouds, gaining insights into their composition, particle size, and distribution.

We can measure the opacity level of clouds lying within the photosphere, and when the data is sufficiently detailed, constrain their microphysical properties as well from their emission, reflection and transmission spectra.

2.2.2 Simulating exoplanet spectra

Modeling exoplanet atmospheres involves setting up the structure and the governing mechanisms in order to simulate the outgoing spectrum, which can then be compared with the observed spectrum to infer the underlying atmosphere. Generally, there are two main types of modeling approaches used in such studies. The first is the theoretical *self-consistent* modeling approach, where certain atmospheric properties such as gravity, effective temperature, C/O, metallicity, etc, are assumed. The corresponding atmospheric structure is then derived by solving the physical equations that govern it, and the outgoing spectrum is generated. The second is the *free parameter* modeling approach, which is semi-empirical in nature and makes no prior assumptions about the atmospheric properties, such that the structure is free to take any form to generate a spectrum. Additionally, models with varying levels of freedom also exist, which self-consistently derive parts of the atmospheric structure while freely parameterizing the rest. The development of both these approaches occurs in tandem, as each informs and refines the other.

Self-consistent models

Self-consistent modeling is conducted in two ways, namely the simpler 1D plane-parallel models and the 3D general circulation models (GCM).

In the 1D models, assumptions are made about the object's physical properties, such as gravity, atmospheric composition, and stellar irradiation, and the resulting atmospheric structure is computed by solving equations under the assumptions of horizontal homogeneity, hydrostatic and radiative (or radiative-convective) equilibrium (Hubeny, 2017) to generate a spectrum. However, recent works also account for disequilibrium chemistry (Mukherjee et al., 2024b, 2022; Karalidi et al., 2021; Lacy and Burrows, 2023; Phillips et al., 2020).

A grid of spectra is generated based on different initial assumptions about the atmosphere. There are many such grids available, and each is defined within a different range of C/O ratios, metallicities, irradiation, clouds, effective temperatures, etc leading to slightly different treatment of radiative transfer, convection, opacities, etc. In 3D GCMs, the full three-dimensional structure of the atmosphere is simulated, allowing for the modeling of both spectra and dynamical effects, as well as their interplay with chemical processes. While the 1D models are faster and take anywhere between a few hours to a few days, the 3D models take upto 2 months to compute. However, both ensure physically robust results.

Free-parameter models

Free-parameter models (Madhusudhan, 2018) assume no fixed atmospheric properties, and adopt a parameterized structure of the atmosphere. These parameters are "freely" defined. The thermal profile and molecular composition are represented through flexible parameterizations. This method does not always rely on any prior assumption of chemical or radiative equilibrium. These models usually take from seconds upto a minute to compute a spectrum. The caveat to this pace and flexibility is that, sometimes, these parameters can lead to non-physical parameter combinations to fit the data, as verified by theory. In this thesis, we use a free-parameter model called `petitRADTARNS` to generate synthetic spectra.

`petitRADTRANS` is a publicly available Python package to simulate emission and transmission spectra of exoplanet atmospheres in the infrared wavelength region, at both low (wavelength spacing $\lambda/\Delta\lambda = 1000$) and high resolutions (wavelength spacing $\lambda/\Delta\lambda = 1e6$). The low resolution is obtained using the correlated-k method and the high resolution is obtained using the line-by-line method. The package adopts a parameterized temperature structure, using formulations such as the one by Guillot (2010), although other profiles can be custom built and are used in some experiments in this report.

The available line opacities span temperatures between 80-3000K allowing the modeling of a diverse set of objects from the colder terrestrial planets to hot-Jupiters. These opacities can be parameterized to be vertically constant/variable with a slope through the length of the atmosphere. Based on this framework, various chemical models such as free chemistry, chemical equilibrium, and disequilibrium chemistry can be constructed. Further, it includes multiple cloud implementations in a single framework. Clouds can be included using both power-law or real condensation cloud treatments. In case of a microphysical treatment of clouds, amorphous or crystalline, spherical or irregularly-shaped cloud particles are available. The cloud particle size can either be specified through an eddy diffusion coefficient and sedimentation efficiency, or be directly parameterized. The particle size distribution follows a log-normal shape, with a variable width. Further, the abundances of clouds are also treated as parameterized quantities similar to molecular opacities.

This thesis explores a range of atmospheric configurations to model young non-irradiated giant exoplanets and brown dwarfs in `petitRADTRANS`, to generate their emission spectra, including two different thermal profile parameterizations, both equilibrium and disequilibrium chemistries, cloudy and cloud-free scenarios, and multiple spectral resolutions.

2.3 Data analysis

The inference of atmospheric characteristics from spectra is defined as an inverse problem and is called a *retrieval*. There are two main ways in which models are matched with the spectral data based on how uncertainties are obtained on the estimated parameters. These are described below.

2.3.1 Model fitting

Model fitting refers to the process of finding model parameters, say θ, that either minimize an error metric over the observed spectrum, such as least squares, or that maximize the likelihood of observing the spectrum given the model $p(x|\theta)$ (known as maximum likelihood estimation (MLE)). The output of such optimization results in a single point estimate of the model parameters that best fit the observation at hand. Given that there is usually a single observational spectrum for each object, the uncertainty around these model parameters can be calculated by methods such as a Fisher Information Matrix, etc.

Goodness-of-fit is an assessment of how close the model prediction is with the observation. The most common way of measuring a model's "goodness-of-fit" in such retrievals is using the chi-squared (χ^2) statistic (Andrae, 2010; Andrae et al., 2010). The proof of its validity as a measure of model fitting stems from its derivation under the assumption that the standard residuals per bin i, $r_i = x_i - f(\theta_i)/\sigma_i$, where i represents the N wavelength bins, x_i is the observed value, f is the forward model and $f(\theta_i)$ is the predicted value from the forward model, and σ_i is the standard deviation of the residual, per bin, are assumed to be normally distributed with a standard deviation of 1, when θ_i is the "true parameter", there is no model-data mismatch and the noise is truly Gaussian. The Gaussian residuals ensure that the sum of the squares of these standardized residuals follows the chi-squared statistic with $(N - p)$ degrees of freedom, where p is the number of model parameters. If the model is a good fit, the reduced chi-squared statistic ($\chi^2_\nu = \chi^2/(N - p)$) should be approximately 1. It may be overfitting if $\chi^2_\nu \ll 1$, and indicates a poor fit if $\chi^2_\nu \gg 1$.

The validity of the assumption of Gaussian residuals lies in the central limit theorem (CLT), which assumes a Gaussian noise. CLT states that the sum (or average) of a large number of independent and identically distributed random variables, regardless of the original distribution, will tend to follow a normal distribution, provided the variables have finite mean and variance.

Even though the individual and independent factors affecting the observed data, such as the photon counting statistics or instrumental effects like thermal noise, electronic noise, etc, some of which might not be Gaussian, and definitely correlated; for the sake of simplicity, a normal residual distribution is assumed. Here, the model-data match is also implicitly assumed when a model achieves the lowest χ^2 score.

This retrieval method is usually used for comparing the observational spectrum with a pre-existing grid of spectra generated self-consistently, in order to find the best match.

2.3.2 Bayesian atmospheric inference

An alternative way of performing atmospheric retrievals is by Bayesian inference (BI). This is a process of updating the prior hypothesis $p(\theta)$ over the model parameter space using the likelihood $p(x|\theta)$ based on the observation x. This produces an output posterior distribution $p(\theta|x)$ which is normalized by the evidence $p(x)$. The statistical model M includes the noise and the forward model. This is formulated as,

$$P(\theta|x, M) = \frac{P(x|\theta, M)P(\theta|M)}{P(x|M)}. \tag{2.1}$$

Instead of providing a single "best-fit" estimate (as in MLE above), which is considered the "true parameter", Bayesian inference provides the posterior distribution, or the probability of model parameters under the model considered. The posterior distribution encapsulates model uncertainty, which includes model-data mismatch (or model mis-specification), non-Gaussian noise, non-deterministic system, correlations between parameters, biased retrieval, lack of sufficient model parameter information in the data (or parameter non-identifiability), or lack of sufficient prior information (or broad prior), all causing uncertainty in parameter estimation.

Model consistency represents how well the model captures the data generating process, not just the current observation. A few ways of measuring the model consistency in Bayesian inference are by using posterior-predictive checks, coverage tests, and the local classifier two-sample test (L-C2ST), elaborated in Chapter 4, whereas relative consistency between two competing models is provided by the Bayes factor test.

In the exoplanet literature, a χ^2 test is also conducted as a proxy to estimate model consistency. Here, the maximum-a-posteriori (MAP) of each model is used as an MLE. Similar to the model-fitting test, it is assumed to be the "true parameter" and retroactively used to choose the model whose standardized residuals most resemble a Gaussian distribution (or whose reduced χ^2 is closest to 1) as adequate. Provided that usually in astronomy, the priors are broad uniform distributions, the χ^2 test on the MAP is motivated by likelihood assumptions such as model-data match, parameter identifiability, Gaussian noise (with or without correlations),

and a deterministic system where the observed spectrum is a typical realization of the noise distribution rather than a single random sample (implemented using a deterministic simulator in the likelihood evaluation). This leads to interpreting the posterior peak (MAP) as the "true parameter", effectively implying it can be uniquely and confidently determined.

In the ideal case that agrees with the assumptions above, the standardized residuals of the MAP are Gaussian and indeed result in a reduced χ^2 of 1. Therefore, metrics such as the reduced χ^2 test and the BIC criterion over the most probable sample provide approximations for model consistency. Here, the ensemble of standardized residuals will be a mixture of Gaussian distributions with a bigger spread than 1, due to the model uncertainty resulting from the Gaussian likelihood. Similarly, the posterior is Gaussian with a standard deviation comparable to the data noise. However, these assumptions are important to dissect for validity in reality when considering a deterministic simulator with a Gaussian likelihood.

Firstly, model-data match only means that it can reproduce the observed data within its assumed errors, but it does not guarantee that the model is "true" or a complete representation of the underlying system. It might still be a mis-specified or an incomplete description of the system. Which means that sometimes different models may be equally consistent with the data if the distinguishing features lie below the noise level (could be true for less prominent molecules, Welbanks et al. (2025)) and the model that "fits best" based on MAP χ^2 alone can actually be a less adequate model, since it does not consider posterior uncertainty (Gaarn et al., 2023). Secondly, all parameters might not be identifiable under the chosen model and data at hand, and sometimes that can cause degeneracies (especially under broad priors). This can be worsened by increasing the complexity of such models. Therefore, complexity might improve the MAP's resemblance to the observation; however, it does not always lead to "better fits" or a "true value". At best, it may perform no better than a simpler model. Thirdly, the considered Gaussian observation noise is usually reductionist as it considers only measurement noise but neglects model inadequacy and intrinsic system variability. Real systems often exhibit intrinsic variability and this makes the assumption of a deterministic simulator simplistic. This means that posteriors do not have to be Gaussian and can be multi-modal, skewed or non-Gaussian. Not considering the model inadequacy and system variability can skew the MAP from the "true value". On a side-note, it also means that the MAP sample can differ from the observation when stochasticity is considered in the simulator (which is true for some machine learning methods). Thus, above-mentioned caveats such as model-data mismatch, parameter non-identifiability, non-Gaussian noise, system stochasticity, along with broad priors, can bias the MAP from the "true value", and also lead to posteriors having widths that are much larger than observational errors.

A new observation of the same object would generally lead to a different posterior peak even if the model is "true", so the MAP should be interpreted as the most probable parameter for this specific dataset and model, rather than the absolute true value of the system. It is more meaningful to consider the posterior as a distribution of plausible parameter values, reflecting the uncertainty inherent in the data and model, rather than focusing solely on the MAP.

The Bayesian framework is particularly powerful in exoplanet atmosphere studies, as it allows us to rigorously account for uncertainties in the observed spectra and in the model assumptions. The method is data-driven, and the updates on the prior are done iteratively, either by sampling or by training machine learning algorithms. The method can be used either with the free-parameter or self-consistent models. The former usage is the focus of this thesis. Further sections discuss the algorithms that perform Bayesian inference.

2.3.3 Bayesian retrieval algorithms

An ideal Bayesian retrieval algorithm is accurate or asymptotically exact, which means that under ideal conditions, it provides an accurate posterior. It is also fast, potentially due to being able to parallelize. It is scalable, which means as the models become more complex and the number of dimensions of the model parameters increases, it is able to provide accurate retrievals at a computationally feasible pace. Lastly, it is evaluatable, which means that it can be tested on its accuracy using various statistical tests.

Conventional retrieval algorithms in the field are MCMC and NS. These algorithms are briefly described below.

MCMC

Common in exoplanet retrievals, MCMC variants like Metropolis-Hastings estimate the posterior indirectly by drawing samples from a target distribution instead, that is proportional to it. This is because normalizing the posterior requires the evidence,

$$P(x|M) = \int P(x|\theta, M)P(\theta|M)\, d\theta \tag{2.2}$$

which is the likelihood's weighted average. This integral is often intractable since marginalizing over the whole parameter space is either computationally not feasible, or the likelihood does not have a closed-form solution. Therefore, the target distribution (or the un-normalized posterior) is explored instead by constructing several parallel-running Markov chains.

A Markov chain is a type of mathematical model that represents a sequence of random events where the probability/state of the future time-step (or a future iteration), is only dependent on the previous state without holding any memory of the sequence of other states in the past, making it a stochastic process. This is seen in the equation below,

$$P(\theta_{n+1} = \theta_{n+1}|\theta_n = \theta_n, \theta_{n-1} = \theta_{n-1}, \ldots, \theta_0) = P(\theta_{n+1}|\theta_n) \tag{2.3}$$

where, the left-hand side of the equation represents the probability of being in state θ_{n+1} at time step $n+1$, given that the system is in states θ_n at time step n, θ_{n-1} at time step $n-1$, and so forth, all the way back to θ_0 at time step 0. The right-hand side represents the probability of being in state θ_{n+1} at time step $n+1$, given that the system is currently in state θ_n at time step n. This type of exploration of a parameter space is also called a "random-walk".

Each consecutive state in the target distribution is generally selected in regions that contribute more significantly to the evidence integral, generally assigning these samples higher probabilities. This means that there are more samples in the regions of higher probability of the posterior distribution. By the law of large numbers, the empirical distribution of the samples will eventually converge to the target distribution over many iterations.

Although these samples in a traditional Monte Carlo integral are statistically independent, the samples in MCMC are correlated. This is accounted for during error estimation over the model parameters. Additionally, the convergence of MCMC is usually tested using metrics such as the Gelman–Rubin diagnostic (Gelman and Rubin, 1992).

The Metropolis-Hastings (MH) method provides a rejection criterion (or probability) based on which a proposed sample is either accepted or rejected. There are many variants to this such as Gibbs sampling, Hamiltonian Monte Carlo (HMC) and a few others based on how the future samples are accepted and how they traverse through the parameter state space.

MCMC is asymptotically exact when the number of samples in the chain approaches infinity, and some algorithms can be parallelized to enhance pace. It performs at a feasible speed at lower dimensions (until 20). However, since every new dimension expects an exponential increase in the number of samples required to ensure an accurate estimation of the posterior, it is not feasible at higher dimensions. Further, since the nature of the algorithm is sequential, every new retrieval starts from scratch, and this makes it is difficult to evaluate. However, it has proven to be a reliable retrieval algorithm for exoplanet retrievals (Burningham et al., 2017; Madhusudhan et al., 2011, 2014; Line et al., 2013, 2014; Wakeford et al., 2017; Evans et al., 2017; Blecic, 2016; Ballard et al., 2011).

Nested sampling

While MCMC algorithms sample from the posterior distribution using Markov chains, NS (Skilling, 2004) approximates the posterior by moving through the prior in such a way that can be conceptualized as moving through various nested hyper-surfaces (or contours) in the likelihood space. This is achieved by dividing the likelihood into smaller segments and uniformly sampling within these constrained regions defined by specific likelihood thresholds. This process is repeated with different thresholds over time. The algorithm is shown in Algorithm 1.

This idea is rooted in the computation of the evidence $p(x|M)$. To sidestep the intractability, the evidence integral can be solved numerically using Riemann approximation by a finite sum where, the area under the integral is thought of as being equivalent to several 1D concentric/nested "shells" called iso-likelihood contours, that are defined at fixed levels of likelihood thresholds L_i. Each contour occupies a hyper-volume of the parameter space with a prior mass X_i that encompasses sampled θ with likelihood values greater than the likelihood threshold L_i. As the threshold increases, the hyper-volume associated to the contours progressively shrinks, along with the prior mass associated with it. The weight/prior mass between two consecutive nested volumes is given by $w_i = X_{i-1} - X_i$. It can be seen as the probability of the hyper-volume between contours, and can be thought of as a density. The product of this density and the likelihood gives the posterior mass which is also referred to as the typical set. This mass is where most of the integral's contribution comes from.

The integral is rephrased as the area under the 1D function of the likelihood value thresholds and its associated volume is given by,

$$P(x|M) = \int P(x|\theta, M)\, dP(\theta|M) \tag{2.4}$$

$$= \int_0^1 L(X)\, dX \tag{2.5}$$

$$\approx \sum_{i=i}^{N} L_i\, w_i \tag{2.6}$$

The weights derived from this approximation give the posterior estimate automatically, meaning that it directly estimates the typical mass. Convergence is obtained when the contribution of the remaining prior mass to the total evidence becomes negligibly small. The final estimate of evidence can then readily used for model comparison.

NS is more efficient in exploring the parameter space than vanialla MCMC, and especially performs better for complex multi-modal posteriors. However, some variants of MCMC are on par with NS and its variants in this regard. NS is asymptotically exact, however, as all sequential algorithms do, it suffers from scalability issues at higher dimensions and is not evaluatable. Since each retrieval starts from scratch, it is slow if one needs to conduct several retrievals at once. However, it is reliable, and hence its variant *MultiNest*, is the most commonly

used retrieval algorithm for atmospheric characterization (Lavie et al., 2017; Mollière et al., 2020; Todorov et al., 2016; Benneke and Seager, 2013; Waldmann et al., 2015a,b; Oreshenko et al., 2017; MacDonald and Madhusudhan, 2017; Gandhi and Madhusudhan, 2018). In all the *MultiNest* retrievals used in this thesis, the likelihood assumes that there is no stochasticity in the system, that is, the simulator is deterministic. Under this assumption, the posterior peak (MAP) corresponds to the parameter set that exactly reproduces the observed spectrum, meaning it is tied specifically to the observation at hand. This is also usually the case in literature.

Algorithm 1 Nested Sampling Algorithm

Start with N points $\theta_1, \ldots, \theta_N$ sampled from the prior.
for $i = 1, 2, \ldots, j$ **do** // The number of iterations j is chosen based on some convergence criteria.
    $L_i := \min(\text{current likelihood values of the points})$
    $X_i := \exp\left(-\frac{i}{N}\right)$
    $w_i := X_{i-1} - X_i$
    $Z := Z + L_i \cdot w_i$
    Save the point with the least likelihood as a sample point with weight w_i
    Update the point with the least likelihood using some Markov chain Monte Carlo steps according to the prior, accepting only steps that keep the likelihood above L_i
end for
return Z

2.3.4 Other approaches

Apart from the sequential algorithms described above, there have been other Bayesian and non-Bayesian algorithms used for atmospheric retrievals. Some are described below.

Optimal estimation (OE) is a numerical inversion technique grounded in Bayesian principles. It is very fast and operates under the assumption that the posterior is a Gaussian distribution. It is asymptotically exact under ideal conditions.

Random forest (RF) is a type of machine learning model that can be posed as a regression task to predict the atmospheric parameter values based on a given input spectrum in the form of decision trees (Márquez-Neila et al., 2018; Lueber et al., 2023). It needs to be trained only once and can predict the parameters for any input spectrum. It can be evaluated using the "real versus predicted (RvP)" analysis quantified by the coefficient of determination R^2, and it can assess feature importance. Further, the predictions from multiple decision trees can be averaged to account for numerical variations; however, this method does not provide a posterior distribution over the predicted parameters, only point estimates.

Generative adversarial neural network (GAN) are a type of neural networks (Zingales and Waldmann, 2018) that learn the complex underlying probability distribution of observational data. They can be repurposed to perform amortized atmospheric inference. In this setup, one generates new synthetic data samples along with the associated model parameters in a single

frame. This way the network learns to identify samples from the joint distribution of the observational spectrum and model parameters in the frame. Once trained, the trained network enables ”in-painting” to fill in a point estimate of the missing parameters based on any observed spectrum. The training of the network is framed as a classification task. It is trained until the network cannot distinguish between the real images of simulated spectrum with its model parameters versus generated in-painted images anymore. The network has to be trained only once.

Monte Carlo dropout (MCD) is an approximation to variational inference (VI, Jordan et al., 1999) to predict the atmospheric parameter values corresponding to an input spectrum (Soboczenski et al., 2018; Cobb et al., 2019; Ardévol Martínez et al., 2022a). Based on the dropout technique, Gal and Ghahramani (2016) showed that multiple forward passes where different weights are dropped each time result in an output of predictive samples that build a predictive distribution. While Soboczenski et al. (2018) uses this to directly approximate the non-parameterized posterior distribution over the atmospheric model parameters, Cobb et al. (2019) and Ardévol Martínez et al. (2022a) use this to predict the covariance and the mean that parameterize a multivariate Gaussian distribution, thus approximating a tractable posterior distribution over the model parameters.

The next chapter outlines the main research questions driving this work and presents the structure of the thesis designed to answer them.

Chapter 3

Research questions

3.1 Studying Brown Dwarfs in the mid-infrared

Brown dwarfs serve as a bridge between planetary and stellar objects. At lower gravities (3.5-4.5) brown dwarfs with similar masses at identical temperatures are expected to resemble the atmospheres of directly imaged giant exoplanets that lie far from the host star. This provides interesting brown dwarf/planet analogs such as Y dwarf WISE 0855 (Kühnle et al., 2024) and exoplanet Eps Ind Ab (Matthews et al., 2024), which have very similar effective temperatures. Such a resemblance implies similar convective interiors with outer envelopes of dominantly H+He with other minor opacity species, such that many brown dwarf atmospheric assumptions of radiative, convective, hydrostatic, and chemical equilibrium are valid starting points to study exoplanet atmospheres. They are also both weakly irradiated (unlike hot-Jupiters), the latter owing to their distance from the host star, resulting in a negative temperature gradient along the height of the atmosphere with little to no thermal inversions. As a result, both their spectra show absorption features from key molecules like CO, CH_4, and H_2O. They can also exhibit non-equilibrium chemistry and thick clouds in their atmospheres. However, exoplanets at the same temperature have been observed to have more clouds because the transition from a cloud-free to a cloudy atmosphere occurs at lower temperatures for exoplanets. This is due to their lower surface gravity because of their mass and/or youth, which causes cloud formation to happen at lower temperatures compared to brown dwarfs. These resemblances make brown dwarfs interesting objects to study.

Brown dwarfs are easier to observe than exoplanets because they lack the complication of a bright host star, which would otherwise interfere with their emissions and be challenging to eliminate during high-contrast imaging. However, they are brighter in the mid-infrared range (5 to 30 μm), which is difficult to observe from the ground. Especially wavelengths between 8 to 13 μm are challenging due to huge thermal backgrounds preventing very faint observations, thus restricting such observations to space-based telescopes. Ground-based missions such as Two micron all sky survey (2MASS), DEep Near-Infrared Southern sky survey (DENIS) and Sloan digital sky survey (SDSS), along with early space missions such as Wide-field infrared survey explorer (WISE), discovered brown dwarfs in bulk by exploring the near-infrared region. One of the most significant insights from these studies has been the L/T transition, where

silicate clouds sink deep and methane absorption features emerge. This transition illustrates the complexity of brown dwarf atmospheres, and also exemplifies the broader clarity these efforts have brought to understanding their atmospheres. There have also been significant insights into other spectral types, to uncover the atmospheric processes that shape them. These observations have not only refined brown dwarf science, but also established them as critical benchmarks for understanding giant exoplanets.

On continuing with the recent advancements in detector technology and cooling systems essential for mid-infrared observations, the Mid-Infrared Instrument (MIRI) onboard JWST, for the first time, brings the sensitivity, resolution, and broad wavelength coverage required for detailed emission spectral retrievals in the mid-infrared region. The MIRI instrument comprises a medium-resolution spectrograph (MRS) which enables the study of colder objects such as Jovian and sub-jovian exo-planets and brown dwarfs with spectral types M7 and later, that are much brighter in the mid-infrared.

The mid-infrared region probes higher altitudes in the atmosphere compared to the previously studied near-infrared region due to the wavelength dependence of opacity (see Figure 2.1), offering a new perspective on the exoplanet and brown dwarf atmospheres. Studies on combining the near and mid-infrared regions has allowed us to place improved constraints on bulk physical properties such as surface gravity, radius, mass, luminosity, and chemical composition. It has also enabled validate and improve constraints on the abundances of the main molecules present in such atmospheres, namely H_2O, CO, CO_2, CH_4, NH_3, TiO, VO, HCN and SO_2, some of whose absorption lines dominate this region. This has further facilitated a more accurate calculation of bulk C/O ratio and metallicity, which sheds light on the early environment of brown dwarfs and the location in the disk that exoplanets were formed. The mid-infrared region has also uncovered some of their isotopologues such as $^{15}NH_3$, which aids in a better understanding of formation pathways taken by these objects (Barrado et al., 2023). A robust understanding of the composition has also lead to confirming a potential prevalence of disequilibrium chemistry in their atmospheres, such as by such as increased CO or NH_3, and other physical processes such as the formation of H_2O clouds. Such an enhanced picture of the atmosphere provided by this region, along with more complex physical forward modeling, has also helped in breaking degeneracies in several molecular abundances and physical parameters like radius and gravity. However, uncovering these details requires accurate and fast retrievals, therefore, we explore alternative approaches to conventional algorithms to overcome their shortcomings.

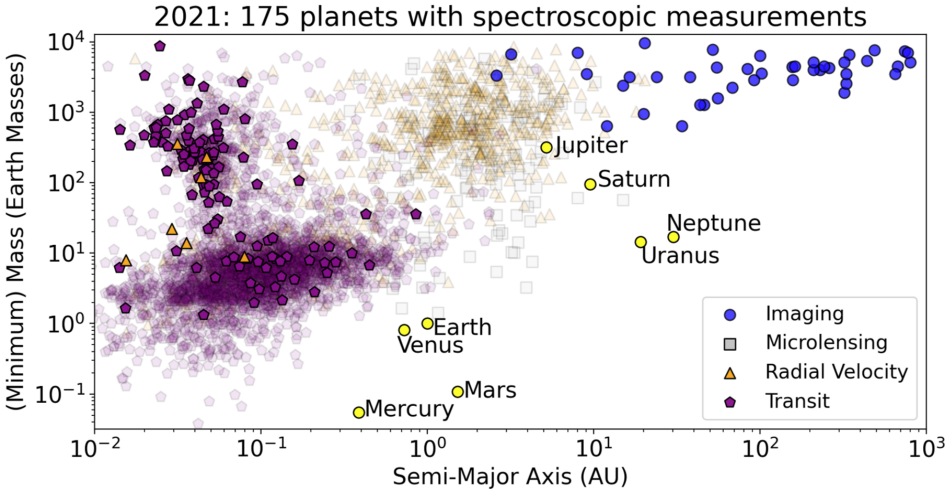


Figure 3.1: Exoplanets with spectroscopic measurements as of 2021 from Currie et al. (2023).

3.2 Shortcomings of conventional algorithms

While the sampling-based algorithms such as MCMC and NS mentioned above are asymptotically exact, they can require extensive computation times, ranging from a few hours to several days for each single retrieval (Cole et al., 2022). This means that processing just a few observations can quickly accumulate to several weeks of computing time, making detailed retrievals for large catalogs of observations impractical. Even considering retrievals on the current inventory (as of 2021) of 175 exoplanets with spectroscopic measurements illustrated in Figure 3.1 from Currie et al. (2023), these approaches becomes computationally expensive. The cost of retrievals will only worsen with observations from JWST and future missions such as ELT, Nancy Grace Roman Space Telescope (NGRST), Large Interferometer For Exoplanets (LIFE), Habitable Worlds Observatory (HWO), and Atmospheric Remote-sensing Infrared Exoplanet Large-survey (ARIEL), the latter of which is expected to generate thousands of transit spectra, making these algorithms largely unfeasible.

Additionally, the computational demands needed to maintain accurate results often scale poorly with the number of model parameters. This issue is particularly pronounced in simulation models with many nuisance parameters. While these parameters' posteriors are typically not of direct interest, they must still be computed because sampling-based approaches require sampling the full joint posterior. This encompasses spectral or noise scaling parameters, some P-T profile parameters that describe the complex function, and cloud parameters that are unobservable and may or may not correspond to a physically meaningful state.

Furthermore, the reliability and statistical rigor of approximations produced by sampling-based algorithms are challenging to assess. Statistical validation methods, such as simulation-based calibration (Talts et al., 2018) or expected coverage (Hermans et al., 2020), require repeated inferences, which are not feasible within a reasonable time frame. Therefore, improving these shortcomings is a main motivator of this thesis.

3.3 Research questions

In this thesis, we aim to answer two central questions that drive both the methodological and observational aspects of our work, namely:

1. How can we address and mitigate the key shortcomings of sequential atmospheric retrieval algorithms such as MCMC and NS, i.e., relatively slow retrievals that suffer from the dimensionality curse and cannot be evaluated for statistical rigor?
2. What new insights can we gather into the atmospheric structure, composition and formation of Y-type brown dwarfs from their mid-infrared spectrum observed by JWST/MIRI?

3.4 This thesis

We address the shortcomings of the conventional algorithms by exploring a machine learning–based inference method, belonging to the simulation-based inference (SBI) family, to perform atmospheric retrievals. SBI (Cranmer et al., 2020), as the name suggests, is an umbrella term for inference methods that use simulations from complex models to perform statistical inference instead of relying on an explicitly defined likelihood function. This enables inference for simulators that have intractable likelihoods. Doing so also enables improved speed, scalability, and testability. We then use SBI to perform an in-depth study of the atmospheres of several brown dwarfs from the MIRI and the CRIRES+ programs. This work is structured in the following way.

- In Chapter 4, we introduce the principle of SBI, highlight its advantages, and describe the range of algorithms it includes, especially recent ones that use probabilistic models in machine learning. We specifically focus on the algorithm called NPE, which is used in this thesis for conducting atmospheric retrievals, and provide a detailed description of its principle. Additionally we review previous applications of SBI in exoplanet atmosphere studies, and present a working proof of concept to apply NPE to characterize the atmosphere of a simulated spectrum of HR 8799 e. Further, we benchmark the results against *MultiNest* from Mollière et al. (2020). We show that NPE is not only faster than NS and scalable, but also a reliable retrieval algorithm by performing various validation procedures on the estimated posterior. This work is published in Vasist et al. (2023).

- In Chapter 5 we present the first scientific application of NPE in the field of atmospheric retrieval, by performing combined retrievals on near and mid-infrared spectra of the Y-type brown dwarf WISEP J173835.52+273258.9 (henceforth WISE 1738) from the HST, Gemini and JWST observatories. We constrain its thermal profile and the abundance of major opacity species. We estimate its physical parameters like effective temperature, mass and gravity that lead to values consistent with evolutionary models, along with consistent age estimates with respect to the fast rotation seen. We also uncover dis-equilibrium chemistry in its atmosphere. This work is in review in Vasist et al. 2025 (submitted).
- In Chapter 6, we test the theoretical predictions of the presence of water clouds in the atmosphere of WISE 1738. We use our cloudy model retrieval as a general case study of the different sources of model uncertainty usually seen in estimated posteriors from retrievals. We further implement the Bayes factor test to ascertain which model fits the WISE 1738 spectrum best, and apply importance sampling to our retrievals to correct for bias, hence uncovering the nature of NPE posteriors. We find that the cloud-free model best explains the spectrum at hand.
- In Chapter 7, similar to Chapter 5, we perform combined retrievals of the near and mid-infrared spectra of Y brown dwarf WISEP J182831.08 +265037.8 (henceforth WISE 1828) from the HST and JWST observatories. We benchmark the NPE retrievals with a *MultiNest* retrieval, and use the detection of $^{15}NH_3$ in its atmosphere to interpret its formation history. Although this analysis hints at model-data mismatch, we suspect a cloud-free atmosphere for this brown dwarf based on better predictive power. Additionally, we perform a short preliminary systematic study of the five brown dwarfs including WISE 1738 and WISE 1828, and three other late-T type and Y brown dwarfs namely ROSS 458c, WISE J045853.90+643451.9 (henceforth WISE 0458) and WISEJ0855–0714 (henceforth WISE 0855), to set a precedent for future rapid population studies using NPE/SBI. We compare the inferred properties from their retrievals performed on only mid-infrared spectrum from JWST, along with previous value estimates, to find any trends in their physical and chemical properties. Apart from the expected trend of decreasing luminosity and an upper limit decrease of the HCN abundance with temperature, we also find that the ammonia abundance does not deplete as much as is expected from equilibrium chemistry calculations.
- In Chapter 8 we present a working proof of concept to perform NPE retrievals on a high-resolution Doppler spectrum from CRIRES+ of the L dwarf DENIS J025503.3-470049 (henceforth DENIS J0255), and validate these results. We also perform a comparative study with *MultiNest*. Interestingly, we do not find evidence for ^{13}CO, which is only tentatively detected with low confidence in literature.

- In Chapter 9 we highlight the main results of this work, addressing our research questions. We identify the shortcomings of NPE that we were confronted with during our work, and note the various areas of improvement for the future. We also provide a future scope for SBI in the bigger context of upcoming missions and surveys.

Chapter 4

Simulation based inference

In this chapter, we describe the principle of SBI and how it can overcome the limitations of traditional retrieval algorithms. To bolster this claim, we present a working proof of concept to apply the SBI algorithm NPE to characterize exoplanet atmospheres. Using the radiative transfer model `petitRADTRANS`, we train a neural flow network to quickly estimate the posterior over a simulated spectrum of exoplanet HR 8799 e as a benchmark. We compare this against retrievals computed with `MultiNest`. We find that NPE produces accurate posterior approximations while reducing inference time down to a few seconds. Further, we demonstrate the computational faithfulness of our posterior approximations using inference diagnostics including posterior predictive checks and coverage, taking advantage of the quasi-instantaneous inference time of NPE. We also describe the validation procedure called the L-C2ST, by Linhart et al. 2024. The analysis confirms the reliability of the approximate posteriors produced by NPE. This work is published in Vasist et al. (2023).

4.1 Introduction to Simulation Based Inference

SBI is performed within a framework of a simulator of a complex system that takes a model parameter value vector θ as input, samples a series of physically meaningful internal states of the system represented by latent variables $z \sim p(z|\theta)$ which can themselves be either physical or non-physical in nature, and generates synthetic observations $x \sim p(x|z,\theta)$ as output. When the latent parameters that define the internal states are neither directly nor indirectly observable, and/or their behavior is poorly understood or stochastic, it results in intractable likelihoods, as computing it requires integrating over the joint likelihood $p(x,z|\theta)$ across the latent space such that,

$$p(x|\theta) = \int p(x,z|\theta)dz = \int p(x|z,\theta)p(z|\theta)dz\,, \tag{4.1}$$

In this case, performing Bayesian inference using conventional sampling-based approaches becomes impossible. This is also the case when the likelihood is actually tractable, but computationally too expensive to calculate. Therefore, we seek alternative approaches such as SBI to sidestep the issues brought about by intractability. It achieves this by using the simulator itself to define the data-generating process, thereby implicitly encoding the likelihood.

In atmospheric retrievals however, the self-consistent and free-parameter simulators are almost always deterministic for simplicity, and the gaussian likelihood is accessible. Yet, when data quality will make it possible to account for more details in the cloud physics, more sophisticated simulation models can involve a large number of interfering stochastic processes, resulting in an implicit or intractable likelihood. Possible examples of these processes include the cloud formation mechanisms (e.g., via seeding by nucleation, Lee et al., 2018), their growth (e.g., via coagulation or surface growth, Helling and Woitke, 2006), their diffusion processes and interactions with the surrounding thermodynamic conditions (e.g., by settling and mixing, Woitke et al., 2020), or their evolution with time (e.g., by ionization).

Although addressing intractability is the primary motivation for SBI, many algorithms, especially those that use deep learning, provide other advantages compared to likelihood-based sampling algorithms. For instance, they are *highly parallelizable* and can leverage the usage of graphical processing units (GPU) acceleration, making them computationally efficient. Additionally, the neural networks' ability to generalize better means that they do not need a lot of samples to learn the likelihood, making them highly *scalable* at larger dimensions. Since these networks clearly separate the training and the inference stages, SBI also enables the *reusability of the simulated models* to create more complex models (such as a mixture or combination of models such as patchy clouds), or for different noise models. The latter can be used to systematically understand the effect of the noise model on the retrieval. This is essential since modeling noise is still a challenging feat in exoplanet spectroscopy. Furthermore, they are very expressive and can handle complex distributions such as multiple mode features more efficiently.

Additionally, some SBI algorithms offer *amortization*, where the network is trained only once and can then be used to retrieve multiple observations, which significantly speeds up the inference process. This is especially appealing in the wake of the plethora of spectral observations expected from JWST and future missions such as ARIEL. Performing retrievals on thousands of spectra using conventional algorithms will take decades. However, with amortization, that time is reduced to a couple of hours. Amortization also enables testability of the retrieved posterior for statistical rigor, allowing for important coverage tests and model validation, which is otherwise not feasible using conventional algorithms. These features are what make SBI methods highly appealing, and they are the focus of our current exploration.

The most popular of the SBI algorithms include the approximate Bayesian computation (ABC), Neural posterior estimation (NPE), neural likelihood estimation (NLE), neural ratio estimation (NRE) and their sequential counterparts. In the **ABC** approach, one compares the observed data and the simulated one based on a distance measure that involves some summary statistics. The random forest algorithm mentioned in the previous chapter (Márquez-Neila et al., 2018; Lueber et al., 2023; Sisson et al., 2018) is a form of ABC algorithm. In contrast, **NRE** (Hermans et al., 2020; Durkan et al., 2020), **NLE** (Papamakarios and Murray, 2016; Alsing et al., 2018; Papamakarios et al., 2019) and **NPE**, build surrogates for the likelihood-to-evidence ratio, the likelihood, and the posterior, respectively. In this thesis, we use the algorithm **NPE** for our retrievals.

4.2 Neural posterior estimation (NPE)

NPE enables the parameterization of the posterior using a neural density estimator. A density estimator is a model that estimates the probability density function (PDF) of a random variable based on observed data. Some examples of neural estimators are normalizing flows, auto-regressive models, and variational autoencoders. In this thesis, we use normalizing flows as density estimators.

4.2.1 Normalizing flows

Normalizing flows (Papamakarios et al., 2021) are invertible bijective transformations, denoted g, that transform a random variable u with a simple probability distribution to a random variable v with a more complex distribution, such that $v = g(u)$. The simple distribution is usually a known one, such as a uniform or Gaussian, while the complex distribution is the target distribution to be estimated. By learning these invertible transformations, one can directly estimate the complex posterior distributions. An example of a normalizing flow is illustrated in Fig. 4.1.

A discrete implementation of normalizing flows is based on the change of variables theorem, where the probability density of v is given by,

$$\log p(v) = \log p(t(v)) + \log |\det \mathrm{J}_t(v)| \tag{4.2}$$

$$= \log p(u) - \log \left|\det \mathrm{J}_g(u)\right| \tag{4.3}$$

Here, $u = t(v)$ with $t = g^{-1}$ denoting the inverse transformation. $\mathrm{J}_t = \frac{\partial t}{\partial v}$ and $\mathrm{J}_g = \frac{\partial g}{\partial u}$ represent the Jacobians of the forward and inverse transformations, respectively. The Jacobian determinant $|\det \mathrm{J}|$ accounts for the local changes in volume under the transformation.

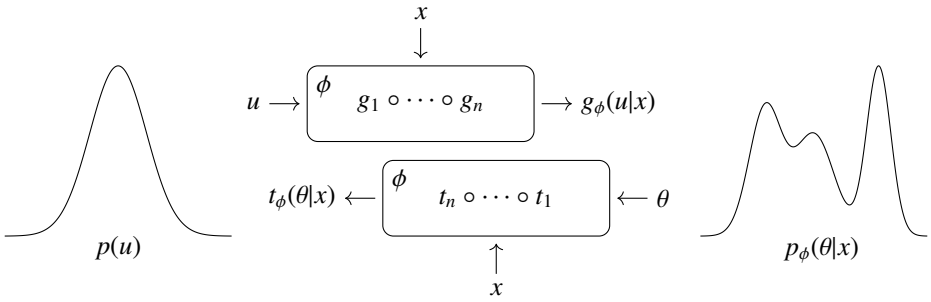


Figure 4.1: Transformation of a random variable u with probability density $p(u)$ toward a variable θ with probability density $p_\phi(\theta|x)$.

Under transformation t, the density "flows" from a complex distribution to a simpler one. When this simpler distribution is chosen to be a standard normal distribution, the transformation effectively "normalizes" the complex distribution, giving rise to the term normalizing flows.

The expressiveness of normalizing flow (NF) can be increased by stacking up the parametric transformations such as $u = t_n \circ t_{n-1} \circ \ldots \circ t_1(v)$. This results in the probability density,

$$\log p(v) = \log p(z_n) + \sum_{i=1}^{n} \log \left| \det \frac{\partial t_i(z_{i-1})}{\partial z_{i-1}} \right| \tag{4.4}$$

where $z_i = t_i(z_{i-1})$ and $z_0 = v$. This stacking enables the estimation of highly complex distributions.

Depending on how the transformations are parameterized, there are three popular implementations of discrete normalizing flows. They are the auto-regressive flows, including the masked auto-regressive flow (MAF) and the neural auto-regressive flow (NAF) (both used in this thesis), and the neural spline flow (NSF).

Auto-regressive flows

Auto-regressive flows form a class of flows that are constructed by parameterizing a single transformation t as a sequence of univariate, conditional transformations τ_i, applied dimension-wise to the input v_i. This structure is auto-regressive, such that,

$$u_i = \tau_i(v_{\leq i}) = \tau_i(c(v_{<i}), v_i) \tag{4.5}$$

Here, c is called the auto-regressive conditioner (Huang et al., 2018).

This formulation results in a triangular Jacobian matrix,

$$J_t = \frac{\partial t}{\partial v} = \begin{pmatrix} \frac{\partial \tau_1}{\partial v_1} & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & 0 & 0 & 0 \\ \frac{\partial \tau_i}{\partial v_1} & \cdots & \frac{\partial \tau_i}{\partial v_i} & 0 & \vdots \\ \vdots & & & \ddots & \frac{\partial \tau_n}{\partial v_n} \end{pmatrix} \tag{4.6}$$

which simplifies the log determinant to the sum of diagonal elements,

$$\log |\det J_t(v)| = \sum_i \log \left| \frac{\partial \tau_i}{\partial v_i}(v_{\leq i}) \right| \tag{4.7}$$

This structure ensures that the Jacobian determinant is tractable to compute. The two auto-regressive flows used in this thesis are MAF and NAF described next.

Masked auto-regressive flow

The Masked auto-regressive flow (Papamakarios et al., 2017) uses affine univariate transformations for each τ_i, such that equation 4.5 takes the form,

$$u_i = \frac{v_i - \mu_i(v_{<i})}{\exp(\sigma_i(v_{<i}))} \tag{4.8}$$

where μ_i and σ_i are unconstrained parametric functions of $v_{<i}$ defined by a masked neural network. This transformation is affine in u_i: shifted by μ_i and scaled by $\exp(\sigma_i)$. All μ and σ parameter values for each τ_i are computed in a single forward pass, as illustrated in Fig. 4.2.

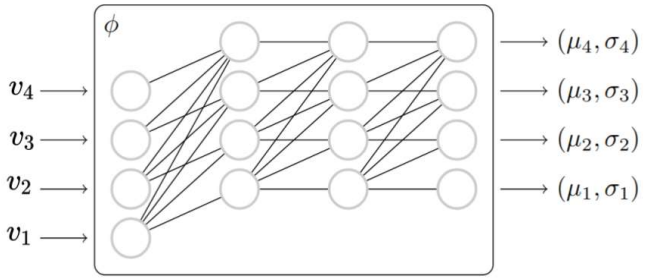


Figure 4.2: A simplified illustration of auto-regressive masking using dropped connections, used to learn the pairs (μ_i, σ_i) in a single MAF transformation (Rozet, 2022). The network is parameterized by weights and biases denoted by variable ϕ.

The resulting Jacobian determinant simplifies to,

$$\log |\det \mathrm{J}_t(v)| = -\sum_i \sigma_i \tag{4.9}$$

Neural auto-regressive flow

Neural auto-regressive flow defines each transformation τ_i as a deep neural network (DNN), and equation 4.5 takes the form,

$$u_i = DNN(c_i(v_{<i}), v_i) \tag{4.10}$$

As with MAF, the conditioning signal c and the output u for each τ_i are computed in a single forward pass. This is shown in the simplified NAF illustration in Fig. 4.3.

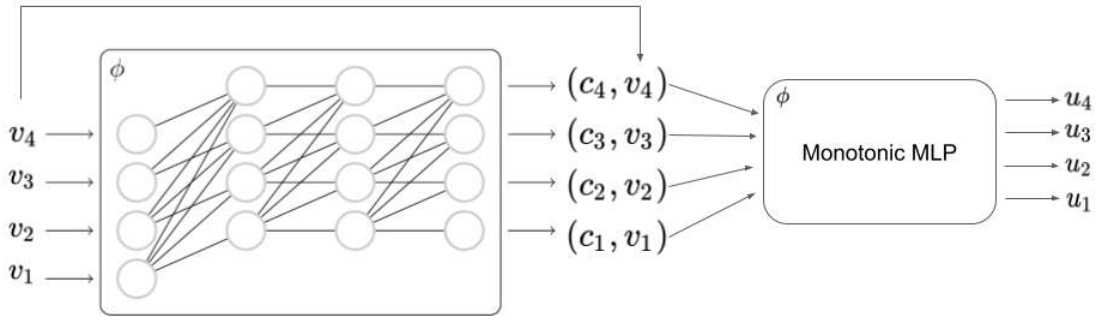


Figure 4.3: A simplified illustration of auto-regressive masking in a neural network, used to compute the conditioning signal c. This signal is then concatenated with the input v before being fed into a monotonic MLP transformer. Both networks are parameterized by weights and biases denoted by variable ϕ.

To ensure that the network is invertible, it is designed to be strictly monotonic. This is accomplished by enforcing strictly positive weights and using activation functions that are themselves strictly monotonic. In this way, the transformations can be parameterized by a neural network to construct new families of distributions.

In practice, to evaluate the log-probability $\log p(v)$ of a sample from the complex distribution, one inverts the transformation to t, to recover u, and computes the density using the change-of-variables formula. This is possible to achieve via a single forward pass through the auto-regressive model, and is fast. In contrast, sampling from the more complex distribution involves drawing a sample from the base distribution u and applying the forward transformation g to obtain v. Since t is auto-regressive, each v_i depends on the previous inputs $v_{<i}$; therefore, it must be computed sequentially in order $v_1, v_2, \ldots v_n$ for each of the n dimensions. This is slightly less efficient.

4.2.2 Parameterizing the posterior

In our application, the posterior density $p_\phi(\theta|x)$ represents a complex distribution over the variable θ (previously introduced as v), which we model through a sequence of transformations applied to a base normal variable u with a known probability density $\mathcal{N}(u)$. As illustrated in Fig. 4.1, these transformations are implemented as invertible neural networks conditioned on the observation x. Hence, they are also called conditional normalizing flows. Here ϕ denotes the set of learnable parameters of the network.

Training

The NPE pipeline is illustrated in Fig. 4.4. The training is carried out by generating a training set $\{(\theta, x)\}$ from a joint simulation model $p(x, \theta) = p(\theta)p(x|\theta)$ of model parameters θ and its corresponding simulated exoplanet spectral observations x. This is then input into a conditional normalizing flow $p_\phi(\theta|x)$, which is composed of an embedding network and a few invertible transformation networks t_i, and trained to estimate the posterior density $p(\theta|x)$.

The weights of the networks are tuned such that the conditional normalizing flow $p_\phi(\theta|x)$ approximates the true posterior distribution $p(\theta|x)$. The training is based on amortized variational inference and amounts to the minimization of the expected forward Kullback-Leibler (KL) divergence between $p(\theta|x)$ and $p_\phi(\theta|x)$ (Agakov, 2004), that is

$$\begin{aligned}\phi^* &= \arg\min_\phi \mathbb{E}_{p(x)} \left[\mathrm{KL}\big(p(\theta|x) \parallel p_\phi(\theta|x)\big)\right] \\ &= \arg\min_\phi \mathbb{E}_{p(x)}\, \mathbb{E}_{p(\theta|x)} \left[\log \frac{p(\theta|x)}{p_\phi(\theta|x)}\right] \\ &= \arg\min_\phi \mathbb{E}_{p(\theta,x)} \big[-\log p_\phi(\theta|x)\big]. \end{aligned} \tag{4.11}$$

Intuitively, this is reducing how "different" the two distributions are. A smaller KL divergence implies that the two distributions are more similar. Thus, training encourages the flow to assign high posterior probability to samples from the true joint distribution, effectively aligning the learned posterior with the true posterior.

Inference

Once trained, it can be conditioned on any observation x_{obs} to generate the posterior, sampling from which is as fast as a forward pass through the normalizing flow. In this way, inference can be repeated for many observations without having to regenerate data or retrain the normalizing flow. Similar algorithms have been used in exoplanet retrievals before, and we outline them next.

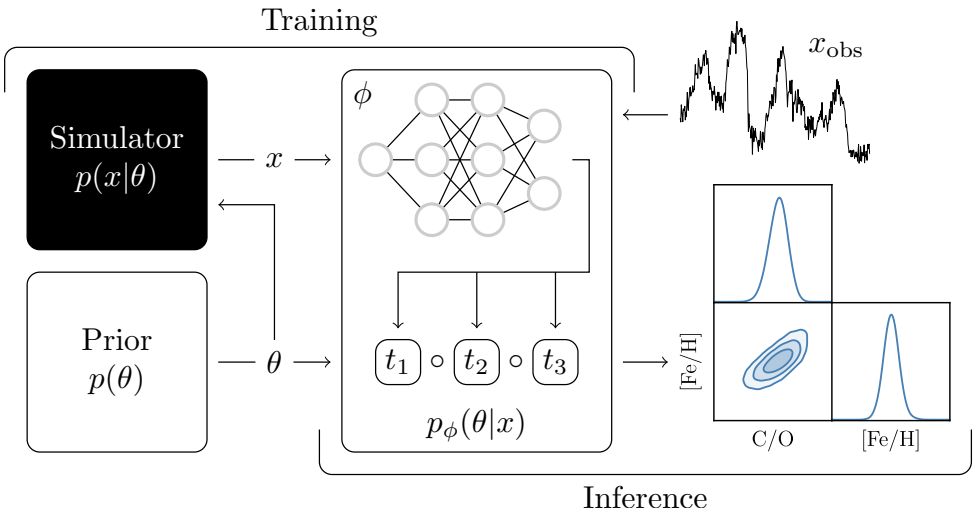


Figure 4.4: Inference pipeline using amortized neural posterior estimation.

4.3 SBI in exoplanet literature

Yip et al. (2022) investigate **variational inference** for atmospheric exoplanet retrieval. They use a similar approach as described above, except that the variational inference is targeted at a single spectrum, using normalizing flows as surrogate models to approximate the posterior distribution. This means that instead of learning a conditional posterior estimator, the flow directly learns the posterior of a specific observation. For this reason, the inference is no longer amortized. The parameters of the normalizing flow are trained by maximizing a variational lower bound on the evidence $p(x)$, provided that the likelihood function associated with the simulation model is both tractable and differentiable.
Pros : It is faster than sequential sampling methods as it reduces the number of likelihood evaluations required.
Cons : It does not provide amortization and needs a likelihood function.

Additionally, Ardévol Martínez et al. (2024) conduct inference using the **sequential neural posterior estimation (SNPE)** algorithm, which is a sequential version of NPE. In this approach, NPE is executed iteratively, where the posterior obtained from each round is used as the prior for the next. This process continues until the posterior distribution converges, i.e., no significant changes are observed between successive iterations.
Pros : It is much faster than NPE, thus making it a better choice for testing initial model adequacy.
Cons : This algorithm is not amortized.

Furthermore, Gebhard et al. (2025) uses **flow matching posterior estimation (FMPE)** to conduct exoplanet retrievals. FMPE algorithm, similar to NPE, uses a normalizing flow to train a posterior estimator; however, it parameterizes the transformations in a continuous way (i.e, not discretely) using continuous normalizing flow (CNF). For more details see Chen et al. (2018). The principle of CNF is that the transformations are modeled by defining it as a time-dependent vector field, which traces the temporal derivative of the trajectory of a sample moving from the base to the target distribution. The network is trained by solving this vector field over time using a standard ordinary differential equation (ODE) solver. This is called flow matching (Wildberger et al., 2023).
Pros : The main advantage of this algorithm is that it provides flexibility in using any neural network to parameterize the vector field. Such lack of architectural constraints offers improved scalability. It has also been shown to be faster than NPE, and in terms of posterior accuracy proven to be on par with NPE and conventional sequential algorithms such as NS and MCMC.

4.4 Amortized SBI, a proof of concept

In order to demonstrate the use of NPE for exoplanet retrievals, we present a proof of concept to characterize the atmosphere of HR 8799 e using its simulated spectrum, and benchmark it with its *MultiNest* retrieval in the further sections.

4.4.1 Scientific context

HR 8799 is a planetary system consisting of an A-type star with four directly imaged planets orbiting it within a massive debris disk at distances from 15 to 70 AU (Wang et al., 2018). Since their discovery, these planets have been extensively studied (Marois et al., 2008, 2010; Currie et al., 2011; Su et al., 2009). Several of these studies suggest that these planets have comparable surface gravities and temperatures, with thick clouds detected in the near-infrared region of the spectrum. To constrain the cloud properties, Mollière et al. (2020) used two different cloud models and, for the first time, performed free retrievals on the combined spectrum of HR 8799 e using VLT interferometer for precision narrow-angle astrometry and interferometric imaging (GRAVITY) data along with archival observations from SPHERE and GPI. The work by Mollière et al. (2020) built upon previous studies (Madhusudhan et al. 2011; Bowler et al. 2010; Marley et al. 2012; Charnay et al. 2018; Lavie et al. 2017). In this proof of concept for the NPE algorithm, we use the same simulated cloudy spectrum of HR 8799 e (see Figure 4.5) used by Mollière et al. (2020) to test the retrieved constraints on its cloud properties, and compare their NS retrieval results with those obtained using NPE. The synthetic spectrum spans a wavelength range from 0.95 to 2.45μm with a continuous wavelength spacing of $\lambda/\Delta\lambda$ =400. As in Mollière et al. (2020), we assume a signal-to-noise ratio of 10 per spectral bin, leading to a

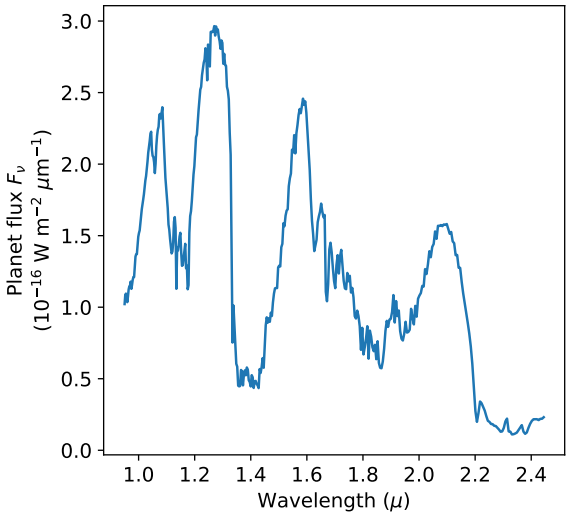


Figure 4.5: Synthetic HR 799 e spectrum.

standard deviation $\sigma = 0.1257e^{-17}$W m^{-2} $^{-}$m^{-1} across all bins for the assumed Gaussian noise. The nominal parameters used to generate the synthetic spectrum are tabulated in Table 4.1. These values were informed by self-consistent atmospheric structures of an exoplanet with identical physical characteristics.

Table 4.1: Parameter values θ_{obs} of the benchmark spectrum x_{obs}.

Parameter	Value	Parameter	Value
T_1	330.6 K	$\log \tilde{X}_{\text{Fe}}$	-0.86
T_2	484.7 K	$\log \tilde{X}_{\text{MgSiO}_3}$	-0.65
T_3	687.6 K	f_{sed}	3
$\log \delta$	-7.51	$\log K_{zz}$	8.5
α	1.39	σ_g	2
T_0	1063.6 K	R_P	1
C/O	0.55	$\log g$	3.75
Fe/H	0	$\log P_{\text{quench}}$	-5

4.4.2 Setup

Radiative transfer simulator

The atmospheric model used in this study, and all others in this thesis, consists of a deterministic atmospheric forward model implemented with `petitRADTRANS`, together with a noise model accounting for measurement noise. (Mollière et al., 2019) is a one-dimensional radiative transfer model that is used to calculate the emission and transmission spectra for exoplanets

with cloudy and cloud-free atmospheres. This model simulates a single atmospheric column consisting of multiple distinct pressure layers, with temperatures determined by a freely parameterized thermal profile. The layers include various opacity sources dispersed throughout the column and are incorporated into the radiative transfer equations.

For this study we use the disequilibrium chemistry emission model predefined in `petitRADTRANS` to compute an emission spectrum based on disequilibrium carbon chemistry, equilibrium clouds, and a spline temperature-pressure profile, defined by 16 parameters in total. We walk through these parameterizations briefly in the following paragraphs.

The temperature structure (illustrated in Figure 4.6) uses both freely variable and physically motivated parameterizations based on atmospheric altitudes. The optical depth defined as $\tau = \delta P^{\alpha}$ is parameterized as a function of the pressure P, δ and α where, the latter two are free parameters. The temperature structure is split into three parts. The mid altitude (photosphere), with an optical depth $\tau > 0.1$, models the temperature according to the Eddington approximation,

$$T(\tau)^4 = \frac{3}{4} T_0^4 \left(\frac{2}{3} + \tau \right)$$

such that, T_0 is a free parameter. In the upper altitude with an optical depth $\tau < 0.1$, the structure is computed by a cubic spline interpolation between free parameters T_1, T_2, and T_3. In low altitudes (troposphere), wherever the atmospheric temperature gradient of the Eddington approximation is greater than the moist adiabatic gradient (i.e, $\nabla_{\rm edd} > \nabla_{\rm ad}$), convection ensues. The $\nabla_{\rm ad}$ is interpolated from a T-P-[Fe/H]-C/O space of a chemical equilibrium table. Within this region, the thermal profile drops linearly with pressure with the adiabatic gradient as $\frac{d \ln P}{d \ln T} = 1/\nabla_{\rm ad}$. Here, the metallicity [Fe/H], and the carbon-to-oxygen number ratio C/O, are also free parameters.

Once the P-T profile is constructed, cloud abundances are calculated in the form of their mass fractions, where they are modified from solar abundances based on the model parameters [Fe/H] and C/O. Silicate and iron clouds are considered, and their respective cloud mass fractions are further scaled with the free scaling parameters $\log \tilde{X}_{\rm Fe}$ and $\log \tilde{X}_{\rm MgSiO_3}$, where $\tilde{X}_i = X_0^i / X_{\rm eq}^i$ is the ratio of the cloud mass fraction X_0^i at the cloud base (i.e., at pressure $P_{\rm base}$) to the mass fraction $X_{\rm eq}^i$ predicted at the same location for the cloud species when assuming equilibrium condensation. The cloud mass fraction decays with altitude based on the free settling parameter $f_{\rm sed}$:

$$X(P) = X_0 \left(\frac{P}{P_{\rm base}} \right)^{f_{\rm sed}} . \tag{4.12}$$

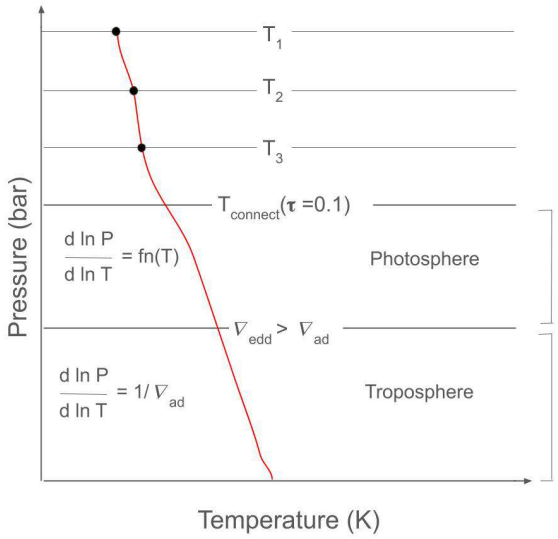


Figure 4.6: Structure of the P- T profile used for the retrieval.

For $P > P_{\text{base}}$ the cloud mass fraction is zero. The other freely defined cloud parameters include the vertical eddy diffusion coefficient K_{zz} and the width of the log normal size distribution σ_g defined in the Ackerman and Marley (2001) cloud model, called Cloud Model 1 in Mollière et al. (2020). Equilibrium chemistry is assumed for the rest of the chemical abundances, which are interpolated from the chemical equilibrium table calculated with `easyCHEM` (Mollière et al., 2017) as a function of T-P-[Fe/H]-C/O. The species considered for this atmosphere are H_2O, CO, CH_4, NH_3, CO_2, H_2S, VO, TiO, PH_3, Na, and K.

The free parameter $\log P_{\text{quench}}$ is used to account for disequilibrium chemistry through atmospheric mixing. For pressures less than P_{quench}, the mass fractions of CH_4, H_2O, and CO are held constant at their values at $P = P_{\text{quench}}$, since these molecules are expected to be susceptible to mixing within the atmosphere of objects with these temperatures (Zahnle and Marley, 2014). The surface gravity ($\log g$) and radius (R_p) of the planet are also considered as free parameters to calculate the emission flux. The radiative transfer equations are then solved using the Feautrier method (Feautrier, 1964) as in the self-consistent `petitCODE` (Mollière et al., 2015, 2017), which also includes isotropic scattering. Following Mollière et al. (2020), we rebinned down the default wavelength spacing $\lambda/\Delta\lambda = 1000$ to a spacing of 400 between 0.95 to 2.45 μm. This was done by generating the binned down correlated-k opacities tables of individual atmospheric absorbers in the resort-rebin fashion (e.g., Mollière et al., 2015; Amundsen et al., 2017), and using them instead of the original opacities to generate linearly binned spectra within the same wavelength range, resulting in vectors of 379 elements. We denote the output spectrum produced by this first simulation stage as $f(\theta)$, where θ contains all 16 model parameters.

Table 4.2: Prior distribution over the model parameters.

Parameter	Prior	Parameter	Prior
T_1	$\mathcal{U}(0, T_2)$	$\log \tilde{X}_{\mathrm{Fe}}$ [b]	$\mathcal{U}(-2.3, 1)$
T_2	$\mathcal{U}(0, T_3)$	$\log \tilde{X}_{\mathrm{MgSiO_3}}$ [b]	$\mathcal{U}(-2.3, 1)$
T_3	$\mathcal{U}(0, T_{\mathrm{connect}})$ [a]	f_{sed}	$\mathcal{U}(0, 10)$
$\log \delta$	$P_{\mathrm{phot}} \sim \mathcal{U}(10^{-3}, 10^2)$ [c]	$\log K_{zz}$	$\mathcal{U}(5, 13)$
α	$\mathcal{U}(1, 2)$	σ_g	$\mathcal{U}(1.05, 3)$
T_0	$\mathcal{U}(300, 2300)$ K	R_P	$\mathcal{U}(0.9, 2)$
C/O	$\mathcal{U}(0.1, 1.6)$	$\log g$	$\mathcal{U}(2, 5.5)$
Fe/H	$\mathcal{U}(-1.5, 1.5)$	$\log P_{\mathrm{quench}}$	$\mathcal{U}(-6, 3)$

Notes. [a] T_{connect} is the uppermost temperature of the photospheric layer that `petitRADTRANS` calculates by setting the optical depth $\tau = 0.1$. [b] Here $\tilde{X}_i = X_0^i / X_{\mathrm{eq}}^i$, where X_{eq} is defined as the mass fraction predicted for the cloud species when assuming equilibrium condensation at the cloud base location. [c] P_{phot} is defined as the pressure where the optical depth $\tau = 1$. The parameter δ is calculated acccordingly.

To account for measurement noise and make the simulation model similar to instrumental data, we consider a Gaussian noise model with identical standard deviation σ for all spectral channels. The spectra $f(\theta)$ generated by `petitRADTRANS` is randomly perturbed with additive noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, where $\epsilon \in \mathbb{R}^L$ with L=379, is a vector of random noise instances in each wavelength bin. Here we assume the same noise variance in each wavelength bin for the sake of simplicity, but more complex noise models (including noise covariance) could be used in our simulator. The final simulator output is given by $x = f(\theta) + \epsilon$.

Prior

We define a 16-dimensional multivariate uniform prior distribution in Table 4.2, with physically motivated ranges for each parameter θ. This prior distribution is the same as the one used by Mollière et al. (2020) and is tuned to represent the specific case of planet HR 8799 e.

Training set

Our training data set is composed of 12 million parameters-spectrum pairs $(\theta, f(\theta))$. It is created by drawing parameters $\theta \sim p(\theta)$ from the prior and passing them through the simulator as shown in Fig. 4.4. We split this dataset into 90 %, 9 %, and 1 % for training, validation, and testing respectively.

Technical details of NPE

We define the posterior estimator $p_\phi(\theta|x)$ as a NAF (Huang et al., 2018) composed of three transformations. We also explored a NSF (Durkan et al., 2020) implementation of the posterior estimator, but in the end implemented a NAF since it gave a lower validation loss. Each transformation of the flow is parameterized by a MLP with five hidden layers of size 512 and exponential linear unit (ELU) activation functions (Clevert et al., 2015). A second network, called the embedding, is used to compress the 379-dimensional spectrum x into a vector of 64 features, which is then used to condition the flow with respect to x. The rationale behind this compression is that it forces the posterior estimator to extract informative features from the spectra instead of memorizing the training data. The embedding network is implemented as a residual MLP (ResMLP), composed of 10 residual blocks (He et al., 2016) of decreasing size (two times 512, three times 256 and five times 128) and also uses ELU activation functions. The ResMLP is followed by the application of the Softclip function. The Softclip function transforms the input flux values into a bounded range between 0 and 100, projected through an S-shaped curve. This ensures that small changes in extremely low or high values result in negligible changes in the output, whereas moderate input flux values produce significant output changes. The Softclip helps to manage the large variations in flux present in the training set.

Before training, random noise realizations are added on-the-fly to the spectra to obtain observations $x = f(\theta) + \epsilon$. Following Eq. 4.11, the flow and embedding networks are trained jointly to minimize the expected negative posterior log-density over the training set. The optimization is carried out through a variant of stochastic gradient descent, namely AdamW (Loshchilov and Hutter, 2017). To improve training without overfitting we use an initial learning rate of 10^{-3} that halves every time the average loss over the validation set does not improve for the last 32 epochs, until it reaches 10^{-6} (Zhang et al., 2021). We also use a high weight decay of 10^{-2}. We train for a total of 1024 epochs during each of which 1024 random batches of 2048 pairs $(\theta, f(\theta))$ are taken from the training set.

The architectural hyper-parameters were adjusted on validation data. We performed extensive hyper-parameter tuning on the flow and embedding network parameters. For the flow, we explored different numbers of transforms and hidden layer dimensions in the range of [3,5] and [256,512], respectively. For the embedding network, we tried different numbers of layers in the ResMLP in the range of [10, 15]. We also explored different activation functions like rectified linear unit (ReLU) and ELU for both networks. We explored different values for the initial learning rate and the minimum learning rate in the ranges of [10^{-5}, 10^{-3}] and [10^{-6}, 10^{-5}], respectively. We analyzed the impact of different schedulers like `ReduceLROnPlateau` and `CosineAnnealingLR`, available in PyTorch, with patience rates between [8,32]. We tried batch sizes between [2^8, 2^{11}] and the number of epochs between [2^8, 2^{10}]. We tuned each hyper-parameter by randomly searching over a grid within their range mentioned above, and studied their impact over $\sim$80 runs in parallel. We selected those that led to lower validation loss

and/or more stable training. Amongst all the parameters that we tuned, the parameter weight decay between $[0, 10^{-2}]$ had the most significant impact on the training. We think this is because of the high variance of the input dataset, where some spectra are six orders of magnitude brighter than the rest. This leads to the skewing of the weights to very high values, which is compensated by weight decay to improve training performance. These technical details are tabulated in Table 4.3. For more details, we refer to the source code of the experiments [*].

Once the architecture is finalized, the posterior log-density $\log p_\phi(\theta|x)$ is computed during each forward pass of training the flow, where ϕ is the set of weights of the neural flow network, and the loss function $\mathbb{E}_{(\theta,x)\sim p(\theta,x)}\left[-\log p_\phi(\theta|x)\right]$ on ϕ is minimized over the whole training set (θ, x) (Papamakarios et al., 2021).

Table 4.3: Technical details of the posterior estimator and training

Flow and embedding		Tuned	
Neural architecture	Details	Hyperparameter	Value
Normalizing flow	NAF	Optimizer	AdamW
Flow transforms	3	Initial learning rate	1×10^{-3}
Transform	MLP	Scheduler	ReduceLROnPlateau
Signal	16	Patience	32
Hidden features (transform)	$5 \times [512]$	LR reduction factor	0.5
Activation (transform)	ELU	Minimum learning rate	1×10^{-6}
Embedding architecture	Residual MLP	Weight decay	1×10^{-2}
Embedding depth	$[512] \times 2 + [256] \times 3 + [128] \times 5$	Batch size	1024
Embedding output	$379 \to 64$	Loss	NPELoss
		Epochs	1024

4.4.3 Results

As a demonstration of atmospheric retrieval with NPE, we present inference results for a synthetic exoplanet spectrum x_{obs} generated with the parameter values θ_{obs} given in Table 4.1, similarly to the benchmark retrieval of Mollière et al. (2020).

Retrieval results are summarized in Fig. 4.7. The corner plot shows 1d and 2d marginal posterior distributions obtained for the benchmark spectrum x_{obs}. The marginal posterior distributions were approximated by sampling sufficiently many times the joint posterior distribution from the normalizing flow, which takes only a few seconds to obtain a smooth corner plot. We observe that the bulk of the marginal posterior distributions (in blue) is centered around the parameter values θ_{obs} (in black) used to generate the spectrum x_{obs}. The figure also illustrates the spread in the posterior P-T profiles with respect to the synthetic observation spectrum.

[*] `https://github.com/MalAstronomy/sbi-ear`

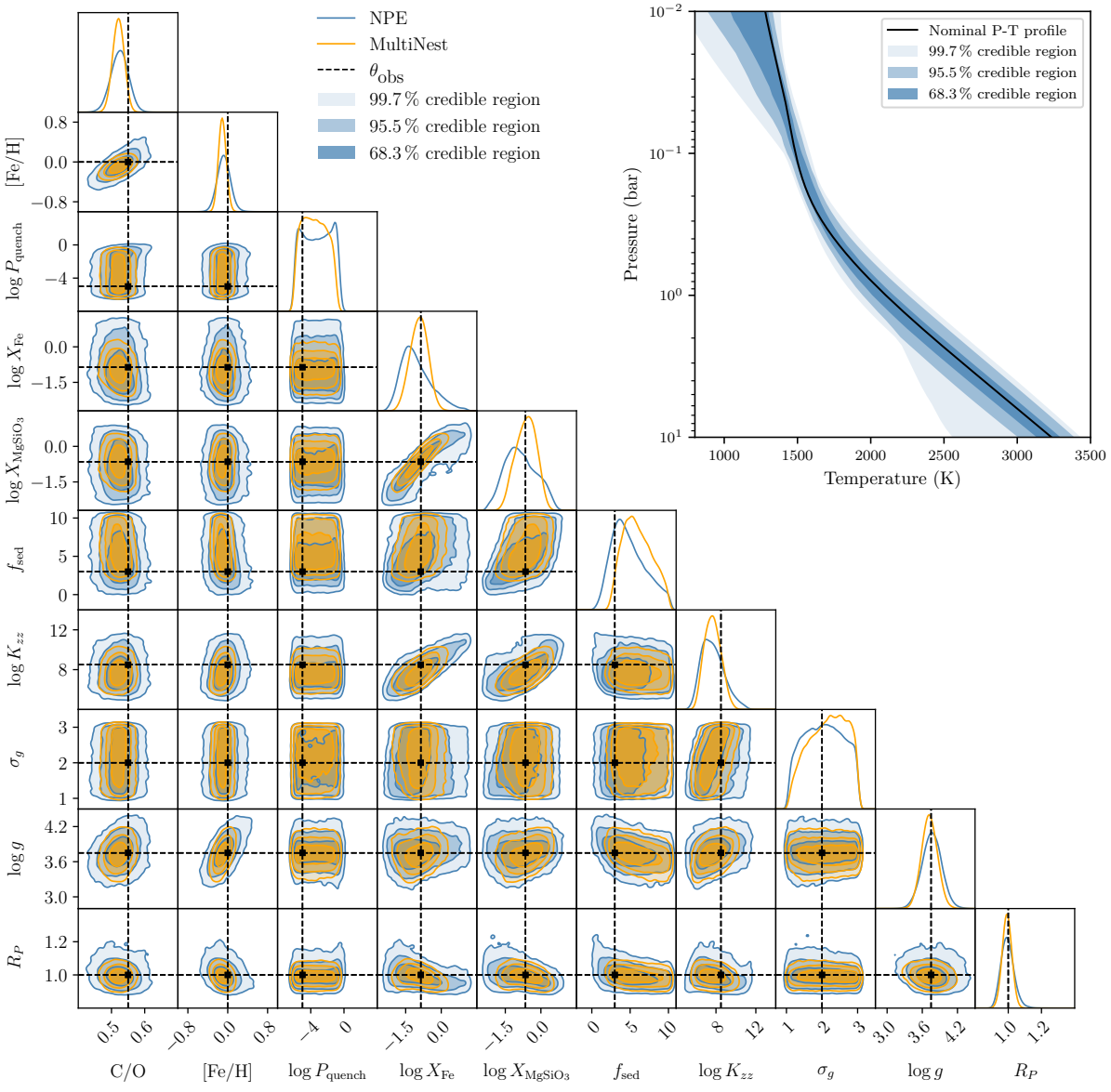


Figure 4.7: Benchmark retrieval using neural posterior estimation. The corner plot shows 1d and 2d marginal posterior distributions obtained for the benchmark spectrum x_{obs} for NPE (in blue) and for nested sampling (in orange). We observe that the nominal parameter values θ_{obs} (in black) are well identified. The top right figure illustrates the posterior distribution of the P-T profiles.

More specifically, we computed posterior P-T profiles for $\theta \sim p_\phi(\theta|x_{\rm obs})$, and show their 68.3 %, 95.5 % and 99.7 % credible regions. We see that the P-T profile for $\theta_{\rm obs}$ (in black) is constrained mostly within the first credible region of the posterior. These results lead us to believe that the NPE posterior approximation produces coherent posterior distributions.

4.4.4 Validation

In Fig. 4.7, we compare the NPE posteriors with those obtained using `MultiNest` (Feroz and Hobson, 2008; Feroz et al., 2009, 2019b; Buchner et al., 2014) for the same noisy synthetic observation. While the results obtained with NPE appear to be coherent with respect to the nominal parameter values $\theta_{\rm obs}$ and the posterior P-T profiles, we see that the approximate marginal posterior distributions computed with `MultiNest`, using a sampling efficiency of 0.8 (recommended for parameter estimation) with 4000 live points, are slightly less dispersed than for NPE. On performing several retrievals with different noise realizations (not shown here), it is seen that, each time, the peaks of the individual parameter posterior distributions shift in a similar way in both retrieval algorithms. This can be seen here in the parameters C/O and Fe/H, similarly shifted slightly to the left. This suggests that these shifts are directly related to the particular noise realization, and that `MultiNest` and NPE behave in a similar way in the presence of noise.

The difference in the posterior widths for the two algorithms motivates a thorough investigation of the computational faithfulness of the NPE posterior approximations using inference diagnostics, including posterior predictive checks and coverage. We take advantage of the quasi-instantaneous inference of NPE to perform these checks.

Consistency plot

The fit of the posterior to the observation can be qualitatively assessed using the posterior predictive distribution $p(x^{'}|x_{obs})$. This represents the distribution of the possible future observations $x^{'}$, given the current observation x_{obs} and the model parameters θ. It is calculated as,

$$p(x'|x_{obs}) = \int p(x'|\theta)p(\theta|x_{obs})\, d\theta$$

where $p(x'|\theta)$ is the likelihood of the new data given the model parameters, and $p(\theta|x_{obs})$ is the estimated posterior distribution. The posterior predictive distribution can be obtained by sampling parameters from the posterior and computing their spectra using the simulator model. One evaluates the quality of the fit by visually comparing the consistency of the posterior predictive distribution with the observation. This is called a consistency plot. If the model is adequate, the data is rich and the observation is not too noisy, the posterior predictive distribution is almost centered around the observation, and narrow, and vice versa.

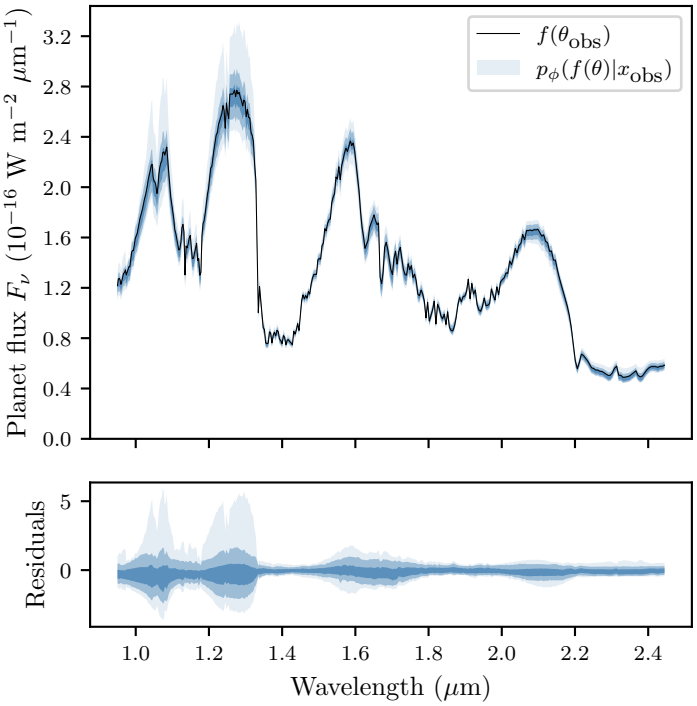


Figure 4.8: *Top.* Posterior predictive distribution $p(f(\theta)|x_{\rm obs})$ of noiseless spectra (without the instrumental noise disturbance ϵ) for the 99.7%, 95% and 68.7% quartiles (hues of blue), overlaid on top of the noiseless observed spectrum $f(\theta_{\rm obs})$ (black line). *Bottom.* Residuals of the posterior predictive samples, normalized by the standard deviation of the noise distribution for each spectral channel.

We perform the qualitative examination of the posterior predictive distribution, here defined as $p_\phi(f(\theta)|x_{\rm obs})$, for spectra without instrumental noise disturbance, which we obtained by sampling parameters from the posterior, $\theta \sim p_\phi(\theta|x_{\rm obs})$, and then computed their spectra $f(\theta)$ with `petitRADTRANS`. Fig. 4.8 shows the consistency plot where we plot the posterior predictive distribution $p(f(\theta)|x_{\rm obs})$ for various quartiles against the noiseless version of the observed spectrum $f(\theta_{\rm obs})$. We observe that (i) the posterior predictive distribution is well constrained, with the 68% quartile distribution mostly within the 1σ noise limit as expected, and (ii) that $f(\theta_{\rm obs})$ is relatively well centered inside the 68% quartile along all bins. Had the posterior distribution $p_\phi(\theta|x_{\rm obs})$ been too wide, we would have observed a much larger spread. Had the bulk of the posterior density been at the wrong place, we would not have observed $f(\theta_{\rm obs})$ to be well inside the distribution. These reassuring diagnostics are a first indication of the good quality of the inference results obtained with NPE. In particular, they demonstrate that the cloud parameter distributions derived by NPE produce spectra consistent with the observed spectrum. In Fig. 4.9, we further demonstrate that the parameter values sampled from the somewhat wider NPE cloud posteriors are actually all leading to cloudy solutions, in good agreement with the synthetic input observation. In this figure, we sampled parameters from

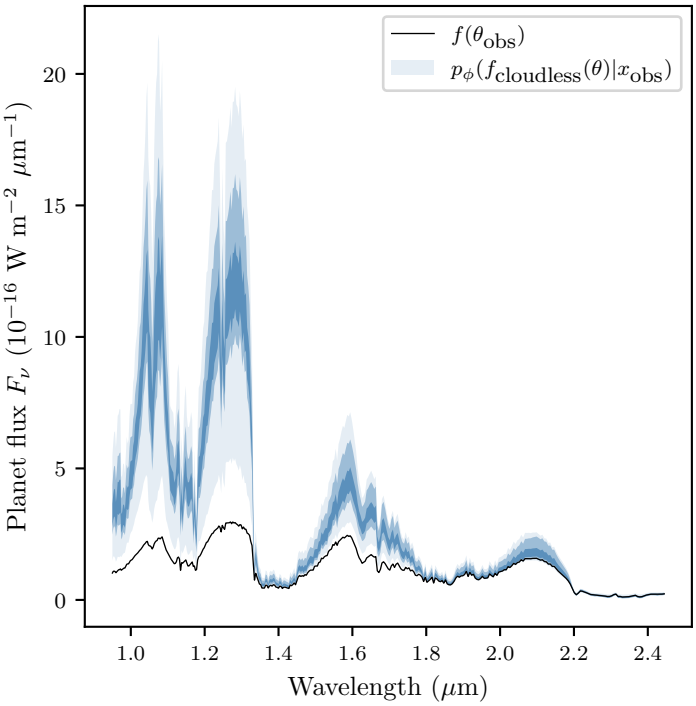


Figure 4.9: Cloudless realizations of the posterior predictive distribution $p(f_{\text{cloudless}}(\theta)|x_{\text{obs}})$ overlaid on top of $f(\theta_{\text{obs}})$, where $f_{\text{cloudless}}$ artificially sets the cloud scaling factors $\log X_{\text{Fe}}$ and $\log X_{\text{MgSiO}_3}$ to a very small value of -10.

the (cloudy) approximate posterior, but then artificially turned off the clouds, by setting the log of the cloud mass fraction scaling factors X_{Fe} and X_{MgSiO_3} to -10, to assess their impact on the spectral shape. We see that these cloudless spectra look significantly different from the cloudy ones shown in Fig. 4.8. This implies that the posterior predictive distribution samples in Fig. 4.8 are indeed affected by clouds. However, since the retrieval was performed on a cloudy model and not a cloud-free one, this result is not conclusive.

Coverage

Following Hermans et al. (2020), we further evaluate the global computational faithfulness of the NPE posterior approximations in terms of expected coverage. We define the expected coverage probability of the $1-\alpha$ highest posterior density regions derived from the posterior $p_\phi(\theta|x)$ as

$$\mathbb{E}_{p(\theta,x)}\left[\mathbb{1}\left(\theta \in \Theta_{p_\phi(\theta|x)}(1-\alpha)\right)\right], \tag{4.13}$$

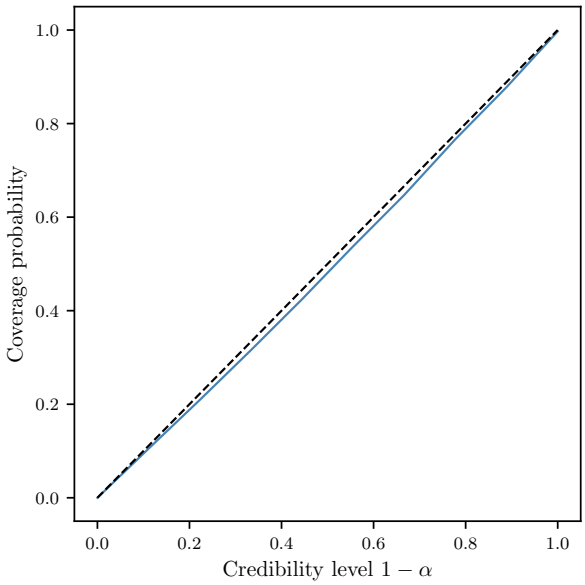


Figure 4.10: Coverage plot assessing the computational faithfulness of $p_\phi(\theta|x)$ in terms of expected coverage. The coverage probability is close to the credibility level $1-\alpha$, which indicates that the posterior approximations produced by NPE are neither significantly overdispersed (the coverage curve would otherwise be above the diagonal) nor significantly underdispersed (the coverage curve would be below the diagonal).

where $\mathbb{1}(\cdot)$ is the indicator function, and where the function $\Theta_{p_\phi(\theta|x)}(1-\alpha)$ yields the $1-\alpha$ highest posterior density region of $p_\phi(\theta|x)$. This diagnostic probes the consistency of the posterior estimator $p_\phi(\theta|x)$ and can be used to assess whether the approximate posterior distributions are overdispersed or underdispersed on average. In other words, it is an evaluation of how realistic the model uncertainty is.

It is estimated by repeatedly sampling (θ, x) from the prior and the simulation models, and then running NPE retrievals on each x. If the posteriors are well calibrated, then the parameter values θ that were used to generate the spectra x should be contained in the $1-\alpha$ highest posterior density regions of the approximate posteriors $p_\phi(\theta|x)$ exactly $(1-\alpha)\%$ of the time. If the coverage probability is smaller than the credibility level $1-\alpha$, then this indicates that the $1-\alpha$ highest posterior density regions are smaller than they should be, which is the sign of overconfident and usually unreliable posterior approximations. On the other hand, if the coverage probability is larger than the credibility level $1-\alpha$, then this indicates that the highest posterior density regions are wider than they should be. In this case, the posterior approximations are said to be conservative. We argue that posterior approximations should rather be conservative to guarantee reliable and meaningful inferences, even when the approximations are not faithful. Indeed, wrongly excluding plausible parameter values of exoplanet spectra could lead to wrong conclusions about the actual nature of the exoplanet, while failing to exclude actually implausible parameter values would only result in a loss of statistical power.

Figure 4.10 summarizes the expected coverage of $p_\phi(\theta|x)$ for credibility levels from 0 to 1. The coverage curve closely fits the diagonal, which indicates that the posterior distributions produced by NPE are well calibrated – even though we note a trend for the posteriors to be very slightly underdispersed.

Unfortunately, running the same coverage diagnostic for a sampling algorithm such as MCMC or nested sampling is not possible within a reasonable computation time, since it requires the repeated construction of posterior distributions over many distinct random realizations x in order to approximate the expectation in Eq. 4.13. For this reason, we cannot conclude whether `MultiNest` is computationally faithful in terms of expected coverage. However, given that the approximate posterior distribution produced by `MultiNest` in Fig. 4.7 is slightly narrower than the NPE posterior distribution, it suggests that `MultiNest` is slightly more underdispersed than NPE. This conclusion is in line with the analysis of `MultiNest` posterior distributions in Ardévol Martínez et al. (2022b), where 4000 retrievals were performed on simulated observations using a convolutional neural network (CNN) and `MultiNest`, which on comparison suggests that `MultiNest` tends to underestimate the uncertainties of the parameter it retrieves.

Local classifier two-sample test (L-C2ST)

Since coverage is a global validation method, it serves as a necessary but not sufficient condition for a valid inference algorithm. A coverage check that fails indicates that the inference is invalid, while passing coverage checks does not guarantee that the posterior estimation is accurate. This is because coverage deals with cumulative probability and does not account for local inconsistencies. This limitation motivates the use of a local validation procedure called the local classifier two-sample test (L-C2ST, Linhart et al., 2024). L-C2ST allows for the local evaluation of a posterior estimator at any given observation. In case of an inconsistency, L-C2ST is also able to graphically show how to improve the estimator.

The test involves training a classifier to distinguish between samples drawn from the true joint distribution $p(\theta, x)$ (class 0) and those drawn from the estimated posterior $q(\theta|x)p(x)$ (class 1). For a normalizing flow, this translates to learning to differentiate samples from $\mathcal{N}(0, \mathbf{I}_m)p(x)$ (class 0) and $p(T_\phi^{-1}(\theta; x_o)|x_o)$ (class 1) where T_ϕ is the forward transform of the flow. The classifier is trained under the null hypothesis, which asserts that the two distributions are indistinguishable. Formally, the null hypothesis under the normalizing flow (NF) is expressed as:

$$H_{NF,0}(x_{\text{obs}}) : \; p\left(T_\phi^{-1}(\theta; x_{\text{obs}})|x_{\text{obs}}\right) = \mathcal{N}(0, \mathbf{I}_m)$$

where, x_{obs} is any observation over which one is evaluating the quality of the retrieval.

The classifier is also trained once on the observed data where the training set retains the relationship between variables to account for real-world variability. Finally, the classifier is evaluated on a single observation x_{obs} based on metrics that rely on the L-C2ST statistic. The statistic is the mean squared error (MSE) between 0.5 and the predicted probabilities from the classifier of being in class 0 over the dataset $(\theta_{\text{obs}}, x_{\text{obs}})$.

The first metric is the p-value, which is the proportion of the times the L-C2ST statistic under the null hypothesis is greater than the L-C2ST statistic at the observation x_{obs}. This is computed as the empirical mean over statistics obtained from several trials under the null hypothesis:

$$p\text{-value} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(L_{\text{C2ST}}^{(i)} \geq L_{\text{C2ST}}(x_{\text{obs}}) \right),$$

where, $L_{\text{C2ST}}(x_{\text{obs}})$ is the L-C2ST statistic at the observation x_{obs}, $L_{\text{C2ST}}^{(i)}$ are the statistics under the null hypothesis, and $\mathbb{I}(\cdot)$ is the indicator function. The distribution under the null hypothesis is called the T-distribution. If the posterior estimate is not close to the true posterior, the classifier will identify a significant difference between the two classes, resulting in higher values of the statistic and hence very small p-values. If this value is less than the assumed significance level, it indicates that the null hypothesis does not hold. Further, the pp-plot, a variation of the coverage plot, helps assess the overall trend of bias or the potential over- or under-confidence of the estimated posterior. This validation tool is applied to an NPE retrieved posterior in Section 5.1.

4.4.5 Computational cost

One of the main advantages of NPE is its amortization of the inference procedure. Once trained, inference does not require on-the-fly simulations and can be repeated several times with different observations at very low computational cost. We demonstrate the true potential of NPE by performing 1000 retrievals and comparing how long it would take for `MultiNest` to produce comparable results. The 1000 observations were produced by randomly sampling parameters values θ from the prior distribution and rendering them through the forward simulation model to produce $x \sim p(x|\theta)$. We then retrieved their corresponding approximate posterior distributions $p_\phi(\theta|x)$. A single retrieval consists of sampling 30740 posterior parameter vectors (as many as `MultiNest` yields) and rendering the corner plot, took respectively 6 and 10 seconds in average. In total, 1000 retrievals would take approximately 4.5 hours. With the upfront generation of the dataset (17 hours on 1000 CPUs) and the training of the neural network (13 hours on a standard NVIDIA GTX 1080 Ti GPU), we reach a total computing time of 34.5 hours. By contrast, each retrieval with `MultiNest` takes around 134 hours on a cluster of 440 CPUs (totaling about 60000 CPU hours) so that retrieving atmospheric parameters on 1000 spectra

would require an extrapolated time of 134000 hours (15 years). In summary, NPE is around 4000 times faster for a thousand retrievals, and almost 30000 times faster if we do not take the upfront generation and training into account. An example of amortization over real data is shown in section 7.2.

It is important to note that the computational speedup comes with the overhead cost of building the training set (one per atmospheric model) and training NPE on it. In our case, simulating a single parameter-spectrum pair $(\theta, f(\theta))$ took around 5 seconds, which results in a total of 17000 CPU hours for the generation of the 12 millions pairs used in this study. The actual wall-clock time, however, can be largely reduced by simulating the pairs in parallel on a large computing cluster, contrary to the on-the-fly and sequential simulations required in MCMC or nested sampling methods. In our case, the training set was generated in less than 17 hours using a cluster of 1000 CPUs. Generating as many samples may not be necessary in all cases, since sufficiently good performance is likely to be possible from smaller training sets. Instead, the main challenge with amortized inference will be to identify a simulation model that is general enough to be applicable and valid in many situations, so that the whole training process does not need to be repeated for each individual case. This may be possible for studies focusing on specific classes of planets, such as hot Jupiters observed in transit, or self-luminous giant planets observed with direct imaging. We also note that, when performing retrievals on a single object, a given training data set can potentially be reused several times when exploring various levels of wavelength binning or different noise models in the retrieval. In this case, only the cost of the NPE training needs to be paid several times. This is usually common since brighter objects have a different measurement noise compared to the dimmer ones. Although, since we often scale our observation noise, the noise model can be generalized by introducing scaling factors as free training parameters over standard noise models, for similar objects (by spectral type or brightness).

4.5 Conclusion

In this chapter, we present a simulation-based inference algorithm called NPE to perform Bayesian retrievals of exoplanet atmospheres. Unlike the commonly used nested sampling and MCMC methods, which perform sequential sampling to construct a joint posterior of all model parameters using an explicit and tractable likelihood function, NPE relies on normalizing flows to estimate the posterior in an amortized way, without requiring an explicit or tractable likelihood. This offers several benefits over standard algorithms.

First, NPE is time-efficient due to amortization. The inference network needs to be trained only once, and the same network can be used to perform quasi-instantaneous retrievals over several observations without starting from scratch. We demonstrated this by performing 1000 retrievals with synthetic observations constructed by sampling randomly from the prior. This proced-

ure took 34.5 hours in total, leading to a computational speed-up of 4000 over `MultiNest`. This gain in speed makes it computationally feasible to retrieve thousands of spectral observations we expect to receive from JWST and future missions such as ARIEL, making it a viable tool for population studies. Additionally, the initial overhead cost of simulations was around 17 hours, which can be easily compensated as the number of observations increases. In the case where several simulation models f need to be tested for the retrieval on a given observation, NPE still provides a speed-up of over a factor of four (134/30). When sufficiently parallelized, `MultiNest` matches the computational speed of a single NPE retrieval. This enables the exploration of several different simulation models over limited observations in a reasonable time, thus making NPE a viable retrieval option with an added advantage of amortization.

Second, NPE is scalable. Since the inference network is trained on the parameters of interest only, performance does not deteriorate as quickly as sampling-based algorithms that must navigate the full joint posterior over both the parameters of interest and the nuisance parameters. This is especially important for future simulation models that are likely to include a large number of nuisance parameters.

Lastly, NPE is testable. Since the inference of many observations takes only seconds to perform, one can easily check the validity of NPE by performing posterior predictive checks and producing coverage plots, which are almost impossible to achieve in the case of sequential algorithms. The results presented in this study confirm that NPE provides computationally faithful posteriors, without any simplifying assumption on the shape of the posterior, yet with a possible sign of being slightly under-dispersed. While such tests cannot be performed with `Multinest` to provide a fair comparison, our mock retrievals suggest that NPE may be less under-dispersed and more faithful than `Multinest`. The prospect of subjecting these retrievals to evaluation metrics such as posterior predictive checks and coverage plots ensures a statistical rigor to the associated results. This establishes NPE as a robust algorithm to perform time-efficient retrievals in the future.

Chapter 5

Characterizing WISE 1738 from its JWST/MIRI spectrum, a cloud-free approach

In this chapter, we present the first scientific application of NPE. We study the atmosphere of the Y0 spectral standard WISE 1738 for the first time in the mid-infrared region using data from JWST/MIRI. This allows us to probe higher in the atmosphere than previous observations. In order to achieve an accurate characterization of the WISE 1738 atmosphere, we perform a cloud-free retrieval on the combined near and mid-infrared data from the HST/WFC3, GNIRS/Gemini and JWST/MIRI. We constrain the brown dwarf's physical properties such as mass, radius and surface gravity, age and find consistency with the evolutionary models. We also constrain the primary molecular abundances usually seen in brown dwarf atmospheres such as H_2O, CO_2, NH_3 and CH_4 and find evidence of disequilibrium chemistry from the constraints on CO and CO_2 which are expected to be depleted under theoretical equilibrium assumptions. We compute super-solar C/O and metallicity in their atmosphere and find a vague interpretation of the formation pathway using the tracer $^{14}N/^{15}N$ ratio. Finally, we validate this retrieval using the tools described in the previous chapter.

5.1 Context

Among over 20 confirmed Y dwarfs (Kirkpatrick et al., 2019b), WISE 1738 was one of the first ultracool dwarfs identified using the Wide-field Infrared Survey Explorer (WISE; Kirkpatrick et al., 2011). Its discovery was announced by Cushing et al. (2011), with a spectrum obtained using the infrared channel of the WFC3 (Kimble et al., 2008) onboard the HST at a resolution of approximately 130. With a calculated temperature of approximately 400 K, it lies at the boundary between T and Y dwarfs. The steep flux drop in the blue wing of the *H*-band, associated with NH_3 absorption at 1.49 μm, led to its classification as a Y0 spectral standard (Cushing

et al., 2011). Since then, it has been extensively studied in the near-infrared. Early atmospheric characterization was performed by comparing the data to self-consistent radiative-convective-thermochemical equilibrium models (e.g., Allard et al., 1996; Marley et al., 1996; Tsuji et al., 1996; Burrows et al., 2001). An early spectral fit by Schneider et al. (2015) using cloud-free and chloride, sulphide, and water cloud models from Saumon et al. (2012) and Morley et al. (2012, 2014) on HST/WFC3 data, did not achieve satisfactory agreement between models and observations. The discrepancies were attributed to the assumption of equilibrium chemistry, which can significantly affect estimates of effective temperature, surface gravity, and cloud properties (Phillips et al., 2024; Mukherjee et al., 2024b).

Subsequent higher-resolution ($\lambda/\Delta\lambda \approx 2800$) near-infrared spectra from instrument GNIRS (Elias et al., 2006a) at the Gemini Observatory highlighted the critical role of disequilibrium chemistry (Leggett et al., 2015, 2016b, 2017) based on model comparisons with Saumon et al. (2012); Tremblin et al. (2015); Morley et al. (2012, 2014). Additionally, light curve variability (approximately 3%) observed in the Spitzer 4.5 μm band, corresponding to a rotation period of $6.0\pm$ 0.1 hours, was attributed to patchy KCl and Na_2S clouds (Leggett et al., 2016a). Cloud-free models generally still fit better for this source, specifically models with $T_{\rm eff}$ between 400 K and 425 K, vertical mixing with eddy diffusion coefficient K_{zz} = 10^6 cm s^{-2} and log g = 4.0, with solar and super-solar metallicity ([m/H] = $+0.2$) (Leggett et al., 2016b).

Zalesky et al. (2019) conducted free retrievals on the HST spectrum of WISE 1738 and used abundances from the retrievals to compare with grids to make inferences about disequilibrium chemistry. The retrieved *P-T* profile was consistent with radiative-convective equilibrium and suggested a cloud-free atmosphere, even though the *P-T* profile intersected Na_2S and KCl condensation curves within the photosphere and showed water condensates at higher altitudes. However, the retrieved parameters were inconsistent with evolutionary models, indicating unusually high surface gravity and mass.

Here, we investigate the atmosphere of the Y0 spectral standard WISE 1738 using, for the first time, a mid-infrared spectrum obtained with JWST's MIRI instrument, alongside near-infrared data from HST/WFC3 and Gemini/GNIRS.

5.2 Observations and data processing

JWST/MIRI spectrum

The MIRI instrument onboard the JWST (Wright et al., 2015, 2023) includes the Medium Resolution Spectrometer (MRS, Wright et al., 2023), which provides medium-resolution spectroscopy over the mid-infrared wavelength range of approximately 5 to 28 μm (Argyriou et al., 2023). The presented MRS data are part of the guaranteed time observation (GTO) program "MIRI

Spectroscopic Observations of Brown Dwarfs" under the observation ID: 1278 led by Pierre-Olivier Lagage. The MIRI/MRS data of WISE 1738 were obtained on July 18th 2023 at 17:44:58 UTC. The observation ran in the FASTR1 read out pattern and the two point dither pattern with four exposures, one integration per exposure and 110 groups per integration.

The time between each frame was 2.78 sec, as well as the time between groups, resulting in a total exposure time of 610.5 sec. During the mid-time of the exposure, the telescope pointed at RA 17^{h} 38^{m} 24^{s} and DEC $+27^{\circ}$ $30'$ $0''$.

The data were downloaded from the Mikulski Archive for Space Telescopes (MAST, DOI: 10.17909/vzcg-p593) and processed through the MIRI pipeline. The pipeline (Bushouse et al., 2023) consists of multiple stages, where the first stage includes the ramp fitting to calibrate the raw data to flux units. The second stage assigns a world coordinate system, applies a flat field, straylight, fringe and photometric correction. After this step, the background is removed by subtracting the two dithers from each other (for details, see Barrado et al. 2023). The last stage converts the detector data set to a 3D cube using the 'drizzle' weighting algorithm after assigning a coordinate system to it and running an outlier detection to flag remaining bad pixels. In the end, the spectrum is extracted by centring an aperture at the source with a radius of one full width at half maximum (FWHM) of the point spread function (PSF) of the object. An additional one dimensional fringe correction in the extraction function is performed to reduce additional fringing effects. The data reduction was done using the JWST pipeline version 1.12.5, CRDS version 11.17.10 and context file version jwst_1149.pmap.

The output MIRI spectrum was obtained from channels 1A to 3C, ranging between 4.9 μm and 17.9 μm. Each channel was rebinned to the `petitRADTRANS` wavelength spacing of $\lambda/\Delta\lambda = 1000$. Accounting for the overlaps at the edges of each channel by averaging over the repeated wavelengths, and stitching them all together, resulted in the final vector used for retrievals. The spectrum is shown in Figure 5.1 in black. Visual inspection already suggests the unambiguous presence of molecules such as CO, H_2O, CH_4, NH_3 and CO_2 as indicated in the figure.

HST/WFC3 spectrum

The near-infrared spectrum for WISE 1738 was obtained from the data observed on the infrared channel of the WFC3 (Kimble et al., 2008) on-board the HST, as a part of its Cycle 18 program (GO-12330, PI: J. D. Kirkpatrick) in 2011. Further details on how it was acquired can be found from the discovery paper by Cushing et al. (2011). The HST spectrum covers a wavelength range of 1.07–1.70 μm at a resolving power of approximately 130. The spectrum was obtained from Figure 8 of Schneider et al. (2015).

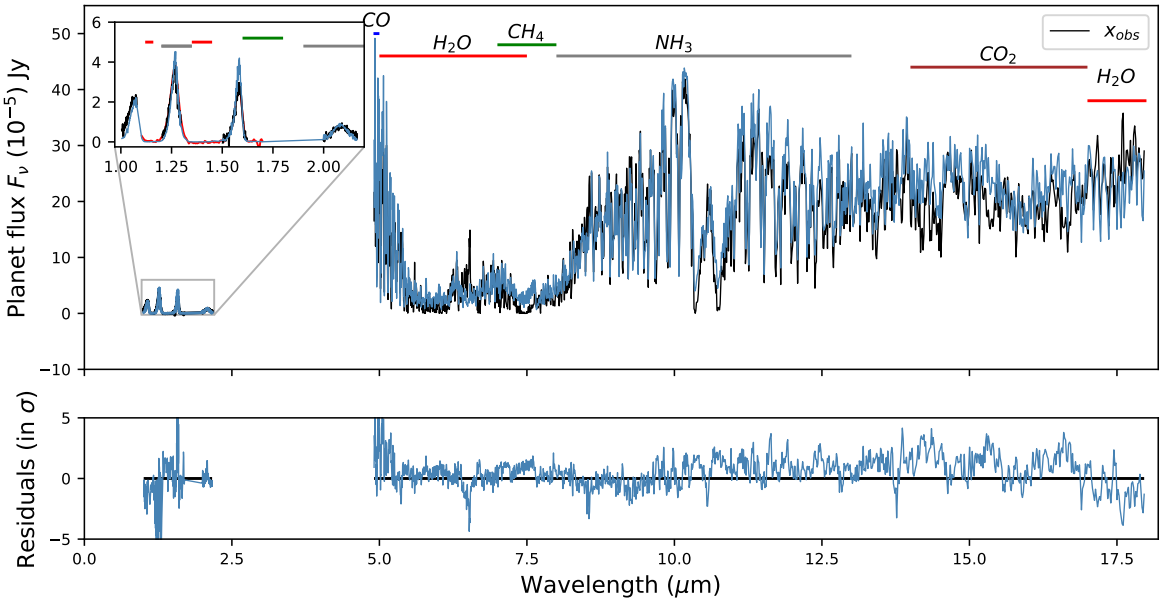


Figure 5.1: *Top.* WFC3 (J and H bands, red), GNIRS (Y, J, H and K bands, black), and MIRI (black) observations x_{obs}, overlaid with the simulated noiseless spectrum $f(\theta)$ associated with the most probable parameters from the posterior. *Bottom.* Residuals of the sample normalized by the inflated standard deviation of the noise distribution for each spectral channel.

Gemini/GNIRS spectrum

A higher-resolution near infrared spectrum for WISE 1738 was also obtained from the data obtained by the Gemini North telescope using the GNIRS (Elias et al., 2006b), at a resolving power of approximately 2800, via the program GN-2014A-Q-64. Further information on this can be found in the work by Leggett et al. (2016b). The spectrum covers a wavelength range of 0.993–1.087 μm, 1.191–1.305 μm, 1.589–1.631 μm and 1.985–2.175 μm, i.e., spanning the Y, J, H and K bands. For the retrievals in this work, this spectrum is rebinned down to a resolution $\lambda/\Delta\lambda = 1000$ to match the default resolution of `petitRADTRANS`.

5.3 Setup, cloud-free approach

5.3.1 Radiative transfer simulator

This experiment is setup similarly to that in Section 4.4.2, with a parameterized thermal profile and the addition of main absorbers suspected in the atmosphere. The version of `petitRADTRANS` used here is 2.6.3.

Here we adopt this setup to model a cloud-free atmosphere. This is because, although Line et al. (2017) found an alkali depletion trend consistent with the theoretical trend of Na_2S and KCl condensation curves intersecting with the thermal profile, previous studies on late-T (Line et al., 2017) and early Y dwarfs (Zalesky et al., 2019) have found no strong evidence for the presence of optically thick Na_2S and KCl clouds, which is contrary to expectations for cooler brown dwarfs (Kirkpatrick, 2005). Further, so far Y dwarfs have been well characterized without the need for water or ammonia clouds (Kühnle et al., 2024; Barrado et al., 2023; Zalesky et al., 2019).

The cloud-free model is parameterized using 26 parameters, denoted as θ. The *P-T* profile is calculated on a pressure grid containing various levels between 10^{-6} and 1000 bar. Within this grid, 10 nodes are equidistantly defined in log-space. The temperature at the bottom-most node is set as a free parameter T_{bottom}, such that it can take any value uniformly distributed between 100 to 9000 K. The temperature at each upper node is calculated as a parameterized fraction uniformly distributed between 0.2-1.0 of the temperature at the node immediately below it. These 9 node fractions are defined as $T_{\text{nodes[i-ix]}}$. Once the temperatures at all the nodes are calculated, the entire profile is constructed by quadratically interpolating between them.

The primary absorber species typical in Y-dwarf atmospheres with an effective temperature (T_{eff}) of approximately 400 K—such as CH_4, H_2O, H_2S, CO_2, CO, and NH_3—are considered (see Figure 5 in Leggett et al., 2015), along with the isotopologue $^{15}NH_3$ (Barrado et al., 2023). Our model additionally incorporates HCN and PH_3 (Visscher et al., 2006; Zahnle and Marley, 2014) and TiO and VO, the latter two of which are notably present in higher temperature brown dwarfs, such as late-type M dwarfs (Kirkpatrick et al., 1999). Sources of continuum opacity such as H_2-H_2 and He-H_2 collision-induced absorption bands are also considered. The logarithm (of base 10) of the abundance of each opacity species, expressed as mass fractions, is treated as a free parameter. Brown dwarf atmospheres are expected to exhibit significant turbulence, affecting species such as CH_4, CO, CO_2, N_2, and NH_3 (e.g., Noll et al., 1997; Saumon et al., 2000; Golimowski et al., 2004; Leggett et al., 2007; Visscher and Moses, 2011; Zahnle and Marley, 2014). Further Mukherjee et al. (2022) theoretically estimates (log) mixing values to be between 6-7 for brown dwarfs between 400-500 K. Consequently, we assume the considered abundances remain constant throughout the pressure column, with values (in $\log_{10}$ units) uniformly distributed between -10 and 0. We note that recent studies such as Rowland et al. (2023) have demonstrated that this approach may oversimplify atmospheric complexities, potentially leading to inaccurate retrievals of gravity, metallicity, and C/O ratios, irrespective of the parameterization of the *P-T* profile. Varying abundances along the atmosphere could be incorporated into future retrievals that utilize broad wavelength ranges enabled by combined approaches.

Table 5.1: Prior distribution for the 26 model parameters.

Parameter	Prior	Parameter	Prior
R_P	$\mathcal{U}[0.5, 3)$	H_2O	$\mathcal{U}[-10, 0)$
M_P	$\mathcal{U}[1, 50)$	CO_2	$\mathcal{U}[-10, 0)$
T_{bottom} [a]	$\mathcal{U}[100, 9000)$	CO	$\mathcal{U}[-10, 0)$
$T_{\text{nodes[i-ix]}}$ [b]	$\mathcal{U}[0.2, 1)$	CH_4	$\mathcal{U}[-10, 0)$
b_w [c]	$\mathcal{U}[-17, -11)$	NH_3	$\mathcal{U}[-10, 0)$
b_g [c]	$\mathcal{U}[-17, -11)$	PH_3	$\mathcal{U}[-10, 0)$
b_m [c]	$\mathcal{U}[-15, -7)$	H_2S	$\mathcal{U}[-10, 0)$
		$^{15}NH_3$	$\mathcal{U}[-10, 0)$
		HCN	$\mathcal{U}[-10, 0)$
		TiO	$\mathcal{U}[-10, 0)$
		VO	$\mathcal{U}[-10, 0)$

Notes. All the abundances are mass fractions in $\log_{10}$ units.
[a] T_{bottom} is the temperature at the bottom-most node in the pressure grid. [b] $T_{\text{nodes[i-ix]}}$ are the subsequent fractions of the previous node temperatures. [c] The b factor for the instruments are additive noise factors, in log value, by which the square of the measured error bars are exaggerated in each bin of the spectrum. This embodies the uncertainty in the estimated error of each instrument or model inaccuracies.

The spectra are calculated with the radiative transfer routines implemented in `petitRADTRANS`. The planet mass M_p and radius R_p are also considered to be free parameters and are used to calculate the emission flux. The noise model is identical to that defined in Section 4.4.2, except that the standard deviation σ_N across all wavelength bins throughout the combined spectral length of the WFC3+GNIRS+MIRI observations is not identical. They are uniquely obtained from the calibrated noise measurements of these respective instruments. To account for potential systematics not removed from the data, noise due to combining datasets from instruments each with different systematics, underestimating the measurement noise from the pipeline, making a wrong Gaussian assumption of the noise behavior, using an inadequate model, either due to 1D oversimplification or using a wrong model, and randomness in the source, etc., an additional scaling factor b is added to inflate the standard deviation on the noise measured for each instrument such that the total error s is given by $s^2 = \sigma_N^2 + 10^b$ (Line et al., 2015). The free parameters b_w, b_g and b_m pertaining to the WFC3, GNIRS and MIRI instruments respectively, are set to take values uniformly distributed in the range of [-17, -11), [-17, -11) and [-15, -7). These parameter ranges scale the maximum standard deviations of their respective instruments by factors of 1.35, 1.1 and 33 times respectively. A lower prior range chosen for the WFC3 and GNIRS than for the MIRI b factors is motivated empirically based on our previous retrievals on this source that consistently prefer an insignificant increase in the noise scaling over the near-infrared range, potentially due to relatively robust and well-understood (preliminary) error bars on these older instruments.

Table 5.2: Technical details of the posterior estimator and training used for all WISE 1738 retrievals.

Flow and embedding		Tuned	
Neural architecture	Details	Hyperparameter	Value
Normalizing flow	NAF	Optimizer	AdamW
Flow transforms	3	Initial learning rate	0.001
Transform	MLP	Scheduler	ReduceLROnPlateau
Signal	16	Patience	32
Hidden features (transform)	$5 \times [512]$	LR reduction factor	0.5
Activation (transform)	ELU	Minimum learning rate	1×10^{-8}
Embedding architecture MIR	Residual MLP	Weight decay	1×10^{-2}
Embedding depth MIR	$[512] \times 2 + [256] \times 3 + [128] \times 5$	Batch size	1024
Embedding output MIR	$1298 \to 64$	Loss	NPELoss
Embedding architecture NIR	Residual MLP	Epochs	100
Embedding depth NIR	$[512] \times 3 + [256] \times 5 + [128] \times 7$		
Embedding output NIR	$129 + 305 \to 8$		

5.3.2 Prior

The prior distribution in the case of a cloud-free model is a 26-dimensional multivariate uniform distribution $p(\theta)$ with physically motivated ranges for each parameter listed in Table 5.1.

5.3.3 Training set

The training set consists of approximately 3.7 million pairs of parameters and spectra $(\theta, f(\theta))$, providing around 1.8 value points per dimension if one were to use a regular 26-dimensional grid. The sample pairs in this training set are split into 90%, 9% and 1% for training, validation and testing respectively. The training set includes combined simulations of the near-infrared and the mid-infrared wavelengths.

For the training set, atmospheric models between the wavelength ranges of 0.98-2.2 μm in the near-infrared and 4.9-18 μm in the mid-infrared are simulated to match the default wavelength spacing of $\lambda/\Delta\lambda = 1000$ in `petitRADTRANS`. The WFC3 and GNIRS spectra have overlapping coverage in the near-infrared region, although they have different wavelength spacings. To match the WFC3 wavelengths, the spectra in the NIR range are convolved to the WFC3 spectral resolution and rebinned to a spacing of 130.

To generate the GNIRS component of the training set, the NIR spectra at the `petitRADTRANS` simulated default wavelength spacing of $\lambda/\Delta\lambda$ =1000 is retained, but masked to match the rebinned GNIRS observations. Similarly, the mid-infrared spectra are generated to match the MIRI observations. All three component spectra are combined to generate the training set. The simulations took around 3200 CPU hours.

5.3.4 Technical details on NPE

12 optimal neural auto-regressive flow architectures are found by conducting 128 parallel runs, each with a configuration randomly chosen from a uniform grid over all the architectural and training-based hyper-parameters, identical to Section 4.4.2, with an additional embedding layer which is used to represent the near-infrared part of the spectrum. The hyper-parameters are adjusted on the validation data. Of these 12, the run leading to lower validation loss and/or more stable training across all atmospheric models is preferred. The optimal hyper-parameters used for studies in this chapter are tabulated in Table 5.2. The process of tuning took 24 hours.

The flow is trained for a total of 100 epochs, during which 1700 and 170 random batches of 1024 and 256 pairs each of $(\theta, f(\theta))$ are used from the training and validation sets. The time taken to train the posterior estimator $p_\phi(\theta|x)$ is approximately 5.8 hours

5.4 Results

After training, the posterior estimator is conditioned on the spectrum of WISE 1738, which is then sampled 20,469 times from the joint posterior to generate the 1D and 2D marginal posteriors. This takes approximately 1 minute on average. Therefore, the total time required to perform a single retrieval, including the overhead cost of training dataset generation, hyper-parameter tuning, and training takes approximately 30 hours.

The results of the NPE retrievals on the combined WFC3+GNIRS+MIRI spectral observations x_{obs} of WISE 1738 are presented in the form of 1D and 2D marginal posterior distributions in Figure 5.2. These are approximated by extensively sampling from the estimated joint posterior distribution by means of forward passes through the trained normalizing flow. These samples are then used to construct the 68.3%, 95.5% and 99.7% credible regions of the *P-T* profile. Along with the constrained parameters, we also plot the derived (i.e, not retrieved) posterior distributions of the $^{14}NH_3/^{15}NH_3$ ratio and surface gravity. The figure also includes the posterior distribution of the *P-T* profile, shown in the inset. The emission contribution function is overlaid on top of the *P-T* profile in dashed lines, brightly highlighting the regions of the atmosphere that are probed by the observations. It also includes the equilibrium state water ice, ammonia, metal chloride and sulfide condensation curves (Lodders and Fegley, 2006) in blue, purple, red and green, respectively. From the plot, it can be seen that the profile is narrower in regions outside the contribution function and this is due to a more confident prior predictive distribution over those regions, due to the the prior's enforcement of smoother profiles across its space on each temperature node. Further, it can be seen that the water condensation curve intersects with the *P-T* profile above the upper limit of the probed photosphere, and the metal and chloride clouds intersect it within the near-infrared photosphere below.

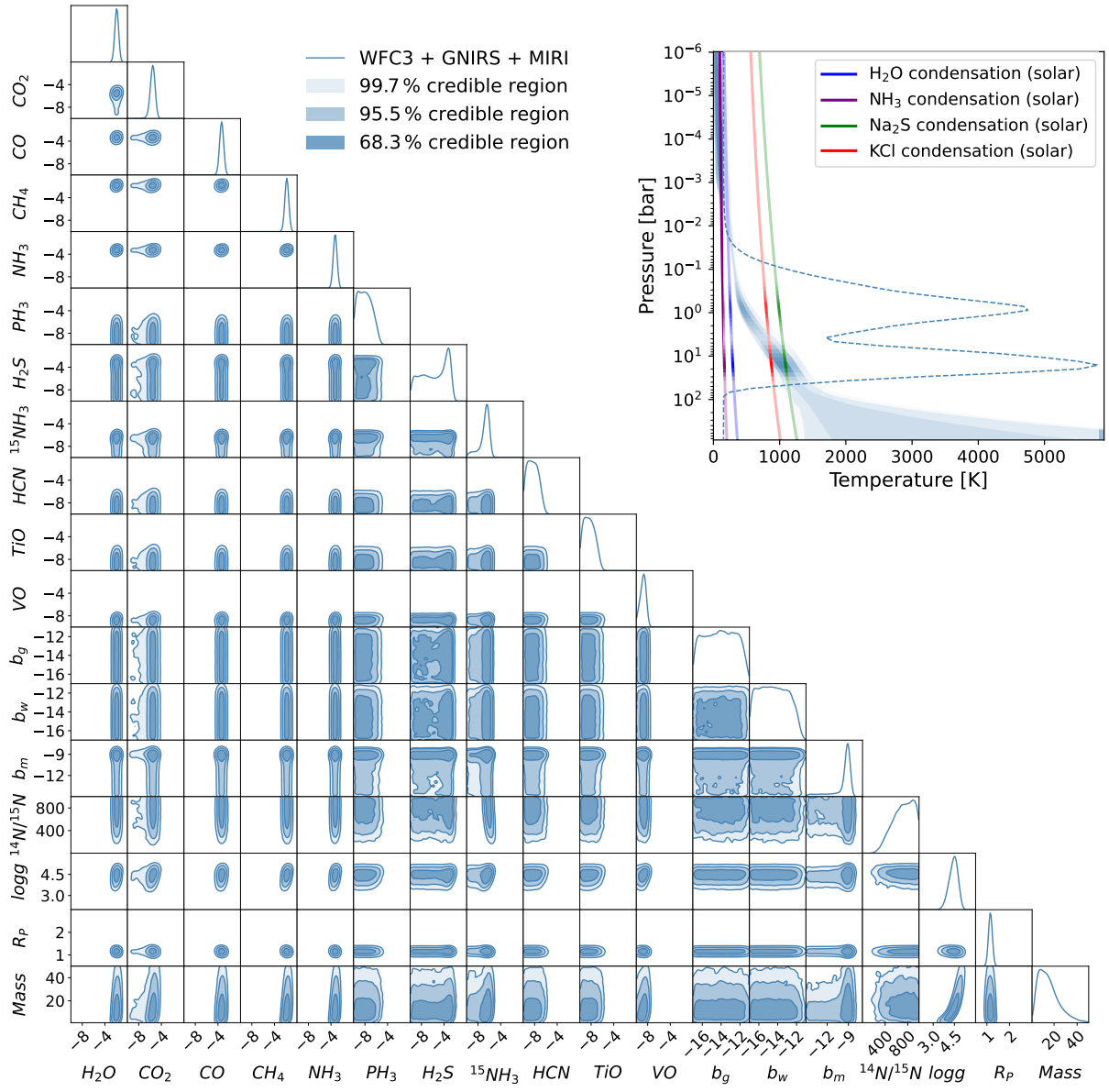


Figure 5.2: *Left*. Cloud-free retrieval using neural posterior estimation on the WISE 1738 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1738 observations x_{obs}, WFC3+GNIRS+MIRI spectra. *Right*. The top right figure illustrates the posterior distribution of the *P-T* profiles, that has the emission contribution function overlaid on top of it in shades of white highlighting visibility. The equilibrium state water ice, ammonia, chloride and sulfide condensation curves are plotted along the profile in blue, purple, red and green respectively for solar metallicity [M/H] and C/O ratio.

Table 5.3: Retrieved atmospheric (log) abundances as volume mixing ratios.

Retrieval	H_2O	CH_4	CO	CO_2	NH_3	$^{15}NH_3$
Z19-free[a]	$-2.87^{+0.08}_{-0.08}$	$-2.75^{+0.12}_{-0.10}$	-3.3	-4.1	$-4.21^{+0.10}_{-0.09}$	-
Z19-constr.[b]	$-2.97^{+0.09}_{-0.12}$	$-2.89^{+0.12}_{-0.13}$	-3.79	-3.83	$-4.34^{+0.12}_{-0.13}$	-
This work[c]	$-2.86^{+0.11}_{-0.11}$	$-2.72^{+0.14}_{-0.15}$	$-4.49^{+0.18}_{-0.18}$	$-6.87^{+0.25}_{-0.31}$	$-4.2^{+0.12}_{-0.12}$	$-7.53^{+0.43}_{-1.15}$

Notes. Log abundances are expressed units of volume mixing ratios.
[(a)] Called the "free" model with 31 parameters. This incorporates a 80 $M_{\rm Jup}$ mass prior upper limit. The 3σ upper limits are from Table 4 of Zalesky et al. (2019). [(b)] Called the "constrained" model, also with 31 parameters. Here the previously used upper mass limit is removed and restraints are applied on the priors of radius and $\log g$, as $0.7 < R/R_{\rm Jup} < 2.0$ and $3.5 < \log(g) < 5.5$ respectively. The 3σ upper limits are computed from the marginal posteriors obtained by Zalesky et al. (2019). [(c)] 3σ upper limits computed from the marginal posteriors obtained in our work. The volume mixing ratios in this work were calculated by converting the mass fractions from the cloud-free retrieval.

We obtain clear constraints on the abundances of H_2O, CO_2, CO, CH_4 and NH_3. All the constrained abundance values are documented in Table 5.3. We identify upper bounds on $^{15}NH_3$, PH_3, H_2S, HCN, TiO and VO with their 3σ limits -5.68, -5.31, -2.50, -6.15, -6.08 and -8.00 respectively, implying a non-robust detection (Line et al., 2015, 2017). The 3σ limits are computed as the 99.85% upper percentile corresponding to a symmetric 99.7% interval. We also observe that the $^{14}NH_3/^{15}NH_3$ ratio is not well constrained, and gives a 3σ lower bound of 275. All the retrieved and computed physical parameters are documented in Table 5.4. The most probable sample from the posterior, and its normalized residuals which is normalized to the inflated standard deviation are displayed in Fig. 5.1.

The noise scaling factors (or b factors) for WFC3 and GNIRS are not constrained. Instead they have an upper bound at -12. However, values of b factor < -12 donot have any significant impact on the error bars. This implies that the retrieval does not seek to enhance observational uncertainty in the near-infrared region to find good model fits. The retrieval prefers a highly confident observation. This means that any deviation of the model from the observation is purely due to the changes in the model parameters. There is no numbing of the impact of the changes in the model parameters with noise. Further, the drop in marginal values from -12 to -11 is a real effect that persists even when using a broader prior range. However, for MIRI, the error estimates are scaled with a b factor around -9. This scaling factor implies that the uncertainty is approximately 4 times higher in the largest error bar compared to the measured value. This means that the retrieval prefers a less confident observation to find good model fits in this wavelength range. These results align with prior expectations since the WFC3 and GNIRS have been iteratively configured to account and correct for redeemable sources of instrumental noise, while the configuration of MIRI is an ongoing process.

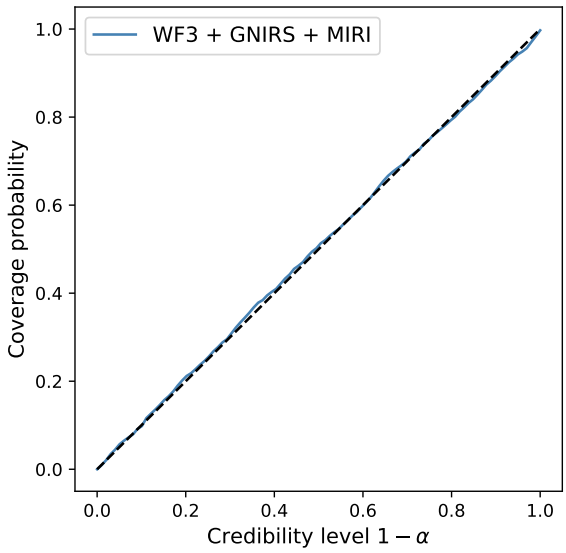


Figure 5.3: Coverage plot for the cloud-free posterior estimator.

5.5 Validation

We perform various validation tests to validate the trained NPE posterior estimator using coverage checks, consistency plots, and the L-C2ST evaluations, as shown below.

Coverage

The coverage plot for the cloud-free posterior estimator is plotted in Figure 5.3. It shows that the estimated posterior is appropriately dispersed.

Consistency plot

We also illustrate consistency plots plotted for the posteriors obtained from retrieving spectra spanning various wavelength ranges, from combined to separate near and mid-infrared wavelengths in Figures 5.4 and 5.5. We see that the posteriors conditioned on shorter wavelength ranges are consistent only in the regions they were trained on, while being significantly inconsistent in the wider regions of the spectrum. This implies a bias in retrievals for smaller wavelength regions.

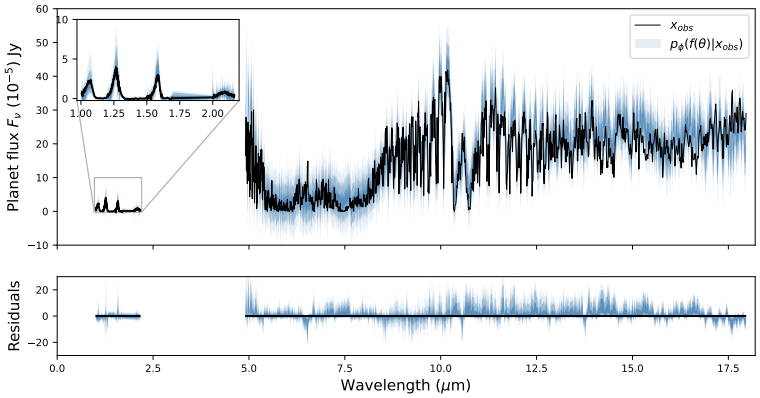


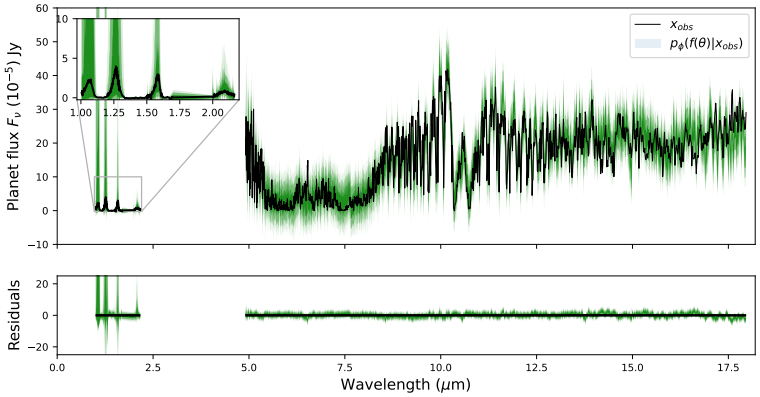
Figure 5.4: WFC3+GNIRS+MIRI consistency plot.

L-C2ST

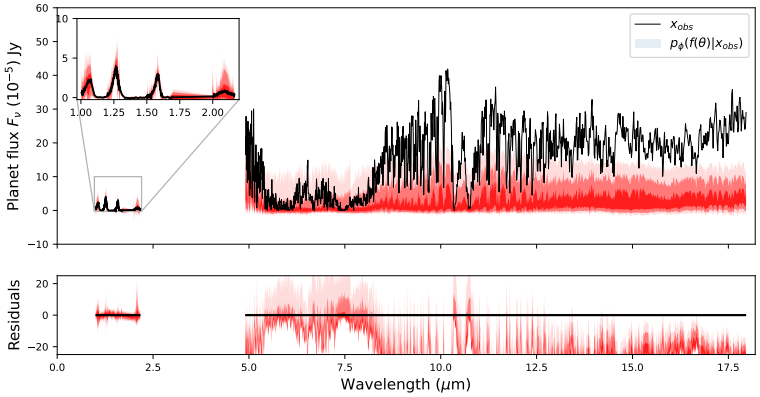
The binary classifier for the L-C2ST test (see section 4.4.4) is implemented as an ensemble of 15 MLP from `scikit-learn`. The classifier is initialized with two hidden layers, each having a number of neurons equal to 10 times the number of input features (ndim = 26). The ReLU activation function is used in the hidden layers, and the Adam optimizer is employed to adjust the model's weights. The training process runs for a maximum of 100 iterations. The classifier is trained on 50k samples from the training set. When trained, the classifier uses back-propagation to minimize the binary cross-entropy loss function, adapting its weights to improve predictions of one of two possible outcomes (e.g., 0 or 1) based on the input data.

Using the trained classifier, we calculate the p-values for the estimated posteriors retrieved on the WISE 1738's spectrum, the most probable simulated observation, and a random sample from the prior. The three spectra are illustrated in Figure 5.6. Each of these posteriors is tested to see if they follow the null hypothesis, which proposes that the estimated posteriors are indistinguishable from the true posterior. In each case, the T-distribution shown in Figure 5.7 indicates a p-value well within the rejection threshold of 0.05, suggesting that the null hypothesis cannot be rejected. This implies that the estimated posterior for these observations is close to the true posterior.

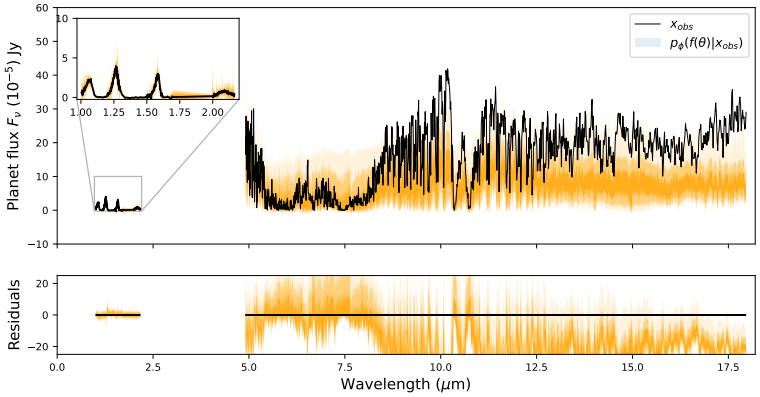
Next, we present the pp-plots for these observations (see Figure 5.8). The pp-plot, a variation of the coverage plot, helps assess the overall trend of bias or the potential over- or under-confidence of the estimated posterior. For all three estimated posteriors, the red curve is entirely within the gray confidence region, suggesting that it is neither significantly over-dispersed nor under-dispersed. However, there is a slight rightward bias (a small deviation from the dashed vertical line) in the latter two cases, indicating that some estimated parameters may be slightly higher than the nominal value. In contrast, the posterior pertaining to the most probable sample shows less bias.



(a) MIRI-only consistency plot.



(b) WFC3-only consistency plot.



(c) GNIRS-only consistency plot.

Figure 5.5: Comparison of consistency plots from different retrievals on the WISE 1738 spectrum. In descending order, we compare retrievals using the full WFC3+GNIRS+MIRI data (in steelblue), using only MIRI data (in green), using only WFC3 data (in red) and only GNIRS data (in orange). Each plot shows the posterior predictive distribution $p(f(\theta) + \epsilon|x_{\text{obs}})$ of noisy simulated spectra for different confidence levels, overlaid on the observed spectra.

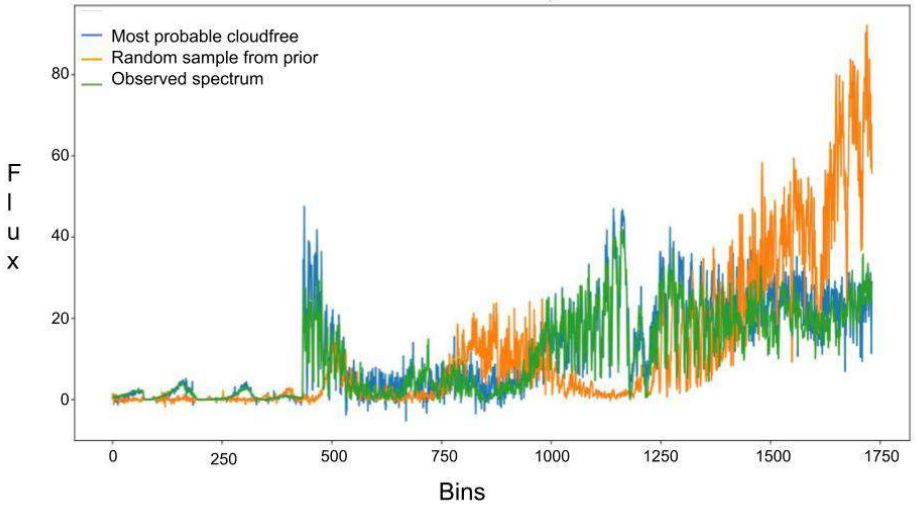


Figure 5.6: The three observations for which we evaluate the estimated posterior.

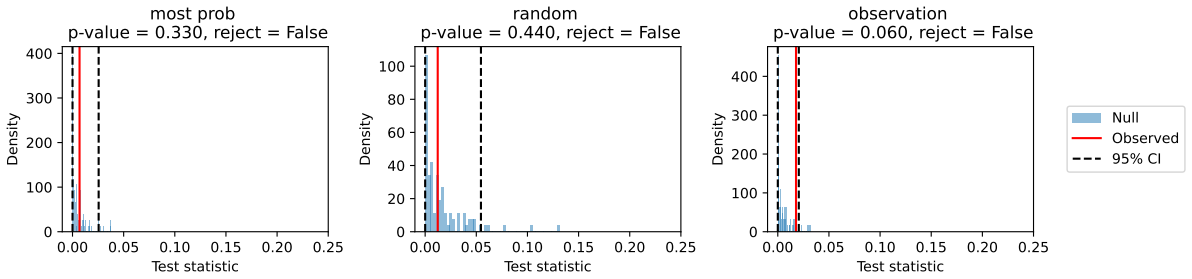


Figure 5.7: T distribution plot. The p-values are computed as the proportion of times the L-C2ST statistic under the null hypothesis is greater than the L-C2ST statistic at the observation x_{obs}.

5.6 Discussion

5.6.1 Combined retrieval vs near-infrared retrievals

Previous studies, such as those by Cushing et al. (2011); Schneider et al. (2015); Zalesky et al. (2019), used WFC3 data to investigate the atmosphere of WISE 1738 while Leggett et al. (2016a, 2017) used the GNIRS data for their analysis. Here, we combine these near-infrared datasets with mid-infrared observations for the first time. The combined retrieval of the WFC3, GNIRS, and MIRI spectra provides a more comprehensive view of the atmosphere by effectively expanding the range of probed pressure levels. This fills gaps even within the near infrared wavelength range by considering data at different resolutions/observational conditions. The

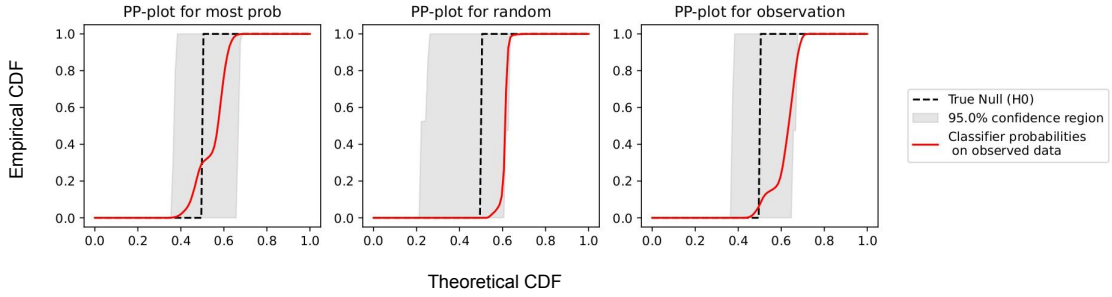


Figure 5.8: PP-plot. Cumulative distribution function (CDF) for the three posterior estimates.

Table 5.4: Summary of previous model fits/retrievals attempting to characterize WISE 1738.

Study	$T_{\rm eff}$ (K)	$\log g$ (cm/s^2)	Mass ($M_{\rm Jup}$)	Radius ($R_{\rm Jup}$)	Age (Gyr)	$\log K_{zz}$	Distance (pc)	C/O	[M/H]
C11	350–400	4.75–5.0	20	0.86–0.94	–	4	3.4–7.3[d]	–	solar
S15	400	4.0–4.5	5–14[a]	0.47[b]	0.6–3	–	–	–	solar
L16b	425 ± 25	4.0 ± 0.25	3–9[a]	–	0.15–1	6	7.8 ± 0.6[e]	–	0, +0.2
L17	410–440	4.0–4.5	5–14[a]	1.0–1.2[a]	0.3–3	6	10.5[f]	–	-0.05 ± 0.25
Z19-free	371^{+27}_{-29}	$5.43^{+0.13}_{-0.17}$	59^{+15}_{-22}	$0.71^{+0.05}_{-0.05}$	> 10	–	7.34 ± 0.22[g]	$1.32^{+0.1h}_{-0.09}$	$0.35^{+0.10h}_{-0.09}$
Z19-constr	371^{+33}_{-30}	$5.20^{+0.20}_{-0.29}$	34^{+20}_{-17}	$0.73^{+0.04}_{-0.03}$	> 10	–	7.34 ± 0.22[g]	$1.2^{+0.09h}_{-0.03}$	$0.23^{+0.11h}_{-0.13}$
This work	402^{+12}_{-9}	$4.43^{+0.26}_{-0.34}$	13^{+11}_{-7}	$1.14^{+0.03}_{-0.03}$	1–4[c]	–	7.34 ± 0.22[g]	$1.35^{+0.39}_{-0.31}$	$0.34^{+0.12}_{-0.11}$

Notes. The studies by Cushing et al. (2011) (C11), Schneider et al. (2015) (S15), and Zalesky et al. (2019) (Z19) utilize data from WFC3 (Cushing et al., 2011) to perform their analysis, whereas Leggett et al. (2016b) (L16b) and Leggett et al. (2017) (L17) rely on data from GNIRS (Leggett et al., 2016a).
[(a)] Estimated from the evolutionary model of Saumon and Marley (2008). [(b)] Computed by multiplying the retrieved ratio 6.445×10^{-2} $R_{\rm Jup}$/pc by the distance measure from Martin et al. (2018). [(c)] The age is estimated using the evolutionary model from Marley et al. (2021). [(d)] Distance measure from Cushing et al. (2011). [(e)] Distance measure from Kirkpatrick et al. (2011). [(f)] Distance measure from Beichman et al. (2014). [(g)] Distance measure from Martin et al. (2018). [(h)] Computed from the posteriors of (Zalesky et al., 2019) without sequestration.

advantage of combining datasets is demonstrated through a comparative study, where we perform single retrievals on WFC3, GNIRS and MIRI data separately, as well as a combined retrieval on WFC3, GNIRS, and MIRI data. The results of this comparison are shown in Figure 5.9.

The comparison reveals that the abundances of CO, CO_2 and $^{15}NH_3$ are more tightly constrained in the combined retrievals than in those based solely on the near-infrared range. This is discussed in more detail in the following subsections. Additionally, although not significant, we see a similarly slight improvement in the confidence of constraints on the remaining abundances, suggesting that these features are strong in all datasets and hence easier to constrain. Similarly, the constraints on the *P-T* profile also improve, with the near-infrared region

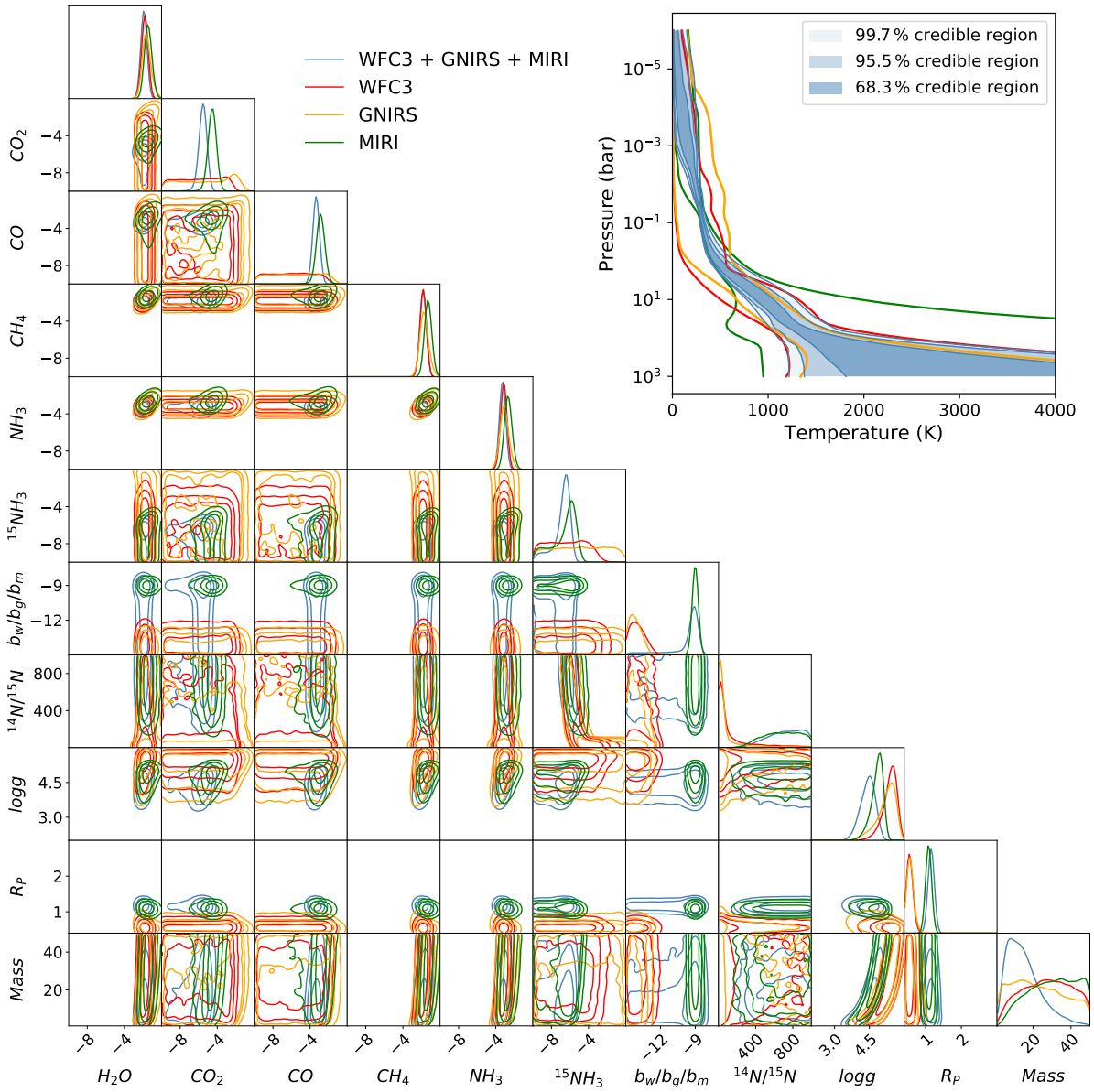


Figure 5.9: Comparing individual spectral retrievals of WISE 1738 across different wavelength regions with the combined retrievals. The corner plot shows 1D and 2D marginal posterior distributions obtained for the WFC3, GNIRS and MIRI spectrum along with the combined WFC3+GNIRS+MIRI spectra. The top right figure illustrates the posterior distribution of the $P-T$ profile of the combined retrieval, while also highlighting the 99.7% credible intervals of the three independent retrievals.

providing tighter constraints on the deeper atmospheric layers, while the mid-infrared region contributes to tighter constraints in the upper layers. Although the values of the major opacity species such as H_2O, CH_4 and NH_3 agree with the constraints from (Zalesky et al., 2019), their corresponding uncertainties are difficult to compare due to differences in their setups. However, the retrieved posterior validation tests can be found in Section 5.5. Furthermore, we

do not include alkalis in our final combined retrievals since we find that the MIRI only retrievals do not offer a bound on them, hence offering no added information about their contents from Zalesky et al. (2019) who obtain a 3σ upper bound at -5.2. The most significant improvement, however, is in the constraints on mass, radius and surface gravity.

The free retrievals conducted exclusively on the near-infrared spectrum, both in our study and in the work by Zalesky et al. (2019), result in higher estimates for mass (40 M_J) and surface gravity (5.5 cms^{-2}), along with a lower estimate for radius (0.7 R_J). In contrast, the combined retrievals result in lower masses and gravities, and larger radii that align better with the predictions from the evolutionary models (discussed more in Sect. 5.6.3). To explain this difference, we produce consistency plots across the entire wavelength range (see Figure 5.5), which reveal that in each case, while retrievals remain consistent within the originally retrieved spectral region, they fail to predict consistent spectra in the extended (not retrieved) regions. This is in contrast with the consistency plot obtained for the combined retrieval in Figure 5.4, which exhibits consistency across the extended wavelength range. The enhanced consistency obtained in the latter case demonstrates how biased our characterizations can become when data coverage is limited. It also suggests that when focusing on narrow wavelength regions, multiple competing hypotheses can fit the spectral shape. Therefore, the MIRI data provide critical additional information, enabling improved constraints on these physical parameters. These findings highlight the necessity of incorporating broader datasets to achieve accurate characterizations of brown dwarfs.

Interestingly, in the combined retrievals, while the normalized residuals of the consistency plot are centered around the horizontal black line at zero within the 1-10 μm wavelength range, we observe a slight offset above the zero line in the 10-16 μm range. This offset is not observed in retrievals performed exclusively on each individual dataset. Although not significant, this suggests a challenge in reconciling the near-infrared and MIRI spectra, which may stem from either systematic noise effects that are not accounted for in the data calibration, or the absence of a more comprehensive forward model that accounts not only for bulk chemical processes but also for localized chemistry. One such instance is the deep atmospheric dynamics driven by fingering convection under dis-equilibrium chemistry, which can cause compositional gradients to impact the different regions of the spectrum differently, as discussed by Tremblin et al. (2015). Local effects become increasingly important while analyzing such long spectral wavelength ranges that provide deeper insights into larger parts of the atmosphere. Alternatively, the slight discrepancy could arise from an unknown process not accounted for in the forward models such as a missing opacity source deeper in the atmosphere (Morley et al., 2018; Beiler et al., 2024b). Additionally, it could also be the models attempting to adjust for variations in the temperature continuum slopes, or from suspected near-infrared variability, estimated to be between 5% and 30% in the near-infrared Leggett et al. (2016b).

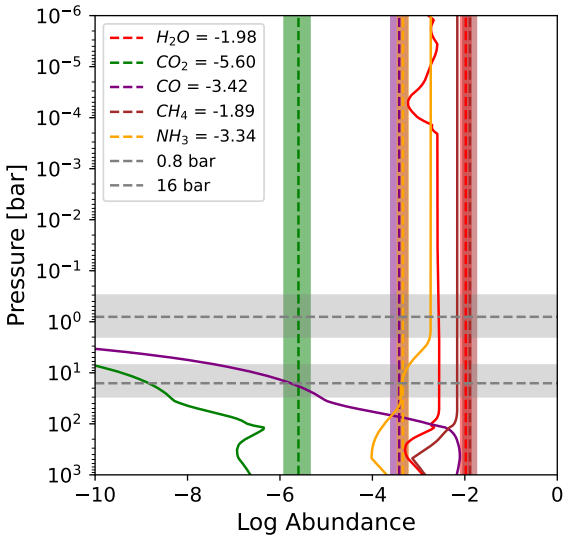


Figure 5.10: Chemical equilibrium abundances for an atmosphere with identical composition as the retrieval suggests for WISE 1738, having a C/O = 1.35 and [M/H] = 0.34 (in solid lines), calculated using the retrieved most probable *P-T* profile. These are compared with the retrieved molecular abundances in mass fractions (in dashed lines, including 1σ uncertainties as colored bars) for key opacity-contributing species: H_2O (red), CO_2 (green), CO (purple), CH_4 (brown), and NH_3 (orange). The grey regions are the estimated 1σ of the emission contribution functions for the MIRI probed photosphere (at the top) and the near-infrared probed photosphere (at the bottom).

5.6.2 Disequilibrium chemistry

Elemental abundances in substellar objects are key to understanding their evolutionary history, as they govern atmospheric opacity and cooling rates (Burrows et al., 2001). These abundance patterns suggest either star-like formation by gravitational collapse or planet-like formation through gravitational instability, situating the sub-stellar brown dwarfs in the gap between higher-mass stars and lower-mass planets. Unlike stars, which directly reveal atomic abundances, the cooler atmospheres of brown dwarfs exhibit molecular abundances that can be used to determine elemental compositions. Additionally, molecular abundances not only reflect the atmospheric chemistry but also the dynamics at play within these atmospheres. In some instances, these molecular abundances are sensitive to equilibrium condensate rainout and vertical disequilibrium mixing (Burrows et al., 2001; Sharp and Burrows, 2007). In Figure 5.10, we compare the retrieved molecular abundances, expressed as mass mixing ratios, for the key opacity-contributing species H_2O, CO_2, CO, CH_4, and NH_3 with the easyCHEM chemical equilibrium calculations for an atmosphere with a composition identical to the retrieval of WISE 1738, having a C/O = 1.35 and [M/H] = 0.34, tabulated in `petitRADTRANS` (see, Mollière et al., 2017; Lei and Mollière, 2024), in order to gain insights into the atmospheric dynamics of WISE 1738.

The retrievals constrain the abundances for CO and CO_2, even though their chemical equilibrium abundances are expected to be depleted at pressures lower than 1 bar. The fact that these species can be constrained to values higher than expected for chemical equilibrium using data from a higher photospheric region (illuminated by the MIRI spectrum), while remaining unconstrained in the near-infrared retrievals (unexpected under chemical equilibrium), provides clear evidence of disequilibrium mixing. This is explored in detail by Beiler et al. (2024b), who investigate the over-abundance of CO_2 in brown dwarf atmospheres and suggest vertical mixing as a reason for this enhancement, despite an expectation to be quenched. These significant deviations suggest that additional processes, such as vertical mixing or chemical kinetics, need to be incorporated to fully explain the observed abundances with self-consistent models.

In contrast, while the retrieved opacity of CH_4 aligns well with equilibrium values, consistent with previous studies (Burrows and Sharp, 1999; Sharp and Burrows, 2007), H_2O is retrieved to have a higher value. However, the retrieved abundance of NH_3 is lower than its equilibrium value in the MIRI probed photosphere, which is likely depleted due to quenching (Zahnle and Marley, 2014), but is consistent with the near-infrared probed photosphere lower in the atmosphere. Consistency with equilibrium abundance seen in CH_4 confirms a constant abundance profile for the molecule within the dis-equilibrium state of the atmosphere.

Additionally, we calculate a ratio of $^{14}NH_3/^{15}NH_3$ which has recently been proposed as a new tracer for formation history (Barrado et al., 2023). Our measurement, which is broadly constrained with a lower bound value of 200, largely overlaps with the range found for WISE 1828 (670^{+390}_{-211}) and the solar system. However, it also includes much lower estimates, aligning more closely with the estimate of 332^{+39}_{-43} for WISE 0855 reported by Kühnle et al. (2024) and the ISM. Such broad values provide an ambiguous understanding of its formation history. Additional measurements of isotopologues in the atmosphere of WISE 1738 could help us understand the connections to formation scenarios further.

5.6.3 Evolution of WISE 1738

We estimate the effective temperature of WISE 1738. This calculation relies on the bolometric fluxes over a broad wavelength range of 0.8 to 50 μm, derived from the retrieved posterior. Assuming a distance of 7.34 pc (Martin et al., 2018) and the retrieved radius posterior, the effective temperature is determined to be 402^{+12}_{-9} K, which is in agreement with previous estimates in Table 5.4. Additionally, the $\log g$ is calculated using the retrieved values for mass and radius, yielding a result of $4.43^{+0.26}_{-0.34}$.

The evolutionary models presented by Saumon and Marley (2008) and Marley et al. (2021) suggest that a brown dwarf with an effective temperature of approximately 400 K and a $\log g$ between 4.1 and 4.7 cm s^{-2} would have a mass in the range of 6–20 M_J, a radius between 0.97 and 1.1 R_J, with a bolometric luminosity $\log L/L_\odot$ between -6 and -7. The retrieved

mass of 13^{+11}_{-7} M_J, a radius of 1.14 R_J, and bolometric luminosity of $-6.52^{+0.05}_{-0.04}$ align reasonably well with these theoretical predictions, supporting consistency between observed and modeled parameters. Further, we estimate an age spanning 1 to 4 Gyrs using the evolutionary model from Marley et al. (2021), which is consistent with the rotation period of 6 hours measured for WISE 1738 (Leggett et al., 2016a), as brown dwarfs are expected to spin faster as they age (Bouvier et al., 2014).

5.6.4 C/O and metallicty

We determine the carbon-to-oxygen (C/O) ratio to be $1.35^{+0.39}_{-0.31}$, and metallicity [M/H] to be $0.34^{+0.12}_{-0.11}$, under the considered atmospheric model, assuming no oxygen sequestration in the atmosphere. However, the derived atmospheric abundances may not reflect the true chemical composition of the atmosphere because some chemical elements can be used up in cloud particles and/or are removed from the atmosphere due to rainout. Atmospheric oxygen is typically depleted by 20–30% relative to intrinsic values due to sequestration in condensates, with the extent of depletion depending on the intrinsic metallicity and C/O ratio (Line et al., 2015). This process increases the atmospheric C/O ratio while reducing the overall atmospheric metallicity, as oxygen is a dominant metal. Specifically, we account for enstatite and forsterite condensation (Fegley and Lodders, 1994) by adjusting the abundances of oxygen-bearing molecules by a factor of 1.3, following the approach in Zalesky et al. (2019). This adjustment is equivalent to the removal of 3.28 oxygen atoms for every silicon atom (Burrows and Sharp, 1999). Incorporating this sequestration, we recalculate the C/O ratio and metallicity [M/H] as $1.04^{+0.30}_{-0.24}$ and $0.40^{+0.12}_{-0.10}$, respectively. These values lie within 2σ and 4σ of solar values, respectively. Such super-solar C/O ratios and metallicity have been observed in late T-dwarfs, as reported by Line et al. (2017); Zalesky et al. (2019). This ratio also aligns with the upper range of the local FGK population, which extends to a C/O ratio and metallicity upto values 1.4 and 0.6 respectively (Zalesky et al., 2019; Hinkel et al., 2014).

In case the brown dwarf is formed in isolation, it is thought to form via gravitational collapse that should result in near-solar C/O ratios and metallicity. However, if it is formed around a star and ejected, this could change. It is interesting to note that Pascucci et al. (2013) and more recently Tabone et al. (2023); Arabhavi et al. (2024), find the inner protoplanetary disk to be carbon-rich with molecules such as C_2H_2, HCN, C_6H_6, CO_2, HC_3N, C_2H_6, C_3H_4, C_4H_2, and CH_4 dominating the disk with little traces of H_2O (Arabhavi et al. 2025, submitted), suggesting that such high C/O values could be an artifact of formation. However, further understanding of oxygen sequestration processes and formation scenarios is necessary to explain such a high C/O ratio and metallicity.

5.6.5 Comparison with grid models

We compare the retrieved P–T profile posterior of WISE 1738 to the closest 1D self-consistent atmospheric models in terms of bulk properties such as effective temperature, surface gravity, and metallicity. For this comparison, we use models from grids by Lacy and Burrows (2023) and the Sonora Elf Owl (Mukherjee et al., 2024a). These two sets of grids are modeled with radiative-convective equilibrium, and equilibrium chemistry or vertical mixing-induced disequilibrium chemistry in Y dwarf atmospheres used here. (Lacy and Burrows, 2023) uses coolTLUSTY (Burrows et al., 2008; Hubeny and Lanz, 1995; Sudarsky et al., 2005) to generate models spanning various ranges of effective temperatures, metallicities and surface gravities (see https://doi.org/10.5281/zenodo.7779180). Additionally, Sonora Elf Owl spans sub-solar to super-solar atmospheric Carbon-to-Oxygen ratios and vertical eddy diffusion coefficients and uses PICASO to generate the models (see https://doi.org/10.5281/zenodo.10381249). Both of these models include (water ice) cloudy and clear cases.

6 models from (Lacy and Burrows, 2023) are chosen for comparison. These include both clear and cloudy atmospheres with equilibrium and dis-equilibrium chemistry. In each case, the models have a log g of 4.5, T_{eff} of 400 K and a metallicity of 0.316 Additionally, the cloudy models used include two types of tapering of cloud opacity along the height of the atmosphere (see equation 2 of (Lacy and Burrows, 2023)). This is denoted by the naming convention AEE and E which represent weak and strong tapering factors with values 2 and 6 respectively. In each case, the model cloud particle size used is 10 μm. These models are labeled as AEE10 and E10 respectively. For models with vertical mixing, a mixing coefficient, log kzz, of value 6 is used. Similarly, 5 Sonora Elf models are chosen for comparison. These include log g of 4.5, T_{eff} of 400 K and a metallicity of 0.5, along with various values of the mixing coefficient, log kzz, between 2 and 9.

These grid comparisons are plotted in Figure 5.11. Firstly, we see that there is more variation between grids than that within each grid, which speaks about the impact of the differing treatment of the physical processes solved in each case. Further, we see that Sonora Elf Owl models are compatible with the posterior distribution of WISE 1738, while the Lacy models are not for all variations of configurations chosen.

5.7 Conclusion

While previous retrievals have primarily focused on near-infrared wavelengths, the inclusion of mid-infrared data from JWST/MIRI, allows us to probe higher in these atmospheres offering improved constraints on physical parameters like surface gravity, radius, mass, luminosity and unveiling dis-equilibrium chemistry by more accurately constraining chemical abundances. In

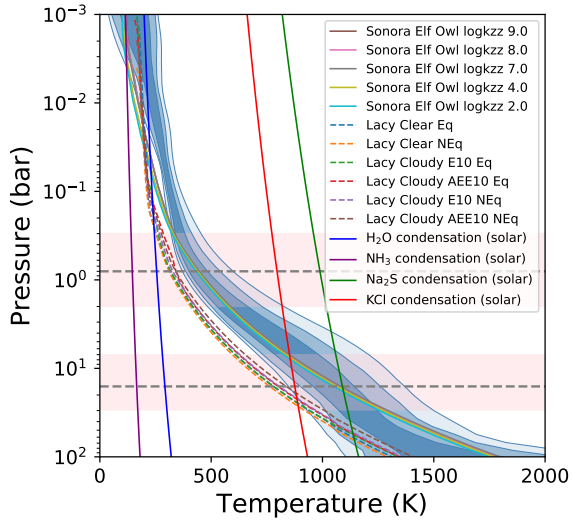


Figure 5.11: Comparison of the *P-T* posterior of WISE 1738 with the nearest possible grid model (in terms of T_{eff}, log g, composition, etc) from (Lacy and Burrows, 2023) and Sonora Elf Owl (Mukherjee et al., 2024a), for several clear and cloudy, and equilibrium and dis-equilibrium conditions. The pink regions are the estimated 1σ of the emission contribution functions for the MIRI probed photosphere (at the top) and the near-infrared probed photosphere (at the bottom). The equilibrium state water ice, ammonia, metal chloride and sulfide condensation curves are plotted along the profile in blue, purple, red and green respectively for solar metallicity and C/O ratio.

this study, we build a comprehensive understanding of the Y dwarf WISE 1738 by retrieving its *P-T* profile, chemical abundances (including an isotolpologue), comparing these findings to evolutionary models, and setting the stage for future investigations into complex atmospheric processes and cloud formation.

We perform atmospheric retrievals using NPE on the combined spectral data from WFC3, GNIRS, and MIRI observations. To estimate their posterior distribution, we train a normalizing flow based on amortized variational inference algorithm, using the atmospheric emission models generated with the simulator `petitRADTRANS`. By combining retrievals over an extended wavelength range, we demonstrate a reduction in the uncertainty in the retrieved *P-T* profile and in some molecular abundances, compared to individual retrievals from each observation. We also caution against potential biases in spectral characterization that may arise from using narrower wavelength ranges, highlighting improved consistency between the estimated posterior and the observations when combined spectral data is used instead of individual spectral retrievals. This confirms trends that we have seen in short wavelength retrievals.

Additionally, we constrain the bulk physical properties of WISE 1738 including mass, radius, surface gravity, and luminosity, with higher accuracy, in agreement with theoretical predictions from evolutionary models. We also estimate the object's age to be between 1 and 4 Gyr which is consistent with the fast rotation period of 6 hrs. However, although not a significant effect, reconciling the near-infrared and mid-infrared regions proves challenging, suggesting uncalibrated systematic noise or the presence of an unknown process such as local chemistry, a missing opacity source, or intrinsic variability, that is not yet accounted for.

In addition to the major opacity species such as CH_4 and H_2O, we also estimate the abundances of CO_2, CO, and NH_3, for the first time on this object. These results provide evidence of disequilibrium chemistry in WISE 1738's atmosphere due to vertical mixing, as they could not be constrained by the near-infrared data alone and equilibrium chemistry predicts their depletion below the near-infrared photosphere. This result adds to the evidence of vertical mixing in brown dwarf atmospheres.

Further, we perform grid comparisons with the NPE retrieval. We compare the *P-T* profile NPE posterior of WISE 1738 with several clear and cloudy, equilibrium and dis-equilibrium grid models. We find that the choice of grid greatly affects the resulting *P-T* profiles. Further, we see that the Sonora Elf Owl grid is consistent with the retrieval, whereas the Lacy grid is not, irrespective of the type of atmosphere chosen.

Although the cloud-free model seems mostly consistent with the observations of WISE 1738, in order to investigate the theoretical predictions of the presence of water clouds in its atmosphere, we perform cloudy retrievals on its spectrum in the next chapter.

Chapter 6

Cloudy retrieval of WISE 1738

Following the theoretical predictions for the presence of water ice clouds in the atmosphere of WISE 1738, we perform cloudy retrievals of its spectrum in full and in patchy form. We do not find evidence for either in the atmosphere of WISE 1738. Using the cloudy posterior obtained, we perform several experiments to identify the origin of their uncertainties in order to reduce them. Furthermore, we implement a Bayes factor test to pick the most consistent of these models to the WISE 1738 observation. Finally, we attempt importance sampling (IS) to correct any biases in our NPE posteriors.

6.1 Studying clouds in the atmosphere of WISE 1738

Although the cloud-free analysis performed in the previous chapter provides a good fit to the data, theoretical studies predict that Y dwarfs cooler than late T dwarfs, particularly around 400 K, may develop optically thin, low-lying chloride and sulfide clouds that influence the Y and J bands of the near-infrared spectrum (Figure 7 of Morley et al., 2012), and high-altitude fully opaque or patchy water clouds with small particle sizes (1-5 μm, Morley et al. 2014) that influence the mid-infrared region of the spectrum (Mang et al., 2022) resulting in opacity variations (Leggett et al., 2016b, 2017). Such an impact would become further apparent when considering a broader wavelength coverage such as we do here. While chloride and sulfide clouds have been studied before and have not resulted in consistent fits (Zalesky et al., 2019; Leggett et al., 2016b), water clouds prone to affect the mid-infrared region have not been studied.

6.2 Setup

6.2.1 Radiative transfer simulator

The *Cloudy* atmospheric forward model $f_{\rm cl}$ is implemented in `petitRADTRANS`. The *P-T* profile and molecular species are modeled in the same way as in the cloud-free case, with the addition of parameters describing clouds. The icy water clouds are parameterized using 5 free parameters based on the Ackerman and Marley (2001) model. This includes a cloud log mass fraction, X_{H_2Oc}, of water ice which is included as an added absorber species along with the other molecular species in the atmosphere. The mass fraction is defined at the cloud base located at a base pressure given by $\log P_{\rm base}$. The settling parameter, $f_{\rm sed}$, is a factor by which the cloud mass fraction decreases with increasing altitude as $X(P) = X_{H_2Oc}\left(\frac{P}{P_{\rm base}}\right)^{f_{\rm sed}}$. A vertical eddy diffusion coefficient, $\log k_{zz}$, determines the particle size for a given $f_{\rm sed}$ value. Additionally, the dispersion of particle sizes around the central value is described by the width of the log-normal size distribution, $\sigma_{\rm lnorm}$ (labeled simply as σ in the corner plots). The model is parameterized with a total of 31 parameters.

The *Patchy* atmospheric forward model $f_{\rm pat}$ is also implemented in `petitRADTRANS` similar to the cloudy model. We model patchy water clouds, or water clouds with holes in them (Marley et al., 2010; Morley et al., 2014), by combining two sub-columns of cloudy $F_{\rm cloud}(z)$ and cloud-free atmospheres $F_{\rm clear}(z)$, with identical P-T structures, weighted by a fraction h such that the total flux escaping from the atmosphere is given as,

$$F_{\rm tot}(z) = hF_{\rm clear}(z) + (1-h)F_{\rm cloud}(z). \tag{6.1}$$

Here the parameter h can be thought of as a factor by which the dimming effect of the modeled clouds and the cloud imparted features are linearly scaled. A value of $h = 0$ implies that the clouds dim the spectrum as expected, or 'full cloud cover', whereas a value of $h = 1$, implies a hole in the clouds or 'no cloud cover'. Everything in between implies a linear scaling of the dimming effect, $1-h$ times, or a cloud cover fraction of $1-h$. This is inspired by 3D climate models in terrestrial applications that allow some flux to escape from the bottom layers of the atmosphere even with the presence of optically thick water clouds. The model is parameterized by a total of 32 parameters, which include the 31 parameters identical to the fully cloudy model, along with the additional patchiness parameter h.

In either case, the noise model is parameterized as a Gaussian distribution, where the standard deviation of the measurement noise is scaled in the same manner as in the cloud-free approach using b factors.

6.2.2 Prior

The prior distributions for the cloudy and patchy models are 31 and 32 dimensional multivariate uniform distribution $p(\theta)$, with physically motivated limits for each parameter. Both these models include the 26 parameter priors from the cloud-free setup, with additional prior for the clouds. Parameters X_{H_2Oc}, $\log P_{\text{base}}$, $\log k_{\text{zz}}$, f_{sed} and σ_{lnorm} are set to take values uniformly distributed in the range of $[-10, 0)$, $[-6, 3)$, $[5, 13)$, $[0,10)$ and $[1.05, 3)$ respectively, and the h factor for the patchy model takes values uniformly between [0,1).

6.2.3 Training set

The cloudy and patchy model training set consists of approximately 3.7 million pairs of parameters and spectra $(\theta, f(\theta))$. The training set for the cloudy model share the same θ as for the cloud-free model, with the addition of clouds, constructed using randomly sampled cloud realizations, to compute the corresponding cloudy spectra. The patchy training set is then constructed by combining these cloudy spectra with corresponding cloud-free models based on randomly sampled patchiness parameter. All the post simulation processing, such as spectral binning of the WFC3 spectra, the masking of the GNIRS spectra and combining it with the MIRI spectra is carried out identically as in the cloud-free case.

6.2.4 Technical details on NPE

The same architecture as for the cloud-free model is used in both retrievals. See Table 5.2 for details.

6.3 Results

The results of cloudy and patchy retrievals are illustrated as corner plots in Figures 6.1 and 6.2 respectively. The corner plots include the derived marginal posterior distributions of the ratio and surface gravity and mass, along with the retrieved parameters. The cloudy posterior is an average across 6 posterior estimates on the same neural flow network.

The retrieval results for the cloudy model in Figure 6.1 suggest that similar abundances are constrained (or unconstrained) as in the cloud-free case, along with similar metallicity, C/O ratio and age estimates, with slight variations in $^{15}NH_3$ leading to a lower $^{14}N/^{15}N$ ratio, peaking at 400. Water ice clouds exhibit a mass fraction abundance of $X_{H_2Oc} = -2.46^{+1.28}_{-1.01}$, forming at a base pressure of $\log P_{\text{base}} = -0.94^{+1.48}_{-1.12}$. However, the vertical distribution remains largely unconstrained, as indicated by the f_{sed} parameter. The mixing coefficient suggests high mixing, with $\log k_{\text{zz}} = 7.93^{+1.13}_{-1.33}$, which, on the lower end, overlaps the values reported in Figure 19 of Mukherjee et al. (2022) for brown dwarfs at 400 K, which are typically expected to range

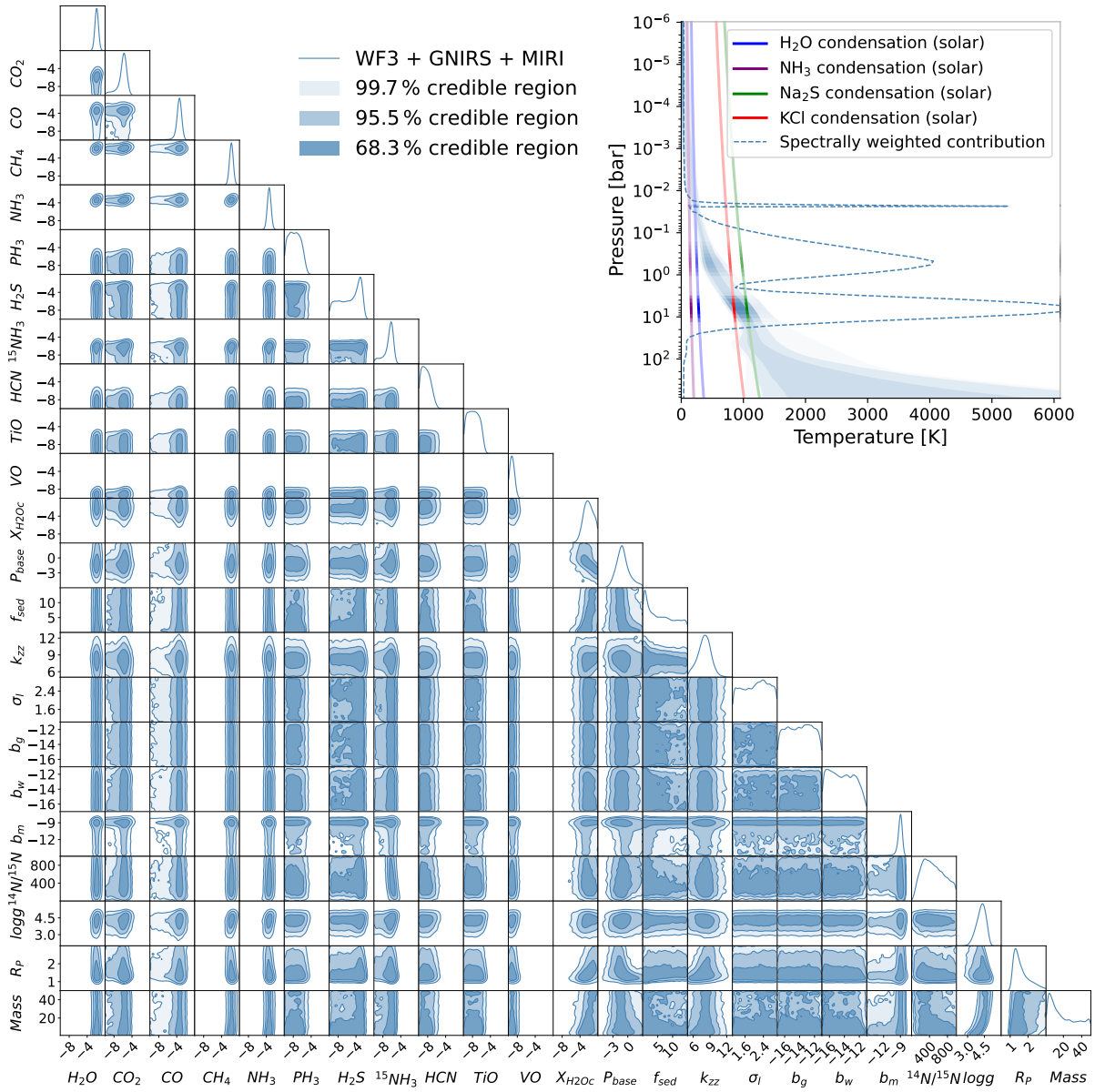


Figure 6.1: *Left.* Cloudy retrieval using neural posterior estimation on the WISE 1738 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1738 observations x_{obs}, WFC3+GNIRS+MIRI spectra. *Right.* The top right figure illustrates the posterior distribution of the P-T profiles, that has the contribution function overlaid on top of it in shades of white highlighting visibility. The equilibrium state water ice and liquid condensation curves are plotted along the profile in red and green (dashed) respectively. The bottom right plot shows the cloud particle size distribution in the atmosphere of WISE 1738 for the most probable cloudy model.

between 6 and 7. The cloud particle size distribution remains poorly constrained, Furthermore,

the mass is not constrained and exhibits a right-skewed distribution. This is similar to the radius, which is also right-skewed but is broadly constrained. The lack of a confident constraint on the radius of WISE 1738, unlike in the cloud-free retrieval, suggests a degeneracy between the cloud parameters and the radius.

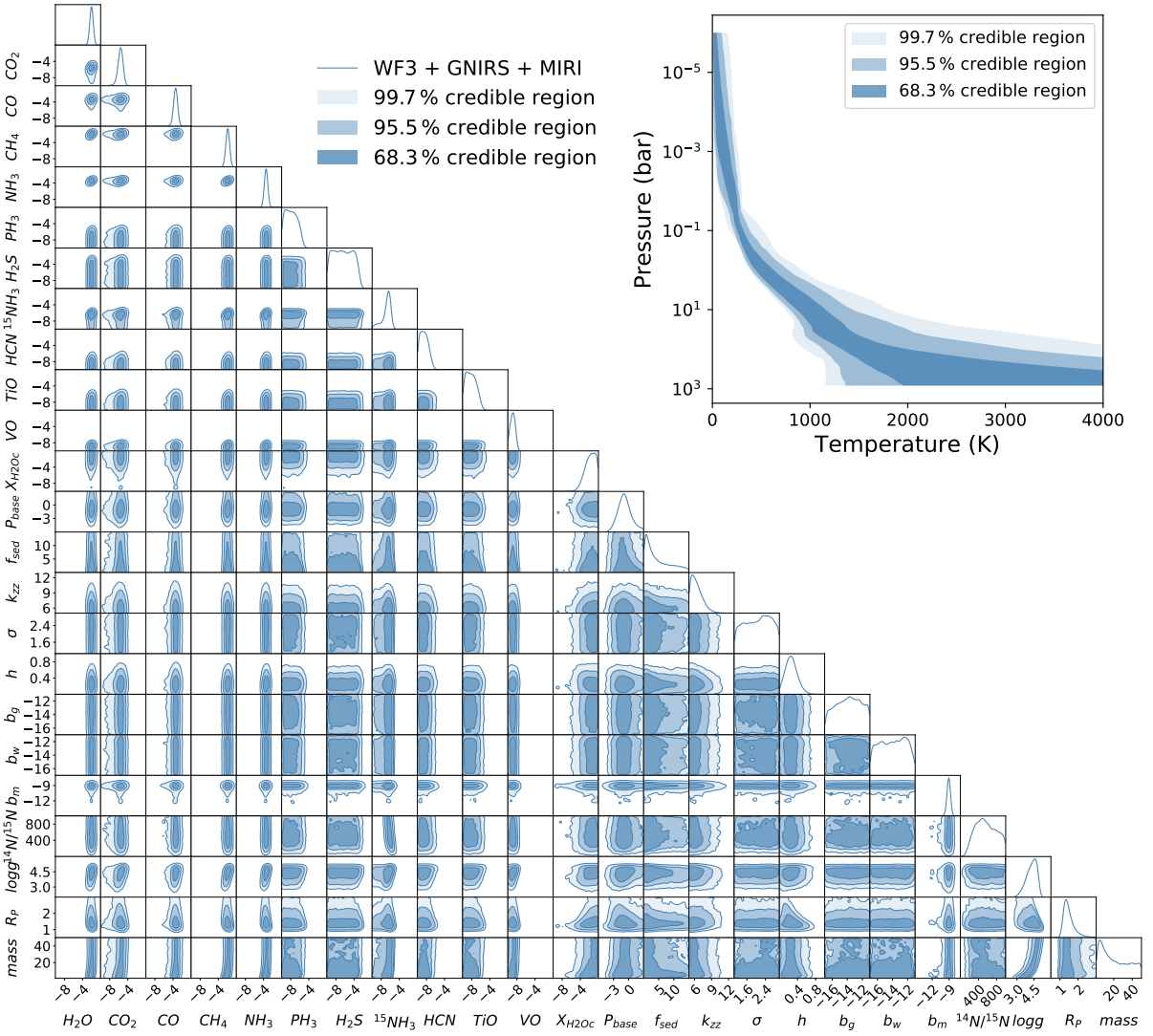


Figure 6.2: *Left*. Patchy model retrieval using neural posterior estimation on the WISE 1738 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1738 observations x_{obs}, WFC3+GNIRS+MIRI spectra. *Right.* The top right figure illustrates the posterior distribution of the P-T profiles.

The patchy model retrieval results are illustrated in Figure 6.2. The results reveal that, just as in the cloudy retrieval, similar abundances are constrained (or unconstrained) as in the cloud-free case, along with similar metallicity, C/O ratio and age estimates, with slight variations in $^{15}NH_3$ leading to a lower $^{14}N/^{15}N$ ratio, peaking at 500. Water ice clouds exhibit a mass fraction abundance of $X_{H_2Oc} = -1.71^{+1.14}_{-1.36}$, forming at a base pressure of $\log P_{\text{base}} = -0.91^{+1.4}_{-1.22}$.

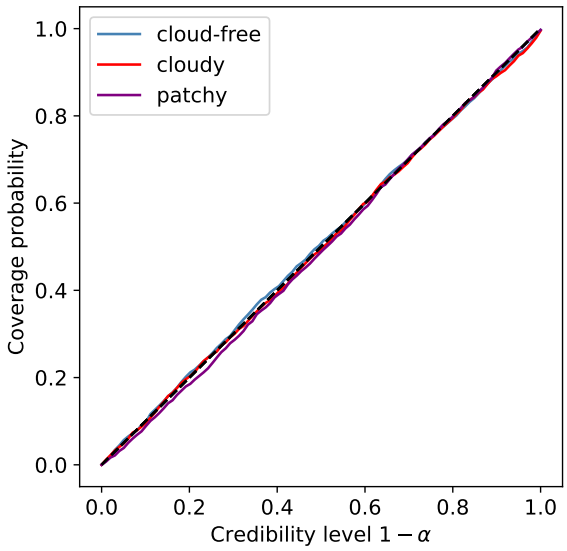


Figure 6.3: Coverage plots for cloud-free, cloudy and patchy model retrievals of WISE 1738 (same as in Fig. 4.10).

However, the vertical distribution and mixing coefficient remain largely unconstrained, as indicated by the $f_{\rm sed}$ and $\log k_{\rm zz}$ respectively. Furthermore, the radius is constrained at $1.4^{+0.35}_{-0.19}$, and it exhibits a right-skewed distribution, implying degeneracy with clouds similar to the cloudy case, albeit not so strongly. Additionally, the h scaling factor is constrained to $0.25^{+0.14}_{-0.12}$, which suggests a hole fraction of 25% or a 75% cloud cover.

6.4 Validation

Coverage

The coverage plot for the cloudy and patchy cloud models are shown along with the cloud-free model in Figure 6.3, illustrating that the respective posterior estimates are appropriately dispersed.

Consistency plot

The consistency plots for the cloudy and patchy models are plotted in Figures 6.4 and 6.5 respectively. The cloudy model plot reveals very broad posteriors, as indicated by the large spread in residuals. We also see a "blanket" of featureless spectra due to thick clouds, which are considered plausible under the estimated posterior. In contrast, the patchy model has a nar-

rower posterior predictive distribution, although the blanket features are similarly observed. Further, the near-infrared region of the spectrum in the patchy model is much wider than that of the cloudy model and vice versa. This suggests that the patchy model better represents the mid-infrared region of the spectrum than the near-infrared region.

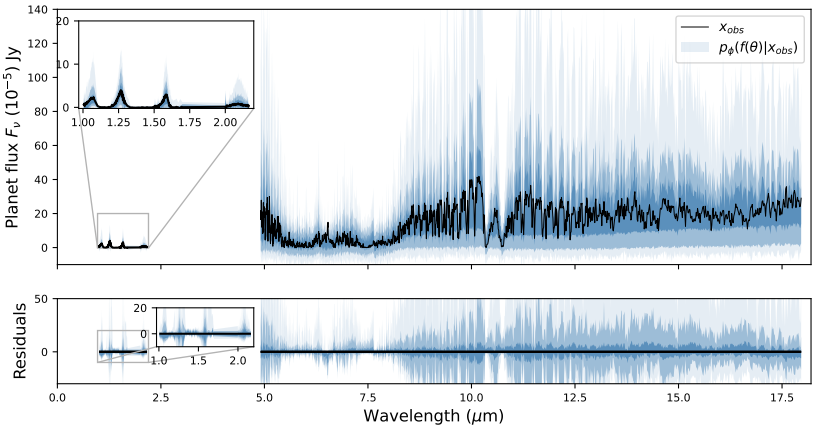


Figure 6.4: WFC3+GNIRS+MIRI cloudy consistency plot. The posterior predictive distribution $p(f_{\text{cl}}(\theta) + \epsilon|x_{\text{obs}})$ of noisy simulations spectra for the 99.7%, 95% and 68.7% quartiles (hues of blue), overlaid on top of the the WFC3+GNIRS+MIRI observation x_{obs} (black line). *Bottom.* Residuals of the posterior predictive samples, normalized by the inflated standard deviation of the noise distribution for each spectral channel and a horizontal line at 0 for reference (in black)

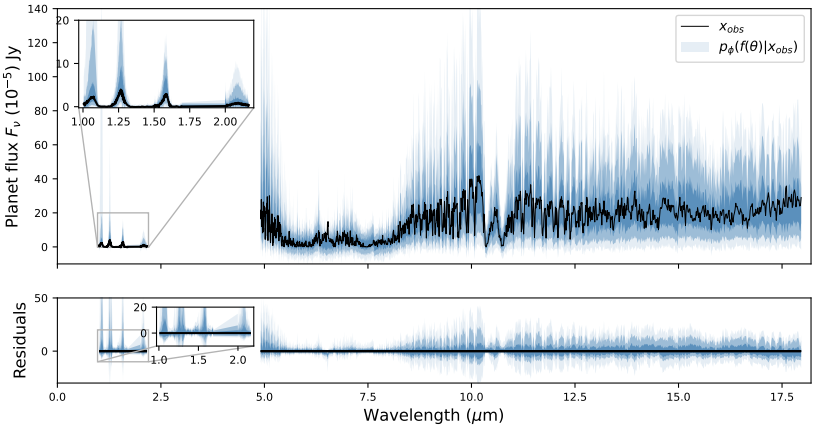


Figure 6.5: Patchy retrieval consistency plot for the brown dwarf WISE 1738. We plot the posterior predictive distribution $p(f(\theta) + \epsilon|x_{\text{obs}})$ of the posterior estimate of the patchy model for different confidence levels, and overlay it on the observation spectrum.

6.5 Discussion

6.5.1 A study of cloudy posterior uncertainty

Since the radius, clouds, certain thermal profile parameters, and mass are the only parameters with significantly broader marginal posteriors (different from the cloud-free case), we conclude that a lack of confident constraints on these parameters is driving the observed uncertainty. To better understand the underlying cause of their width, we investigate the key sources of uncertainty that one encounters in Bayesian model parameter estimation. Our goal is to identify the dominant contributor(s) of uncertainty in the context of our cloudy posterior and to improve upon them.

Posterior uncertainty in our framework can arise from different sources, influencing the robustness of the retrieved properties. Based on its origin, we can broadly classify estimated posterior uncertainty into three categories: methodological, epistemic, and aleatoric (Gansch and Adee, 2023).

Methodological uncertainty

Methodological uncertainty refers to the uncertainty arising from the retrieval algorithm, which may be biased.

This uncertainty only arises in the posterior estimate and is not accounted for in the true posterior, since the inference network is not a perfect approximation of the posterior distribution. The methodological uncertainty could arise from an unsuitable flow neural architecture, or an insufficient number of model parameter-spectral pairs in the training dataset or lack of convergence. Additionally, methodological uncertainty can also result from computational uncertainty (Delaunoy et al., 2024).

To verify the suitability of the architecture, we start by training on the top twelve best architectures (i.e ones leading to lowest validation losses), and also average over 6 runs of the best architecture to improve stability and rid ourselves of uncertainties arising from random initialization. Neither of these efforts reduces the posterior uncertainty. Further, since the architecture we use is identical to the one used for the cloud-free retrieval which results in a much narrower posterior, we conclude that the neural architecture is not the dominant contributor of uncertainty.

To test for convergence, we assess the reliability of the trained posterior estimator by conducting a global coverage test (see in Figure 6.3). While the test implies a valid posterior estimator on average, hence implying convergence, this doesn't test for local validity on the specific observation spectrum of WISE 1738. This can be further tested using the LC2ST test. Additionally, we note that the posterior predictive distribution, although it generalizes badly,

is still consistent with the observations, as it falls within 1σ of the posterior predictive distribution. Furthermore, we also triple the data size, expanding from 3.7 million joint pairs of model parameters and simulated emission spectra to 11.1 million sample pairs to train the flow network for a longer number of epochs. However, this training set expansion does not significantly shrink the posterior width.

Although these tests do not rule out the presence of methodological uncertainty, they rule it out as being a dominant contributor to the broad posterior we observe.

Epistemic uncertainty

Epistemic uncertainty refers to a limited knowledge or imperfect modeling of a system.

This uncertainty arises from incomplete or imperfect representations of a physical system within the model, or incomplete/uninformative data, or limited prior information, which leads to uncertainty in the model's predictions. When the assumed model structure (likelihood, priors, functional form, etc.) does not reflect the true data-generating process, such a case is called model mis-specification. In essence, all models are mis-specified, except some are more mis-specified than others. The uncertainty arising in such cases is reducible through the use of improved forward models, refined priors, and informative data. In an attempt to identify if this dominates our posterior uncertainty, we run retrievals on different models and refine priors.

Forward model :

In general, any inference performed is only valid in the context of the assumed model, since changing the model may change inference results on individual parameters. This holds true even if these parameters are supposed to represent the same physical quantity. This can be seen when we compare the cloudy model retrieval with the patchy case in Figure 6.6. Here we see that cloud parameters do not agree between the two models. Furthermore, this also explains why the cloudy model does not opt for "cloud-free" structures in its retrieval, and that the constraint on radius does not agree between the cloud-free and cloudy-patchy retrievals. This variation exposes epistemic uncertainty in these models.

To improve upon this epistemic uncertainty, we begin by making small changes to our forward model, such as excluding TiO, VO, HCN, and alkali species. This is because certain absorber species may be degenerate with the clouds, which can explain the broad posteriors on them. Since these species are not constrained (i.e not detected), we remove them. However, removing them does not lead to a narrow constraint on the clouds, thus making us look to other aspects of the model.

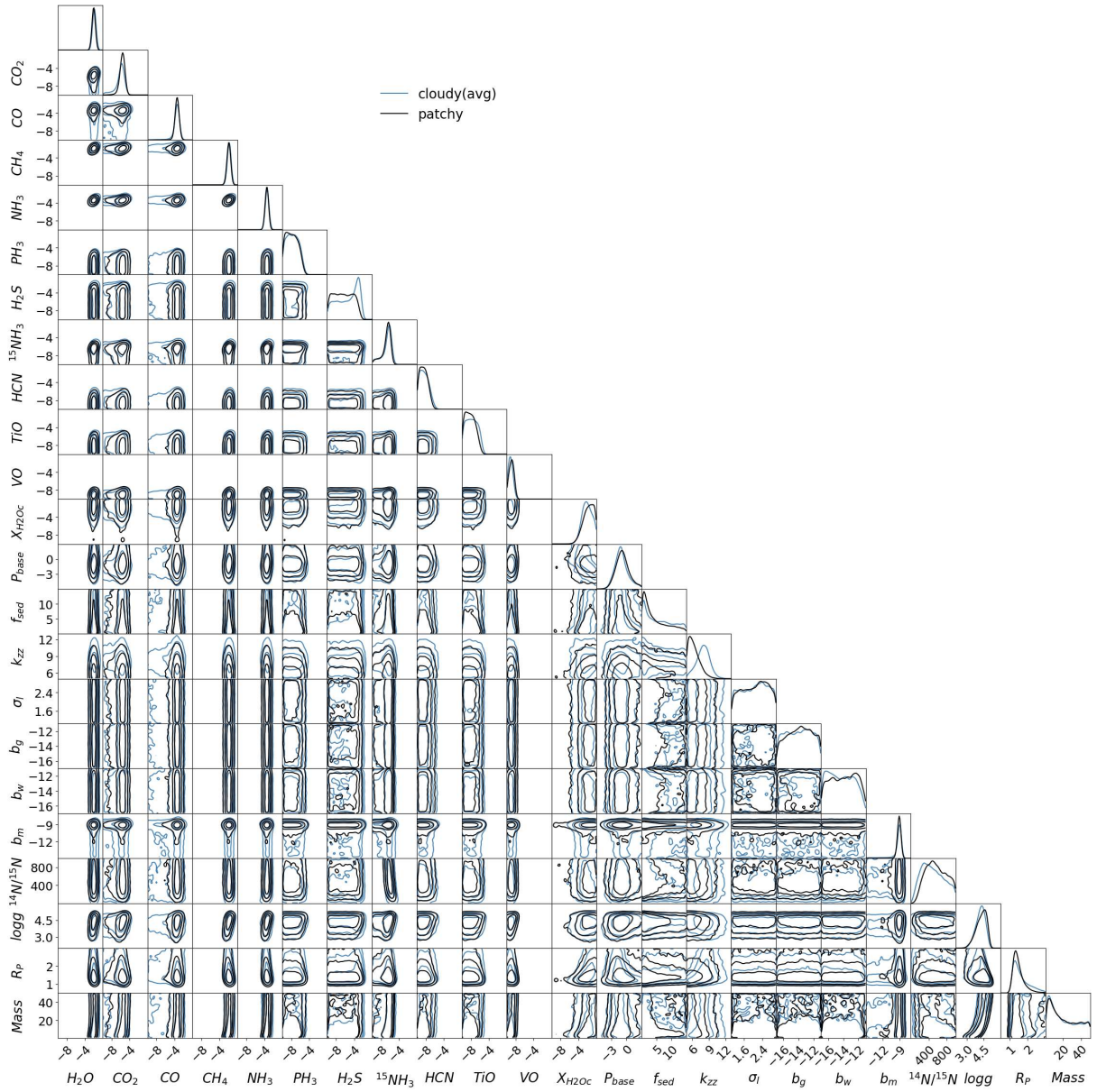


Figure 6.6: Comparison of the cloudy and patchy model retrievals.

Flat likelihoods can lead to flat posteriors. In our model, we use a scaling factor that artificially flattens our likelihood. We do this to accommodate any non-quantified measurement errors that are not accounted for. In order to test if this artificial flattening of likelihood is the main source of posterior uncertainty, we remove the error bar scaling factor from the model. This presents training issues, such as NaNs in the training loss, suggesting that the likelihood is too narrow to retrieve the noisy data at hand. Therefore, we include them in the model; however, we shrink their priors significantly.

Prior :

We use broad uninformed priors in our retrievals in order to not introduce a bias in our characterization of the brown dwarf. However, these broad priors can result in wide posteriors on model parameters, thus we shrink the priors on several parameters. By shrinking the prior on the scaling factor, we improve the model's ability to attribute all the changes in the spectrum purely to changes in the model parameters. Although this does not help parameters that already exhibit a flatter likelihood. However, this approach does not improve the flat marginals or the wide residuals. This suggests that the artificial flattening of the likelihood is not the dominant contributor to the width of the original posterior, but an inherent likelihood of these parameters is, hence making them difficult to constrain.

Additionally, we shrink the prior on the cloud parameters X_{H_2Oc} and P_{base}, to account for the possibility of either thin clouds above the photosphere or thick clouds below it. The thin cloud model does not provide meaningful constraints on cloud properties, as the likelihood in this region is completely flat resembling the prior, meaning that the clouds do not influence the shape of the spectrum at all. In contrast, thick clouds help constrain the cloud properties in a way similar to the full cloudy case. In the thin cloud scenario, the "blanket" features disappear, while they persist in the thick cloud scenario.

Flat parameter likelihood :

Furthermore, to test which parameters have an inherently flatter likelihood, or have weak identifiability, we run retrievals over several simulated spectra from the cloudy prior and compare the retrieved values to the nominal ones. Firstly, we evaluate whether there is enough information in this spectral range to effectively constrain water clouds, provided the noise model is known and there is no model-data mismatch. In our results, we find that while these retrievals are consistent with the nominal case, the *P-T* profile outside the probed photosphere (here from log P_{base} -2 to 2 typical for direct imaging (Marley et al., 2012; Morley et al., 2014; Zalesky et al., 2019)) and cloud properties remain difficult to constrain. This suggests that while the model can align with nominal expectations, the clouds and some of the thermal profile parameters are inherently hard to constrain based on the available data from this wavelength range.

Additionally, we run a retrieval on the same atmosphere with the clouds turned off in the synthetic observation. The absence of clouds is reflected in the lack of constraints on the cloud parameters in the posterior, which resembles the prior distribution. This implies that clouds (or their absence) are accurately detected by this model. Given that clouds are detected in WISE 1738, albeit with huge uncertainties, this raises the question of their presence in its atmosphere. The hypothesis that 'clouds are present but difficult to constrain' is not entirely far-fetched since the water condensation curve intersects with the *P-T* profile at the upper end of the photosphere, (see Figure 6.1).

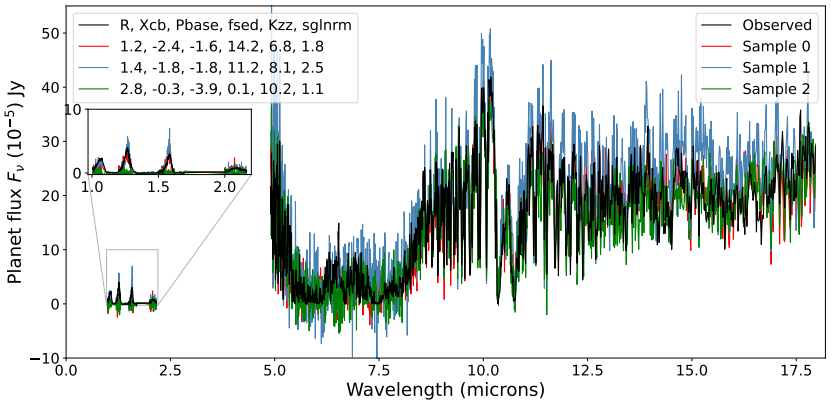


Figure 6.7: Degeneracy between radius and clouds. The three samples from the cloudy posterior of WISE 1738, each with similar likelihoods but different combinations of radius and cloud properties, demonstrate degeneracy between radius and clouds.

Degeneracies :

While flatter likelihoods may hold true for clouds, the thermal profile and mass parameters, it does not hold true for radius. In general, the radius is known to heavily impact the global brightness of the object over longer wavelength regions such as the combined spectrum from the near and mid-infrared. Following this, we identify the 2D marginal posterior over radius and cloud mass fraction (see Figure 6.1) to exhibit a "banana-shaped" distribution, indicating degeneracies between them. Further, we observe that, across all 12 architectures and the six repeated runs on a single architecture, the radius is constrained between 1.2-2.8 $R_{\rm Jup}$ and the water cloud mass fraction X_{H_2Oc}, cloud base pressure $P_{\rm base}$ takes values such that the algorithm explores various atmospheric scenarios with thinner clouds at the photosphere for a brown dwarf with a radius of 1.2 $R_{\rm Jup}$, or very thick clouds above the photosphere for a brown dwarf with a radius of 2.8 $R_{\rm Jup}$, and several possibilities in between. Therefore, even with a tighter likelihood, if the model cannot tell which specific combination of the correlated parameters is correct, it would stretch the posterior. The degeneracy between radius and clouds is illustrated in Figure 6.7. We observe this degeneracy also in the MIRI (cloudy) retrievals of several other brown dwarfs (see Figure 7.10). Degeneracies such as this have been found in other studies too (Novais et al., 2025). In order to break this degeneracy, we reduce the prior on the radius based on theoretical insights from evolutionary models, and find that shrinking the prior on the radius breaks the degeneracy and shrinks the posterior width significantly. However, it still retrieves thick clouds that lead to blanket features.

Aleatoric uncertainty

Aleatoric uncertainty is the inherent variability in the data itself (e.g., measurement noise or randomness in the data collection process or the system being studied) that is modeled using a probability distribution.

Aleatoric uncertainty pertains to detector noise and thermal background variations, wavelength calibration errors, and other instrumental systematics that we choose to model with a Gaussian likelihood. It is inherent to the data collection process and cannot be mitigated, but can be reduced by better instrumentation, increased exposure time, or improved calibration techniques, so in general better quality data.

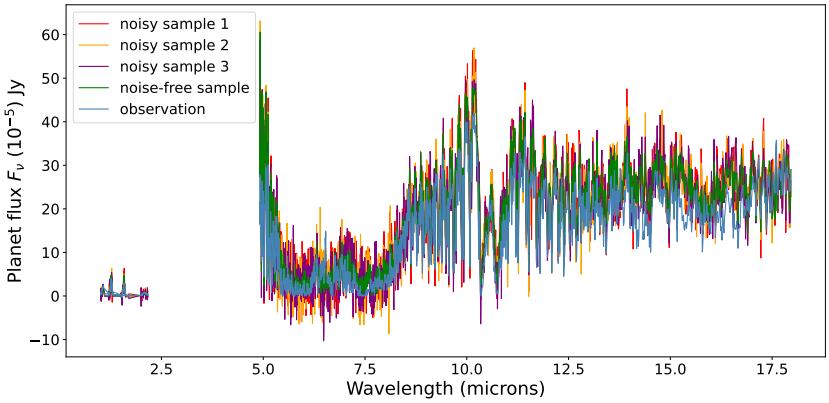


Figure 6.8: Three noisy instances and a noise-free simulated cloudy spectrum corresponding to the most probable sample from the cloudy retrievals, plotted alongside the observed spectrum of WISE 1738.

To investigate whether aleatoric uncertainty is another dominant factor in this case, we perform retrievals on both noise-free and noisy simulated spectra (using three different Gaussian noise instances) corresponding to the most probable posterior sample from the cloudy retrieval, which also best resembles the observation (see Figure 6.8). This allows a clearer assessment of how noise contributes to posterior uncertainty in our retrieval.

Comparing the noise-free synthetic spectral retrieval with the three noisy synthetic spectral retrievals in Figure 6.9 suggests that the $^{14}N/^{15}N$ ratio due to $^{15}NH_3$, H_2S, and the mass, are very sensitive to noise, based on their retrieved variation. In general, the noise-free retrievals are more confident than the noisy ones, as expected. However, the noise-free simulated spectral retrieval still constrains broad marginal posteriors over the cloud parameters, therefore, aleatoric uncertainty is not a dominant contributor to the posterior uncertainty.

More interestingly, none of the noisy or noise-free retrievals show any degeneracy between radius and clouds. Furthermore, the three noisy sample retrievals provide stable constraints on the radius and clouds between the three retrievals, unlike on the WISE 1738 spectrum. This suggests that the reason for unstable solutions between the different WISE 1738 retrieval runs (even within the same architecture) and the degeneracy between radius and clouds could be

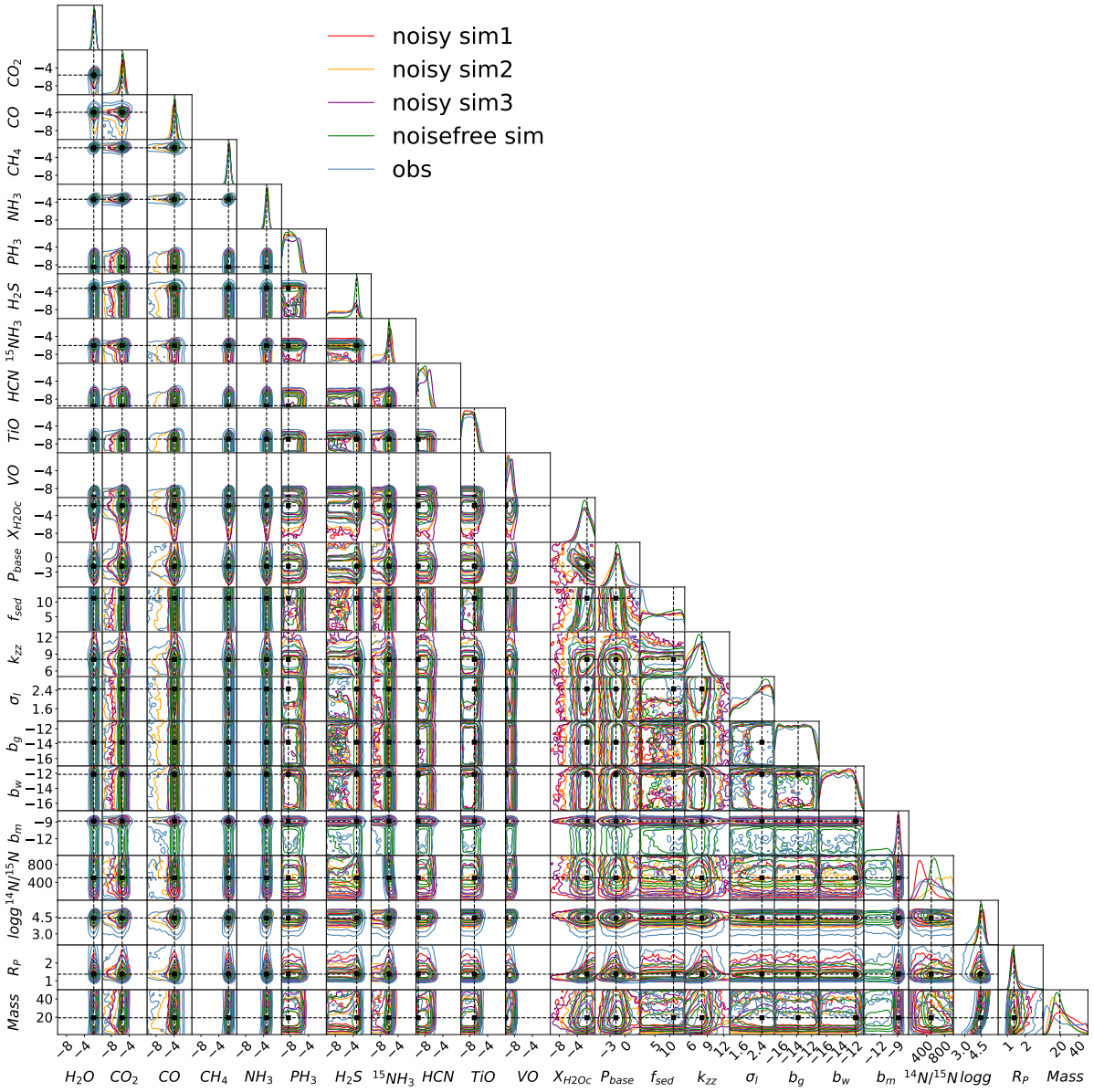


Figure 6.9: Cloudy retrieval using neural posterior estimation on three noisy instances and a noise-free simulated cloudy spectrum corresponding to the most probable sample from the cloudy retrievals of the WISE 1738 spectrum. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the simulated cloudy spectra x_{obs} WISE 1738.

driven by our insufficient assumption of a Gaussian noise model or stochasticity within the system or instrument systematics. The former is an epistemic cause that could be contributing to degeneracies in the posterior that consequently broadens it. Such insufficient assumptions are more pronounced in noisier observations, where we see a positive correlation between degeneracies and constraints on b factor values also in other brown dwarf retrievals. See section 7.2 for more details.

The above experiments suggest that the epistemic uncertainty is a dominant factor in the cloudy posterior. Although this analysis is not exhaustive, it suggests that the broad posteriors can arise from model mis-specification causing an inherent difficulty in constraining cloud/thermal parameters in this wavelength range (i.e flatter likelihood), and degeneracies between the parameters. A retrieval can only constrain parameters when their impact on the observable data is clear and distinguishable from other effects.

This presents an opportunity for improving modeling of the atmosphere of WISE 1738. The atmosphere appears suitable for the formation of Na_2S and KCl clouds, with their condensation curves intersecting the *P-T* profile in the lower photosphere. However, previous retrievals have included these clouds and they have not been detected. Notably, this has not yet been explored using a combined retrieval approach. A comprehensive retrieval study incorporating these clouds would be particularly interesting. Additionally, near-infrared data in the 2–5 μm range, as well as data beyond 18 μm, could provide valuable insights into larger regions of the atmosphere in the future.

6.5.2 Comparing NPE posterior uncertainty with MultiNest

To benchmark the NPE retrievals, we compare the estimated posteriors from a retrieval only on the MIRI spectrum using a similar cloudy forward model and the `MultiNest` algorithm. The results of this comparison are shown in the Figures 6.10 and 6.11. It is however important to note that the forward models slightly differ between the two retrievals in two main ways. Firstly, different line lists are used for the molecules. While the NPE's forward model uses the ExoMol database (Polyansky et al., 2018), the NS forward model uses the HITEMP database (Barber et al., 2006). Secondly, the NS forward model includes a γ factor to smoothen (or regulate) the *P-T* profile to prevent un-physical thermal inversions in the atmosphere, whereas NPE uses an un-regulated *P-T* profile.

The comparison between the two retrievals suggests that the `MultiNest` retrieval constrains a smaller posterior than NPE. This could be either due to differences in their forward models or posterior over-confidence. While it is not computationally efficient to ascertain the dispersion of the `MultiNest` posterior, we ascertain if the NPE posterior is over-dispersed by plotting a coverage plot in Figure 6.11. This suggests that, on average, the NPE posterior is slightly over-

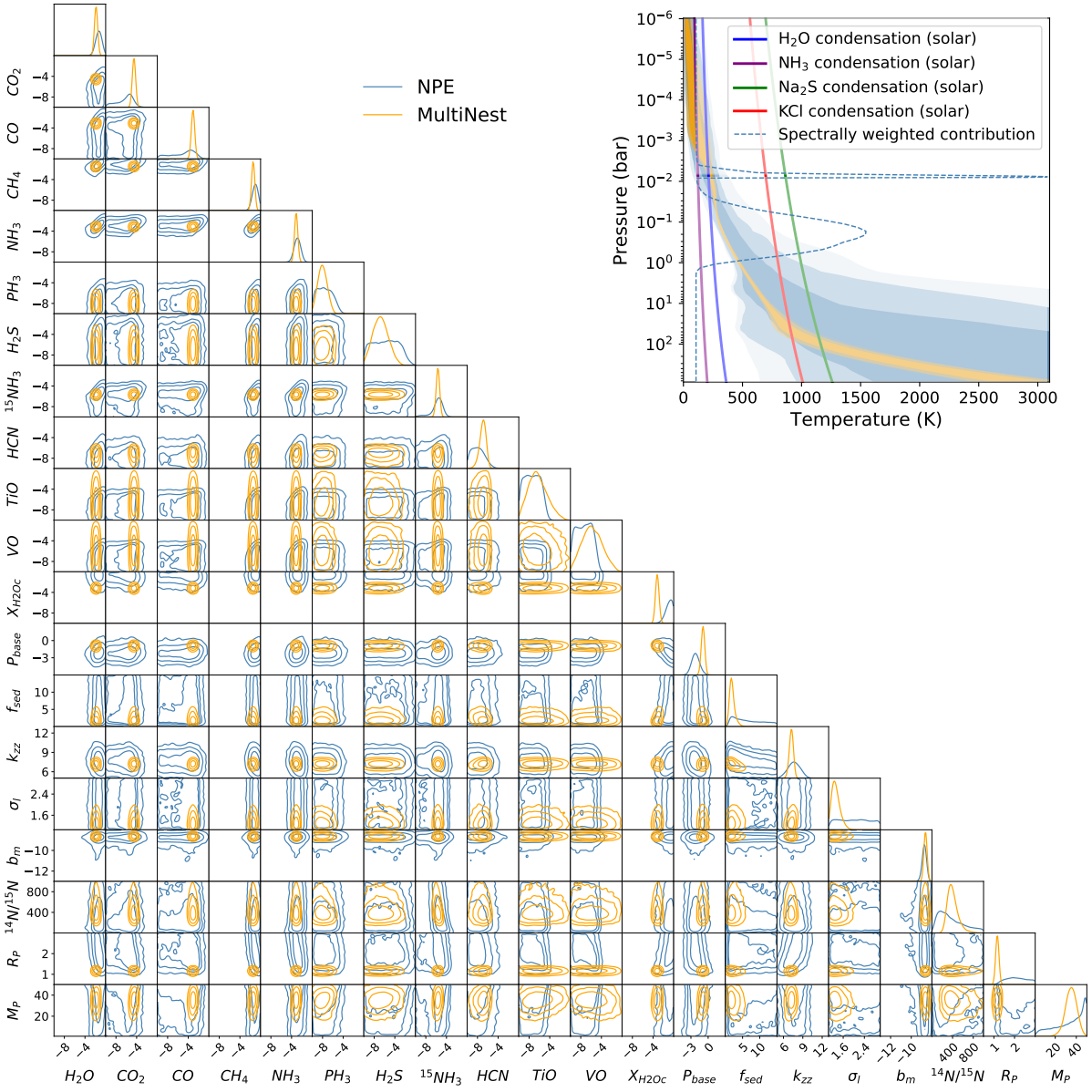


Figure 6.10: *Left.* Cloudy retrieval using neural posterior estimation (in steelblue) and `MultiNest` (in orange) on the WISE 1738 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1738 observations x_{obs}, MIRI spectra. *Right.* The top right figure illustrates the two posterior distributions of their corresponding P-T profiles, that has the contribution function overlaid on top of it in shades of white highlighting visibility. The equilibrium state water ice, ammonia, chloride and sulphide condensation curves are plotted along the profile in blue, purple, red and green (dashed) respectively.

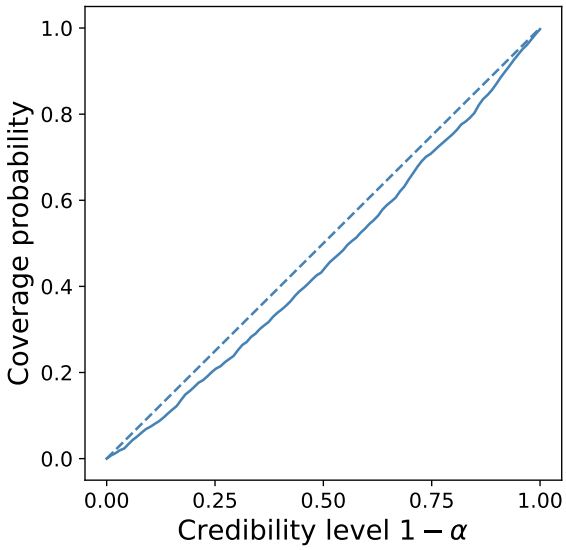


Figure 6.11: Coverage for the NPE retrieval on the MIRI spectrum of WISE 1738 using the cloudy model.

confident, albeit not significantly. However, it is important to note that while coverage verifies posterior widths on average, it does not find local validity. If the NPE posteriors are found to be locally over-dispersed, it could be because it learns the posterior estimator instead of learning the posterior conditioned on a specific observation.

Finding local validity is not straightforward, and even less so in `MultiNest`. Even though some studies such as Vasist et al. (2023); Ardévol Martínez et al. (2022a) have previously hinted at the over-confidence of `MultiNest`, in general, such evaluations are computationally unfeasible and the posterior uncertainties generated are difficult to verify. Recent wide-wavelength studies have found constraints on certain parameters to fall outside the error bars of previously constrained parameters. If the variation in constrained values are not due to epistemic uncertainty, they too hint at `MultiNest` posterior over-confidence. Further, from our experiments we find that the consistent difference in widths between the NPE posterior and the `MultiNest` posterior is more pronounced in cases where there might be a significant model-data mismatch. NPE seems to be more sensitive to such models than `Multinest`, which also suggests `MultiNest` over-confidence. This over-confidence could be due to not sampling enough from the prior space during inference. However, recently there have been efforts to improve the estimation of model uncertainties in `MultiNest` posteriors by Bayesian model averaging using leave-one-out predictive densities and stacking of predictive distributions (Nixon et al., 2024; Wakeford et al., 2017).

6.5.3 Bayes Factor

To quantitatively assess the relative consistency of the different models we have used with the observations, we conduct Bayes Factor analyses. Bayes factor is a statistical measure used in Bayesian inference to compare the relative evidence provided by data for two competing hypotheses or models. Specifically, it is the ratio of the marginal likelihoods (also known as the evidence) of two models, usually denoted as M_0 (the null hypothesis or baseline model) and M_1 (the alternative hypothesis or competing model). The Bayes factor BF_{10} comparing model M_1 against model M_0 is given by:

$$\mathrm{BF}_{10} = \frac{p(\mathrm{Data}|M_1)}{p(\mathrm{Data}|M_0)}$$

where:

- $p(\mathrm{Data}|M_1)$ is the marginal likelihood of the data under model M_1.
- $p(\mathrm{Data}|M_0)$ is the marginal likelihood of the data under model M_0.

The Bayes factor test in NPE isn't as straightforward to access as it is in NS, where the Bayesian evidence is a natural outcome of the algorithm. To address this, we employ a binary classifier to estimate the Bayes factor. This was first introduced in Kim and Rockova (2023), and has since also been further implemented in exoplanet retrievals in the recent study by Lueber et al. (2025), albeit with slight modifications.

To estimate the Bayes factor between two competing models, we first construct a training dataset consisting of two classes that include, samples from the joint distribution $p(\theta, x|M_1)$, where θ represents the parameters of model M_1 and x is the corresponding model output, are labeled as class 1, and, samples from the joint distribution $p(\theta, x|M_0)$, corresponding to model M_0, are labeled as class 0. A CNN classifier f is trained to distinguish between these two classes. Here $p(\theta, x|M_1) + p(\theta, x|M_0) = 1$. The optimal Bayes classifier is given by,

$$f^*(\theta, x|M) = \arg\max \{1 - d^*(\theta, x|M),\ d^*(\theta, x|M)\}$$

where,

$$d^*(\theta, x|M) = \frac{p(\theta, x|M_1)}{p(\theta, x|M_1) + p(\theta, x|M_0)}.$$

Once the classifier is trained, the Bayes factor can be computed by evaluating the classifier on a test sample and taking the ratio of its outputs:

$$\text{Bayes Factor} = \frac{f(\theta, x)}{1 - f(\theta, x)}.$$

This is called the likelihood ratio trick. A $\log_{10}$ Bayes factor greater than 0 indicates that the evidence favors the alternative hypothesis (model in class 1), while a Bayes factor less than 0 indicates a preference for the null hypothesis (model in class 0). A Bayes factor equal to 0 suggests no preference between the two hypotheses. Once trained, the classifier can predict the Bayes factor between them for any given observation. All the Bayes factors (in log scale to the base 10) between the models are tabulated in Figure 6.12 along with their corresponding classifier accuracies calculated on the testset.

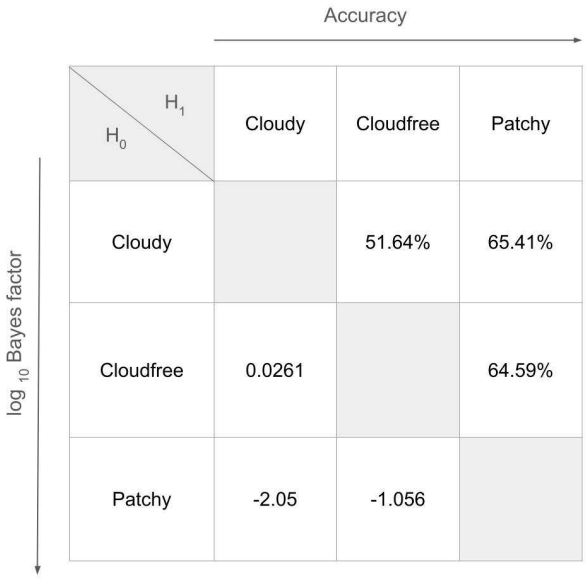


Figure 6.12: $\log_{10}$ Bayes factor values for all the model comparisons are obtained from the evidence provided by the WISE 1738 spectrum (down), along with the accuracies achieved by the classifier (across). The null hypothesis H_0 is given along rows and the alternate hypothesis H_1 across columns. The Bayes Factor values are read as $BF_{column,row}$.

Comparisons between cloud-free and cloudy models result in a $\log_{10}$ Bayes factor close to 0, suggesting that the likelihood of the observed spectrum belonging to each model is the same under both hypotheses being tested. This means that there is no preference for either hypothesis since the evidence provided by the classifier is equally supportive of both. However, the preference for the cloud-free model increases if a distinction is made between cloud-free and very cloudy models. Additionally, comparisons between the patchy model and both the cloud-free and cloudy models indicate a preference for the patchy hypothesis.

The preference for the patchy model, despite having six more parameters than the cloud-free model, contrasts with the less confident parameter estimates compared to the cloud-free case. This suggests that the patchy model is overfitting the data due to its greater model flexibility. Given the limited data, the simpler cloud-free model is preferred, since it generalizes better (i.e narrower posterior predictive distribution). As a result, despite the statistical preference for the

patchy model, the cloud-free model may be the more reliable choice for making predictions. Further, the reduced χ^2 values over the (noise-free) most probable sample from the posterior of the patchy, cloud-free and cloudy models are 1.603, 2.376 and 3.126 respectively, which follows the same preference as the Bayes factor test.

6.6 Importance sampling

Biases may occur in the estimated posterior $q(\theta|x)$ due to using a suboptimal flow network, insufficient training samples, or by achieving incomplete convergence during training. This can be refined through importance sampling (IS) to better resemble the true posterior $p(\theta|x)$. IS is performed by reweighting the estimated posterior samples $q(\theta_i|x)$ using importance weights w_i, which are defined as the ratio of the true posterior and the estimated posterior $w_i \propto \frac{p(\theta_i|x)}{q(\theta_i|x)}$. This boosts the samples that are highly probable under the true posterior if they are underrepresented in the estimate, while down-weighting the less probable samples if they are overrepresented. Since the true posterior $p(\theta_i|x)$ is unknown, and $p(\theta_i|x) \propto p(x|\theta_i)p(\theta_i)$, we calculate the importance weights as $w_i \propto \frac{p(x|\theta_i)p(\theta_i)}{q(\theta_i|x)}$. These weights are normalized to ensure $\sum_{i=1}^{N} w_i = N$.

Although IS improves sample quality, it requires additional likelihood evaluations, which can be parallelized but must remain computationally tractable. In NPE (and other simulation-based inference algorithms), IS acts as a "cheat code" for refining posterior estimates, if the likelihood can be evaluated explicitly. Additionally, IS allows for the estimation of the effective sample size (ESS) as,

$$N_{\text{eff}} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}$$

with sampling efficiency $\epsilon = N_{\text{eff}}/N$. While a high ϵ indicates a well-matched proposal distribution, a lower one doesn't necessarily mean a bad approximation, especially in high dimensions. An efficiency of $\geq 1\%$ is usually considered good. Further, IS can also estimate the Bayesian evidence Z as,

$$Z \approx \frac{1}{N} \sum_{i=1}^{N} w_i,$$

which is useful for model comparison.

A review of this technique can be found in Tokdar and Kass (2010). There have been many applications of importance sampling using various algorithms such as Nested sampling (Lange, 2023; Feroz et al., 2019a), Sequential NPE (Zhang et al., 2023) and flow matching posterior estimation (FMPE, Gebhard et al. 2025). In order to ensure a bias-free posterior estimate, we apply importance sampling to our cloudy retrieval based on the implementation in Gebhard et al. (2025).

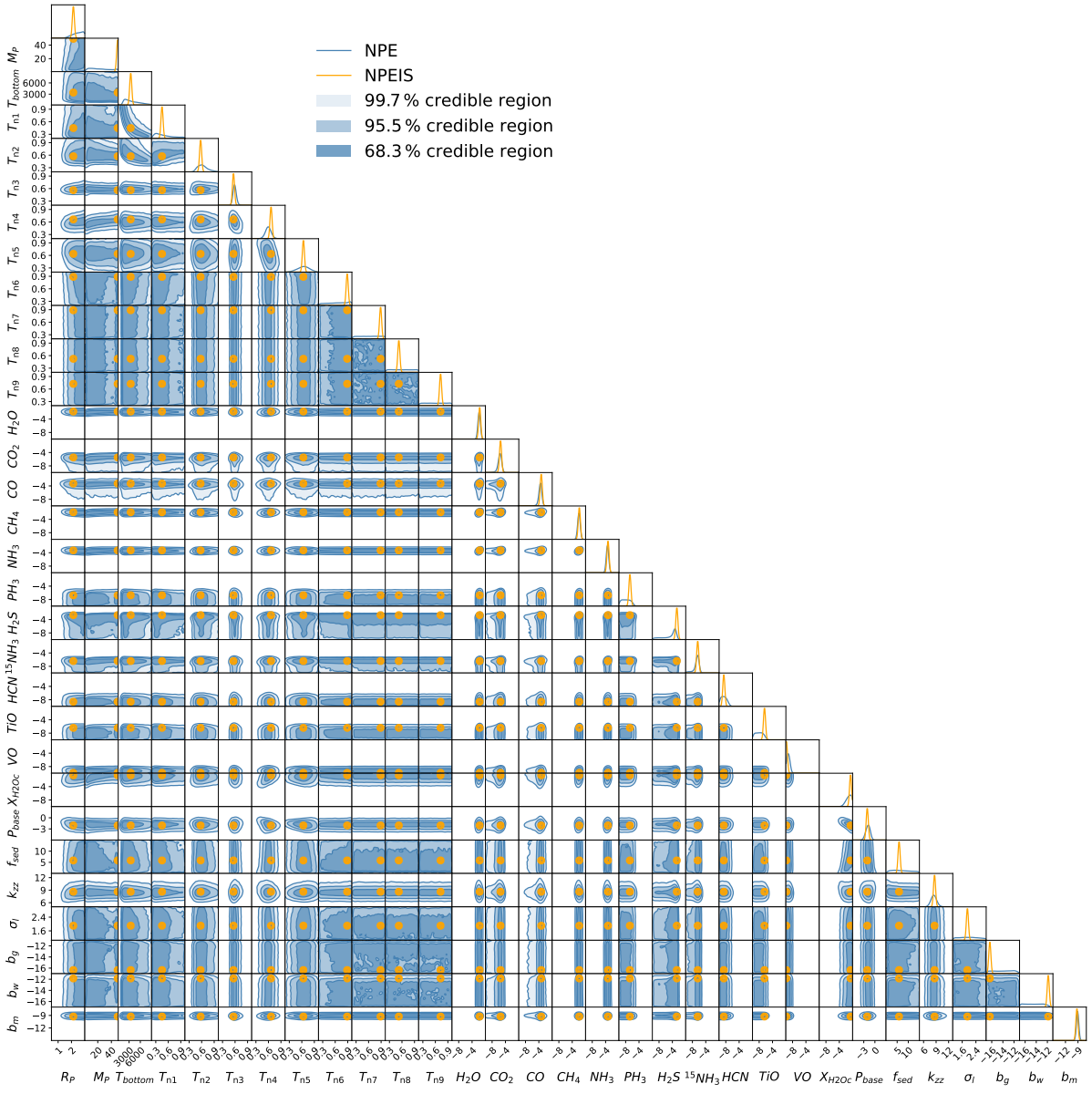


Figure 6.13: Comparing the cloudy NPE retrieval (in steelblue) with its importance-sampled retrieval (in orange) on the WISE 1738 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1738 observations x_{obs}, WFC3+GNIRS+MIRI spectra.

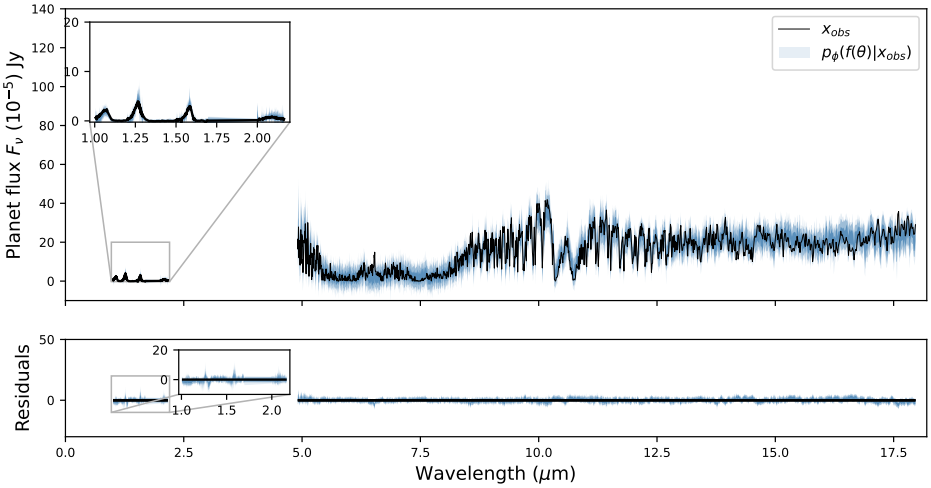


Figure 6.14: Importance samples cloudy consistency plot.

The IS posterior for the cloudy model is presented in Figure 6.13 using 100k samples from the posterior estimate. The IS posterior shrinks significantly compared to the original NPE posterior. Notably, we do not observe any double modes or degeneracies related to radius and clouds, nor do we find flat marginal posteriors for any parameter. This substantial shrinkage suggests that IS is primarily capturing the highest-likelihood regions of the true posterior, indicating either that the weight distribution is highly skewed, or that we aren't sampling enough. This is evident in the sampling efficiency, which is extremely low, almost approaching zero. Further, the corresponding posterior predictive distribution (PPD), shown in Figure 6.14 remains both confident and consistent with the observations. Moreover, the IS posteriors of retrievals using different priors over the cloud parameters, radius and noise scaling, all lead to shifted peaks and narrow posterior predictive distributions similar to Figure 6.14, further proving epistemic uncertainties.

Furthermore, we bias the IS cloudy posterior by applying a clipping procedure where we set weight percentiles beyond 1 and 99 to their respective boundary values. In other words, we assign the least likely samples the same weight as the 1st percentile sample and the most likely samples the same weight as the 99th percentile sample. This essentially reduces the dominance of the highest-likelihood regions and boosts the lower-likelihood regions. A renormalization of the biased weights essentially brings attention to the bulk of the posterior away from the peaks and tails. This outlier *clipped* IS marginal posterior resembles the marginal NPE posterior. This suggests that NPE captures the tails of the posterior distribution along with the bulk. Interestingly enough, the PPD of the clipped IS posterior is consistent with the observations and does not exhibit 'blanket-like' features, which implies that the blanket features are mostly coming

from the tails of the estimated posterior. A further coverage check needs to be performed to test whether this leads to an under-dispersed posterior estimator. Since this involves several thousands of likelihood computations, it was not performed here. However, parallelizing can immensely reduce this overhead.

In general, IS is known to be challenging in higher dimensions and doesn't perform well on real noisy data. This could be due to the fact that, as discussed earlier, noise introduces significant uncertainty and degeneracies in the parameters. In larger dimensions, even the smallest deviation can have a considerable impact. Gebhard et al. (2025) highlights the sensitivity of certain parameters to inaccurate noise modeling. Since we use a Gaussian noise model, which may not accurately capture the true noise characteristics, this mismatch could be contributing to the observed issues. Further, an increase in the number of samples too can be a potential point of improvement. We leave this for future work.

6.7 Conclusion

In this chapter, we perform water ice cloudy retrievals of the WISE 1738 spectrum. We find a broad posterior predictive distribution motivating an uncertainty study of the model. The main sources of posterior uncertainty, namely, methodological, epistemic, and aleatoric, are identified in the context of the brown dwarf retrieval. It is found that the main reason for cloudy retrieval posterior uncertainty, aka model uncertainty, is due to model-data mismatch, since there is not enough information in the spectrum in this wavelength range to constrain cloud parameters, leading to broad posteriors.

We also uncover that such an over-parameterization of the model is causing degeneracies between radius and clouds, further broadening the posterior. This means that the retrieval characterizes a bigger brown dwarf with more opaque clouds or a smaller brown dwarf with more transparent clouds, including some solutions in between. Hence, we conclude that the reasons for wide cloudy model PPDs are epistemic and aleatoric in nature. Further retrievals on the (Gaussian) noisy and noise-free synthetic spectra (of the MAP cloudy posterior sample), that closely resemble the WISE 1738 observation, show stable constraints on radius and clouds without degeneracies. This makes us suspect non-Gaussian noise or hidden systematics or randomness in the source to be the source of these degeneracies.

Once we identify the source of model uncertainty, we attempt to improve the epistemic uncertainty by running a retrieval on a patchy model. This shrinks the posterior uncertainty and characterizes an atmosphere with 75% cloud coverage. We further validate all the retrievals and perform model comparisons between the various models using a Bayes factor test. We find that statistically, the test prefers the patchy model; however, the smaller PPD of the cloud-free model suggests that it has better predictive power. Therefore, a cloud-free atmosphere is most preferred to describe the spectrum of WISE 1738.

To correct for any posterior bias, we use importance sampling to reweigh our posterior with weights proportional to the Gaussian likelihood. This reveals that the blanket-like features arise from the posterior's tails; however, a comprehensive coverage test is still required to determine whether the NPE posterior is genuinely over-dispersed.

Chapter 7

Characterizing other brown dwarfs from the JWST/MIRI GTO program

After an in-depth characterization of WISE 1738, we perform retrievals on four other brown dwarfs with similar spectral types using spectra from the GTO program. Specifically, we perform both cloudy and cloud-free retrievals on the combined HST/WFC3+JWST/MIRI data of WISE 1828 to retrieve its chemical and physical properties. We focus on the estimation of the $^{14}N/^{15}N$ ratio, which provides information about its reliability as a tracer of formation pathway. We also compare these results with those obtained using `MultiNest` and other inference algorithms from Barrado et al. (2023) and mostly find consistency. Additionally, we leverage the property of amortization and perform a preliminary short systematic study on the MIRI spectra of all five brown dwarfs observed within the GTO program namely, ROSS 458c, WISE 458 and WISE 0855, and the previously studied WISE 1738 and WISE 1828, to set a precedent for future population studies using SBI. From these retrievals, we look for any chemical or physical trends that can shed light on their formation conditions.

7.1 Studying the atmosphere of WISE 1828

In the same spectral type as WISE 1738, WISE 1828 is a cold Y brown dwarf with a similar effective temperature of approximately 300-400 K as initially characterized by Cushing et al. (2011). It is located at a distance of approximately 9.9 pc and has a mass estimate of 5-10 M_{Jup} interpolated from the evolutionary models (Beichman et al., 2014; Kirkpatrick et al., 2019a). These characteristics make it a close analogue of gas giant planets. However, unlike planets, isolated brown dwarfs are thought to form through gravitational collapse, and the $^{14}N/^{15}N$ ratio has been proposed as a potential tracer of this formation history (Barrado et al., 2023), providing a means to distinguish between them. A ratio similar to solar values suggests a

stellar-like formation process, while elevated values may indicate a planet-like formation, and subsequent ejection due to gravitational perturbation. This hypothesis can be tested using Y brown dwarf spectra such as WISE 1828, as these objects are both luminous and cool enough for their mid-infrared spectra to be sensitive to the $^{14}NH_3$ and $^{15}NH_3$ isopotologues.

Furthermore, various analyses on its photometry and spectrum revealed WISE 1828 to be too bright and too red for its spectral type (Kirkpatrick et al., 2012; Beichman et al., 2013; Leggett et al., 2013), leading to unsatisfactory model fits simultaneously between wavelength ranges 1-2 and 3-5 μm (Kirkpatrick et al., 2019a). This has led to the claim by several studies (Leggett et al., 2013; Leggett et al., 2017) that WISE 1828 might not be a single source but an unresolved binary system of roughly equal mass objects with an effective temperature of approximately 275-300 K. Recent spectral fits using the new Sonora Bobcat models (Marley et al., 2021) on the spectral data from HST between the wavelength ranges of 0.7-1.7 μm support this claim. However, several studies on the newly obtained data from the Near InfraRed Spectrograph (NIRSpec), Near InfraRed Camera (NIRCam) and MIRI instruments onboard JWST, spanning wavelength regions totaling between 0.6-18 μm, find no evidence for this multiplicity (De Furio et al., 2023; Ardévol Martínez, 2024).

In this section, we leverage the computational efficiency of NPE, which decouples forward model simulation from training. We use models initially simulated for the analysis of WISE 1738 to perform combined spectral retrievals of data from WFC3/HST and MIRI/JWST between wavelength ranges 0.9-1.7 μm and 4.9-18 μm respectively. We characterize its atmospheric composition, confirming the presence of the ammonia isotopologue $^{15}NH_3$, and further derive the $^{14}N/^{15}N$ ratio, finding it to overlap with the values reported in Barrado et al. (2023) and Ardévol Martínez (2024). We also compare the NPE retrievals with the Nested sampling retrieval published in Barrado et al. (2023). This work was carried out as a contribution to that study. Additionally, we investigate the theoretical prediction of water cloud formation proposed by Morley et al. (2014) by performing retrievals over a cloudy model in an amortized framework, in an effort to fix the mismatch between the shorter and longer regions simultaneously. However, we find no evidence for water clouds under the used model, hinting at some missing aspect in modeling. We then compare these results with the analysis conducted in Ardévol Martínez (2024), which examined similar properties using the MRS/MIRI spectrum additionally with the data from PRISM/NIRSpec of the JWST. For all the retrievals we assume WISE 1828 to be a single object, but allow for radius prior to be wide enough to account for an equal-brightness binary scenario.

7.1.1 Setup

Radiative transfer simulator

The *Cloud-free model* atmospheric forward model, f_{cf}, used to generate the emission spectra in this study, is setup similarly to the previously described simulator of WISE 1738. It uses `petitRADTRANS` with identical parameterizations for the thermal profile, the same line lists for the overlapping species, and assumes constant abundances throughout the atmosphere. The molecular absorbers considered here are H_2O, CO_2, CO, CH_4, NH_3, PH_3, H_2S, along with its isotopologue $^{15}NH_3$. Continuum species such as H_2-H_2 and He-H_2 collision-induced absorption bands are also considered. The planet radius R_p and surface gravity $\log g$ are also considered to be free parameters. This is slightly different from the model in WISE 1738 where planet mass M_p was used instead of surface gravity. The model is parameterized using 22 parameters.

The *Cloudy model* atmospheric forward model f_{cl}, is parameterized with a total of 30 parameters. The pressure-temperature profile and abundances are modeled in the same way as in the cloud-free case. Additionally, molecular absorbers HCN, TiO and VO are considered. The abundances of icy water clouds are included as absorber abundances. These icy water clouds are parameterized using 5 free parameters identical to the cloudy case in WISE 1738.

In either case, the noise model is parameterized as a Gaussian distribution, where the standard deviation of the measurement noise is scaled in the same manner as for WISE 1738. The cloud-free and cloudy model scaling parameters b_w and b_m on instruments WFC3 and MIRI take values between $[-15, -7]$ which imply a scaling on the maximum standard deviation on the measurement noise of WISE 1828 spectrum on those instruments by a factor of 371 and 21 times, respectively.

Prior

The prior distributions in the case of the cloud-free and cloudy models are a 22 and 30 dimensional multivariate uniform distribution $p(\theta)$ with physically motivated ranges for each parameter. The cloud-free and cloudy model priors are listed in Table 7.1.

Training set

The cloud-free training set consists of 1.8 million parameter-spectra pairs, while the cloudy model contains 3.7 million pairs, both combining the near-infrared and mid-infrared wavelengths. The cloud-free model is simulated to the full 0.9–1.7 μm wavelength range in the near-infrared, whereas the cloudy model is only simulated to be between 0.98–1.7 μm (which overlaps with the spectral range of WISE 1738), to leverage amortization. As in the dataset of WISE 1738, atmospheric models are simulated with a wavelength spacing of $\lambda/\Delta\lambda = 1000$ using `petitRADTRANS`,

Table 7.1: Prior distribution for the cloud-free and the cloudy model parameters.

Parameter	Prior	Parameter	Prior
R_P	$\mathcal{U}[0.5, 3)$	H_2O	$\mathcal{U}[-10, 0)$
$\log g$	$\mathcal{U}[2.5, 6)$	CO_2	$\mathcal{U}[-10, 0)$
T_{bottom} [a]	$\mathcal{U}[100, 9000)$	CO	$\mathcal{U}[-10, 0)$
$T_{\text{nodes[i-ix]}}$ [b]	$\mathcal{U}[0.2, 1)$	CH_4	$\mathcal{U}[-10, 0)$
b_w [c]	$\mathcal{U}[-15, -7)$	NH_3	$\mathcal{U}[-10, 0)$
b_m [c]	$\mathcal{U}[-15, -7)$	PH_3	$\mathcal{U}[-10, 0)$
		H_2S	$\mathcal{U}[-10, 0)$
		$^{15}NH_3$	$\mathcal{U}[-10, 0)$

Parameter	Prior	Parameter	Prior
R_P	$\mathcal{U}[0.5, 3)$	H_2O	$\mathcal{U}[-10, 0)$
M_P	$\mathcal{U}[1, 50)$	CO_2	$\mathcal{U}[-10, 0)$
T_{bottom} [a]	$\mathcal{U}[100, 9000)$	CO	$\mathcal{U}[-10, 0)$
$T_{\text{nodes[i-ix]}}$ [b]	$\mathcal{U}[0.2, 1)$	CH_4	$\mathcal{U}[-10, 0)$
X_{H_2Oc}	$\mathcal{U}[-10, 0)$	NH_3	$\mathcal{U}[-10, 0)$
$\log P_{\text{base}}$	$\mathcal{U}[-6, 3)$	PH_3	$\mathcal{U}[-10, 0)$
f_{sed}	$\mathcal{U}[5, 13)$	H_2S	$\mathcal{U}[-10, 0)$
$\log k_{zz}$	$\mathcal{U}[0, 10)$	$^{15}NH_3$	$\mathcal{U}[-10, 0)$
σ_{lnorm}	$\mathcal{U}[1.05, 3)$	HCN	$\mathcal{U}[-10, 0)$
b_w [c]	$\mathcal{U}[-15, -7)$	VO	$\mathcal{U}[-10, 0)$
b_m [c]	$\mathcal{U}[-15, -7)$	TiO	$\mathcal{U}[-10, 0)$

Notes. All the abundances are mass fractions in $\log_{10}$ units.
[a] T_{bottom} is the temperature at the bottom-most node in the pressure grid.
[b] $T_{\text{nodes[i-ix]}}$ are the subsequent fractions of the previous node temperatures.
[c] The b factor for the instruments are additive noise factors, in log value, by which the square of the measured error bars are exaggerated in each bin of the spectrum. This embodies the uncertainty in the estimated error of each instrument or model inaccuracies.

ensuring a uniform resolution across the spectra. For the mid-infrared, both models maintain a spectral resolution of $\lambda/\Delta\lambda = 1000$ to match `petitRADTRANS`. Meanwhile, near-infrared spectra from HST are convolved to the observational resolution and rebinned to a spacing of 320 to align with measured data.

To perform our retrievals, here we try a different loss function called the BNPEloss, to improve the dispersion of our final posterior estimates. This results in a new hyperparameter-tuned architecture whose technical details are provided in Table 7.2.

Table 7.2: Technical details of the posterior estimator and training used for WISE 1828 cloud-free model retrieval.

Flow and embedding		Tuned	
Neural architecture	Details	Hyperparameter	Value
Normalizing flow	NAF	Optimizer	AdamW
Flow transforms	5	Initial learning rate	1×10^{-4}
Transform	MLP	Scheduler	ReduceLROnPlateau
Signal	16	Patience	8
Hidden features (transform)	$5 \times [256]$	LR reduction factor	0.5
Activation (transform)	ELU	Minimum learning rate	1×10^{-8}
Embedding architecture MIR	Residual MLP	Weight decay	1×10^{-4}
Embedding depth MIR	$[512] \times 2 + [256] \times 3 + [128] \times 5$	Batch size	256
Embedding output MIR	$1298 \to 64$	Loss	BNPELoss
Embedding architecture NIR	Residual MLP	Epochs	100
Embedding depth NIR	$[512] \times 3 + [256] \times 5 + [128] \times 7$		
Embedding output NIR	$212 \to 16$		

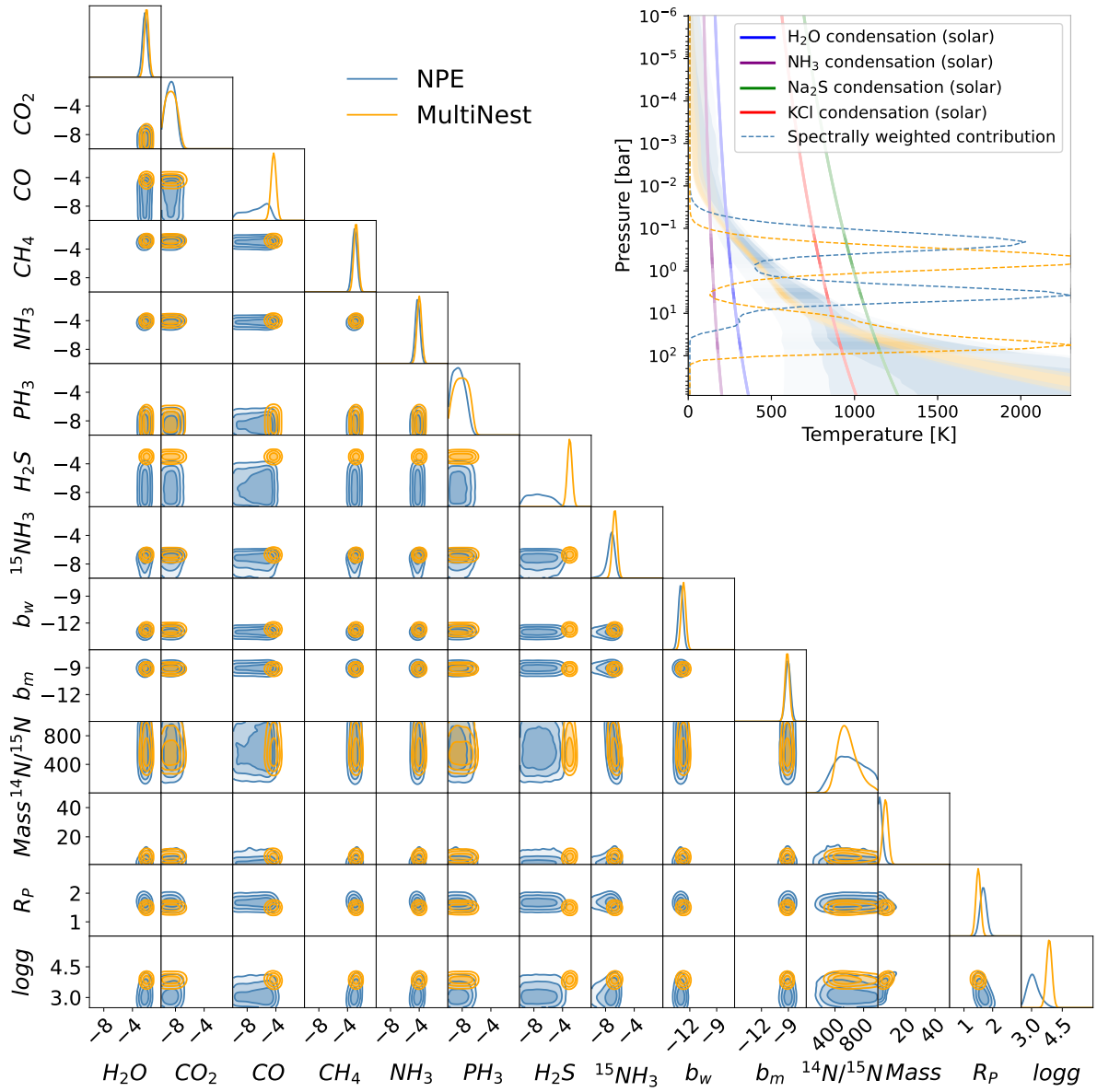


Figure 7.1: *Left*. Cloud-free retrieval using NPE and `MultiNest` on the WISE 1828 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1828 observations x_{obs}, WFC3+MIRI spectra. *Right*. The top right figure illustrates the posterior distribution of the *P-T* profiles, that has the emission contribution function overlaid on top of it in shades of white highlighting visibility. The equilibrium state water ice, ammonia, KCl and Na_2S condensation curves are plotted along the profile in blue and purple, red and green respectively for solar metallicity [M/H] and C/O ratio.

7.1.2 Results

The results of the combined NPE cloud-free and cloudy retrievals over the WFC3+MIRI spectral observations x_{obs} of WISE 1828 are presented as 1D and 2D marginal posterior distributions in Figure 7.1 and 7.2. These posterior samples are generated by performing forward passes through the trained normalizing flow. The corner plots include the derived marginal posterior distributions of the $^{14}NH_3/^{15}NH_3$ ratio and surface gravity and mass, along with the retrieved parameters.

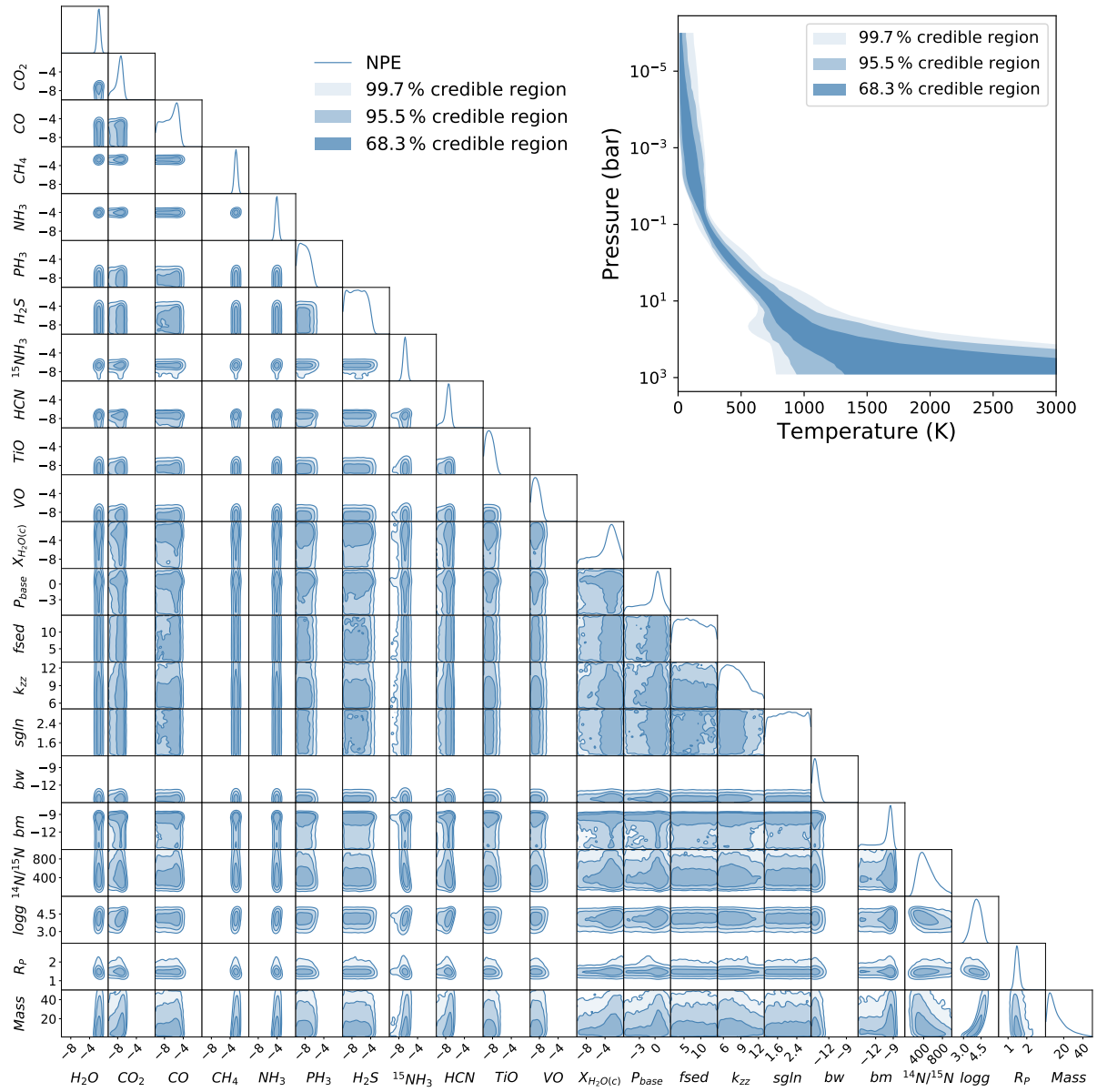


Figure 7.2: *Left*. Cloudy retrieval using neural posterior estimation on the WISE 1828 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the WISE 1828 observations x_{obs}, WFC3+MIRI spectra. The top right figure illustrates the posterior distribution of the *P*-*T* profiles.

The corner plots reveal clear constraints on H_2O, CH_4, NH_3, and $^{15}NH_3$, which in turn constrain the $^{14}NH_3/^{15}NH_3$ ratio. For the remaining molecular species, only upper limits are obtained. The results of the cloud-free NPE retrieval are summarized in Table 7.3, along with results from several other retrievals published in Barrado et al. (2023).

Table 7.3: Physical constraints on WISE J1828

Quantity	NPE	pRT-free	pRT-reg	ARCiS-free	ARCiS-reg	Brewster
$T_{\rm eff}$ (K)	330^{+15}_{-15}	364^{+3}_{-3}	379^{+4}_{-4}	354^{+13}_{-7}	388^{+3}_{-4}	397^{+16}_{-14}
$\log(g)$	$3.02^{+0.31}_{-0.24}$	$3.83^{+0.08}_{-0.08}$	$4.37^{+0.16}_{-0.17}$	$3.38^{+0.1}_{-0.08}$	$4.72^{+0.12}_{-0.14}$	$4.74^{+0.19}_{-0.21}$
R ($R_{\rm Jup}$)	$1.66^{+0.08}_{-0.08}$	$1.45^{+0.02}_{-0.02}$	$1.35^{+0.03}_{-0.03}$	$1.65^{+0.04}_{-0.04}$	$1.35^{+0.03}_{-0.03}$	$1.19^{+0.06}_{-0.06}$
M ($M_{\rm Jup}$)	$1.11^{+1}_{-0.43}$	$5.76^{+1.04}_{-0.02}$	$17.27^{+6.86}_{-2.51}$	$2.69^{+0.58}_{-0.40}$	$38.57^{+10.92}_{-9.66}$	$31.50^{+14.30}_{-11.60}$
$R_{\rm bin}$ ($R_{\rm Jup}$)	$1.17^{+0.06}_{-0.06}$	$1.02^{+0.02}_{-0.02}$	$0.95^{+0.02}_{-0.02}$	$1.17^{+0.03}_{-0.03}$	$0.95^{+0.02}_{-0.02}$	$0.84^{+0.06}_{-0.06}$
$M_{\rm bin}$ ($M_{\rm Jup}$)	$0.56^{+0.5}_{-0.21}$	$2.88^{+0.52}_{-0.45}$	$8.63^{+3.43}_{-2.51}$	$1.34^{+0.29}_{-0.2}$	$19.29^{+5.46}_{-4.83}$	$15.75^{+7.15}_{-5.8}$
[M/H]	$-0.33^{+0.12}_{-0.11}$	$-0.07^{+0.02}_{-0.02}$	$0.06^{+0.04}_{-0.04}$	$-0.34^{+0.03}_{-0.03}$	$0.10^{+0.04}_{-0.02}$	$-0.08^{+0.10}_{-0.09}$
C/O	$0.20^{+0.06}_{-0.04}$	$0.21^{+0.01}_{-0.01}$	$0.20^{+0.02}_{-0.02}$	$0.22^{+0.02}_{-0.02}$	$0.17^{+0.02}_{-0.02}$	$0.73^{+0.08}_{-0.08}$
^{14}N/^{15}N	968^{+1895}_{-489}	560^{+165}_{-115}	642^{+365}_{-192}	949^{+322}_{-208}	591^{+432}_{-190}	–
$\log(H_2O)$	$-3.24^{+0.11}_{-0.12}$	$-3.03^{+0.02}_{-0.02}$	$-2.89^{+0.04}_{-0.04}$	$-3.23^{+0.03}_{-0.03}$	$-2.84^{+0.04}_{-0.04}$	$-3.23^{+0.10}_{-0.10}$
$\log(CH_4)$	$-3.93^{+0.13}_{-0.13}$	$-3.72^{+0.03}_{-0.03}$	$-3.61^{+0.07}_{-0.07}$	$-3.88^{+0.03}_{-0.03}$	$-3.60^{+0.06}_{-0.06}$	$-3.36^{+0.10}_{-0.10}$
$\log(NH_3)$	$-5.16^{+0.10}_{-0.10}$	$-4.93^{+0.02}_{-0.02}$	$-4.77^{+0.06}_{-0.06}$	$-5.06^{+0.03}_{-0.03}$	$-4.71^{+0.05}_{-0.05}$	$-4.58^{+0.10}_{-0.10}$
$\log(^{15}NH_3)$	$-8.15^{+0.31}_{-0.5}$	$-7.67^{+0.10}_{-0.11}$	$-7.58^{+0.17}_{-0.21}$	$-8.03^{+0.11}_{-0.13}$	$-7.49^{+0.18}_{-0.25}$	–

Notes. The table above describes the results of various model retrievals for WISE 1828 from Barrado et al. (2023). The pRT-free forward model has an identical setup to that of NPE apart from the relative near/mid-infrared scaling factor, however it differs from the pRT-reg forward model in that it does not impose any regularization on the *P-T* profile. This is similar to ARCiS-free/reg. Except NPE, all the other retrievals were performed using `MultiNest`. All the abundances are mass fractions in $\log_{10}$ units. The R_{bin} model radii represent WISE 1828 assumed as an equal property binary, calculated by multiplying the radius by $1/\sqrt{2}$. R_{bin} was then used to calculate M_{bin} from the inferred gravity.
[(a)] $T_{\rm bottom}$ is the temperature at the bottom-most node in the pressure grid.
[(b)] $T_{\rm nodes[i-ix]}$ are the subsequent fractions of the previous node temperatures.
[(c)] The b factor for the instruments are additive noise factors, in log value, by which the square of the measured error bars are exaggerated in each bin of the spectrum. This embodies the uncertainty in the estimated error of each instrument or model inaccuracies.

The NPE-derived posterior distribution (in blue) is compared to one of the results from that study, specifically the posterior obtained using the `MultiNest` algorithm with pRT-free forward model (in orange) since they have a near identical setup (as NPE also uses pRT). They differ in that the `MultiNest` pRT-free forward model includes an additional free parameter, a spectral scaling factor, that adjusts the WFC3 flux relative to the MIRI spectrum. This scaling factor follows a uniform prior distribution in the range [0.5, 1.5] but remains unconstrained and is therefore omitted from the corner plot in Figure 7.1. We do not include this parameter in our retrieval since there is little physical justification for it, given that the photometry of both instruments is assumed to be reliable. A comparison of the two corner plots reveals that both retrievals yield identical constraints on all molecular mass fractions except for CO and H_2S. These molecules are well-constrained in the `MultiNest` free-PT retrieval but not in NPE, with H_2S values showing no overlap between the two methods. Additionally, while both retrievals yield a ^{14}N/^{15}N ratio that peaks around 560, the NPE posterior is heavily skewed toward

higher values, shifting the median estimate to 968. Another notable difference arises in the constraint on surface gravity ($\log g$). The NPE retrieval suggests a significantly lower $\log g$ compared to the `MultiNest` results. It is not clear where these differences could be coming from.

A look at the results in Table 7.3 suggests that the NPE retrieval characterizes an atmosphere associated with a larger but cooler and less massive object, leading to a more extended atmosphere due to weaker gravitational force. In contrast, `MultiNest` retrieves a smaller but more massive object, resulting in a denser atmosphere. Despite these differences, the pressure-temperature (PT) profiles from both methods remain largely similar. The C/O ratios retrieved by NPE are nearly identical to those from `MultiNest` as abundances in both C-bearing and O-bearing species are correlated with gravity and hence more robust. This is followed by a significantly lower metallicity (Lew et al., 2024). Additionally, NPE suggests a higher abundance of all molecular species compared to `MultiNest`.

These findings indicate that NPE favors a scenario where a more diffuse atmosphere is enriched in molecular species but has an overall lower metallicity. The increased atmospheric scale height in the NPE retrieval may suggest weaker gravitational retention of heavier elements. On the other hand, `MultiNest`'s characterization of a smaller, more massive object implies a stronger gravitational pull, leading to a more compact atmosphere which also happens to be metal-rich.

The posterior distributions of the *P-T* profiles are also plotted in the inset of this corner plot, with the emission contribution function overlaid on top in dashed lines, highlighting the regions of the atmosphere that are probed by the observations. The similarity in *P-T* profiles indicates that both retrievals compensate in the inferred composition, surface gravity, and atmospheric extent to give a similar thermal structure. The *P-T* plot also includes the H_2O ice, NH_3, Na_2S, and KCl condensation curves (Lodders and Fegley Jr, 2006) plotted in blue, purple, green, and red colors, respectively. The H_2O and KCl curves intersect with the *P-T* profile at the upper and edge of the lower limits of the photosphere respectively, suggesting that there could be thin water ice or chloride clouds in the atmosphere of WISE 1828. The consistency plot of this retrieval is plotted in Figure 7.4. Additionally, the NPE estimate of the posterior distribution of the cloudy model is plotted in Figures 7.2 and 7.4. Water clouds parameters are not constrained.

The scaling factors b_w and b_m are constrained to -13 and -9 in the cloud-free retrieval, while in the cloudy case, they are constrained (in different units) to -14.4 and -9.8, respectively. This corresponds to scaling factors of 1 and 48 for the cloud-free model, and 4 and 1 for the cloudy model, relative to the maximum error bar in the measurement noise of WISE 1828 on these instruments.

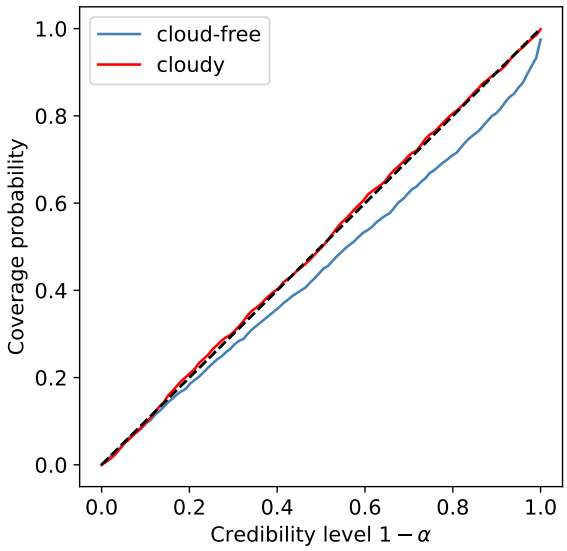


Figure 7.3: Coverage plots for WISE 1828.

7.1.3 Validation

The above results are validated using the coverage test and consistency plots.

Coverage

Coverage plot for retrievals over WFC3+MIRI spectra of WISE 1828 is plotted in Figure 7.3. The plot suggests that the estimators are well calibrated. However, cloud-free model is slightly over-confident, although not significantly.

Consistency plots

Consistency plots for the cloud-free and cloudy NPE posteriors of WISE 1828 are plotted in Figure 7.4 along with the `MultiNest` retrieval. The NPE cloudy PPD is extended to wavelengths that was not retrieved over (illustrated in orange).

The NPE PPD implies that neither of the models explains the data well enough even though they are both consistent with the observations (since the PPD is centered around it). This model inadequacy is suggested from the very broad 2 and 3σ of the PPD, similar to the cloudy and patchy model retrievals of WISE 1738 discussed in Chapter 6. In contrast, the `MultiNest` PPD suggests consistency and adequacy with respect to the observations.

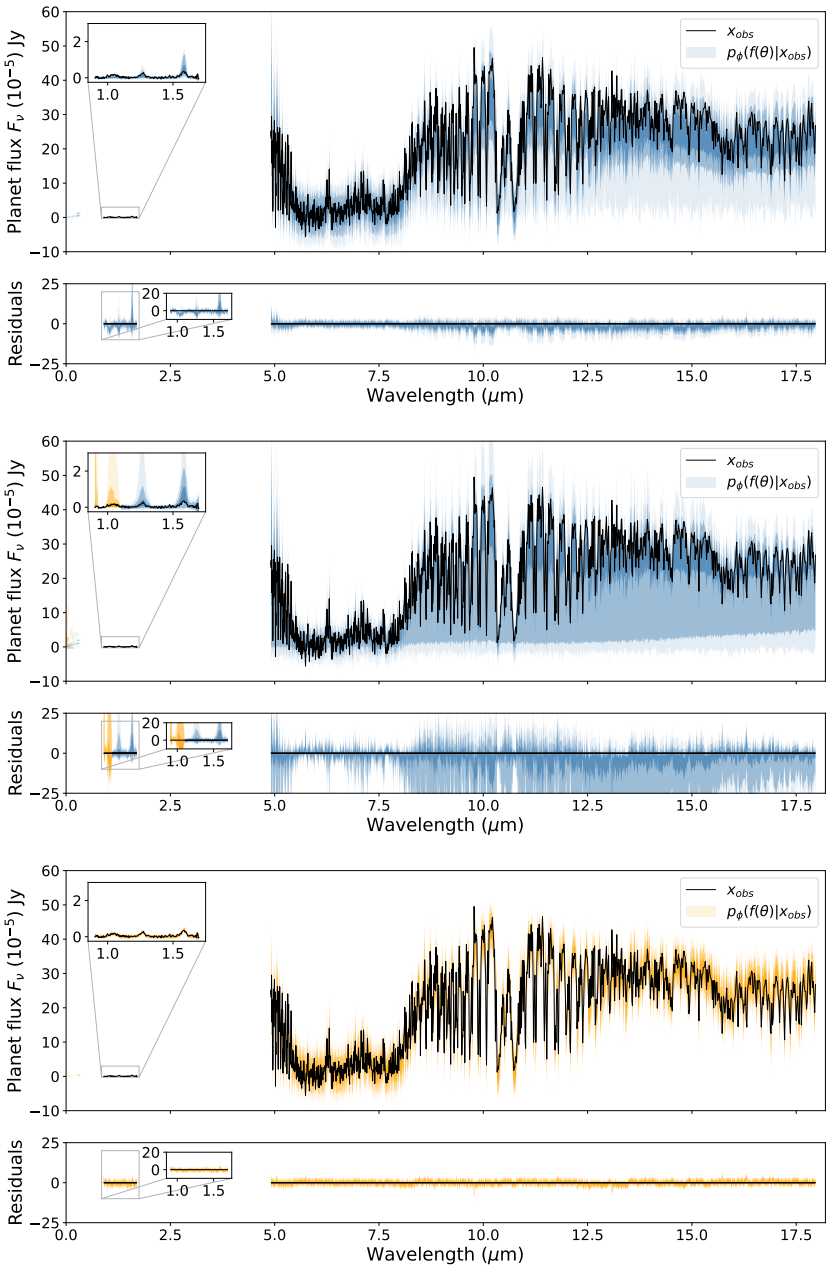


Figure 7.4: WFC3+MIRI consistency plots for the brown dwarf WISE 1828. The posterior predictive distributions $p(f(\theta) + \epsilon | x_{1828})$ obtained from different NPE and *MultiNest* retrievals. The first and second plots pertain to the NPE posterior over cloud-free and cloudy models respectively. The third plot pertains to the *MultiNest* posterior on a near identical cloud-free model. These are overlaid on the WFC3+MIRI observation (black line). The second plot extends to lower near-infrared wavelengths not used in the retrieval (hues of orange).

7.1.4 Discussion

Formation

Due to the heavy right skewing of the derived $^{14}N/^{15}N$ ratio, the constrained median of its derived marginal posterior distribution is 968^{+1895}_{-489}, although the distribution peaks at around 500. This value is consistent with the stellar-like formation of brown dwarfs as expected at the peak (>400). However, it also overlaps with the ISM consistent value (274±18, Ritchey et al. 2015) found by Ardévol Martínez (2024) using the PRISM/NIRSpec data. This can be explained by the huge uncertainty in the NPE's *P-T* profile at the denser regions of the atmosphere which overlaps with both the profiles from Barrado et al. (2023) and Ardévol Martínez (2024). As a result, a wide range of abundance values is found for the trace species $^{15}NH_3$, which only becomes optically thick at higher pressures in order to produce the observed absorption line. This provides an ambiguous result for tracing the formation history of WISE 1828.

Consistency with evolutionary models

The evolutionary models presented by Saumon and Marley (2008) and Marley et al. (2021) indicate that a brown dwarf with an effective temperature between 315-345 K and a surface gravity ($\log g$) between 2.78 and 3.33 cm s^{-2} would have a mass in the range of 0.52–1.04 M_J, a radius between 1.16 and 1.26 R_J, and a bolometric luminosity ($\log L/L_\odot$) between -6 and -6.5. The retrieved mass of $0.56^{+0.5}_{-0.21}\,M_J$, radius of $1.17^{+0.06}_{-0.06}\,R_J$, and bolometric luminosity of $-6.51^{+0.06}_{-0.06}$ are in good agreement with these theoretical predictions, if we assume WISE 1828 to be equal-brightness binary such that the mass and radius are M_{bin} and R_{bin} values from the Table 7.3. Additionally, using the evolutionary model from Marley et al. (2021), we estimate the age to be between 0.04 and 0.1 Gyrs. If not for the binary assumption, the radius and mass values estimated (R and M from the Table 7.3) are much higher compared to the evolutionary models. Consistency with an evolutionary model in general has been challenging for this source (Lew et al., 2024; Barrado et al., 2023; Ardévol Martínez, 2024).

Presence of CO, H_2S and the varying surface gravity

Lew et al. (2024) present bounded constraints on CO and H_2S abundance in WISE 1828, marking the latter, one of the first such constraints in a brown dwarf atmosphere using retrievals on the near-infrared spectrum, alongside Tannock et al. (2021) and Hood et al. (2023). CO is expected to exhibit a strong feature between 3–5 μm, which is missing in our data hence explaining why it is not constrained. However, NS does constrain it. Lew et al. (2024) further suggests that there are degeneracies between surface gravity and atmospheric abundances, and reports a positive correlation with gravity. We see this correlation in the free NPE retrievals. Additionally, the retrieved value of gravity fluctuates significantly depending on the model used within the NPE retrievals, as well as in various other retrievals utilizing multiple datasets

and models from the literature, generally spanning between 3.5 and 5 cm s^{-2} (see Fig. 4.6 of Ardévol Martínez 2024). Such variations in $\log g$ and its impact on the bulk chemistry can either be due to missing physics (as suggested by Ardévol Martínez (2024)) or due to systematics in the noise. This brings into question the validity of the retrieved abundances.

Missing physics

Missing physics in the model might suggest that some key aspect may be missing in the model, with other parameters compensating for these gaps. The model inadequacy is also suggested by the very broad posterior predictive distributions over the combined retrievals of cloud-free and cloudy estimated models, shown in Figure 7.4. The 3σ of the models are huge and the posterior predictive distribution shows "blanket" features similar to the cloudy posteriors of WISE 1738 discussed in Section 6.5.1. As elaborated there, this implies that the predictive power of the retrieved posteriors is poor, potentially hinting at epistemic uncertainty due to model-data mismatch or a broad likelihood. However, these features are not seen in the MIRI only NPE retrievals which produce narrow posteriors. In case it is model-data mismatch, the missing factor could be the presence of chloride (KCl) clouds, as their condensation curve intersects with the lower region of the *P-T* profile (Morley et al., 2012), but, Leggett et al. (2015) does not detect any. A combined retrieval incorporating spectra from longer wavelengths has not been performed and could provide valuable insights.

7.1.5 Inconsistency due to bias in measurements

Another possible reason for the variation in constraints on $\log g$, as well as on H_2S and CO abundances across retrievals, could be biased measurements. These biases may arise from incomplete correction of known systematics, or unmodeled instrumental systematics, or the combination of data from two observatories with differing systematic behaviors. Additional contributing factors might include deviations from a Gaussian noise model or stochastic variations inherent in the data.

Clouds

Water clouds are not well constrained in the narrower wavelength range of the combined retrieval (see Figure 7.2). Since we do not use the entire data for the retrieval, to check for bias, we plot the extended lower near-infrared region posterior predictive distribution in Figure 7.4. The posterior seems consistent here and this results in similarly broad posterior predictive distribution as the rest of the near-infrared region. Additionally, self-consistent cloudy retrievals (Ardévol Martínez, 2024) performed on the PRISM/NIRSpec data do not find water clouds in the WISE 1828 atmosphere, agreeing with our results.

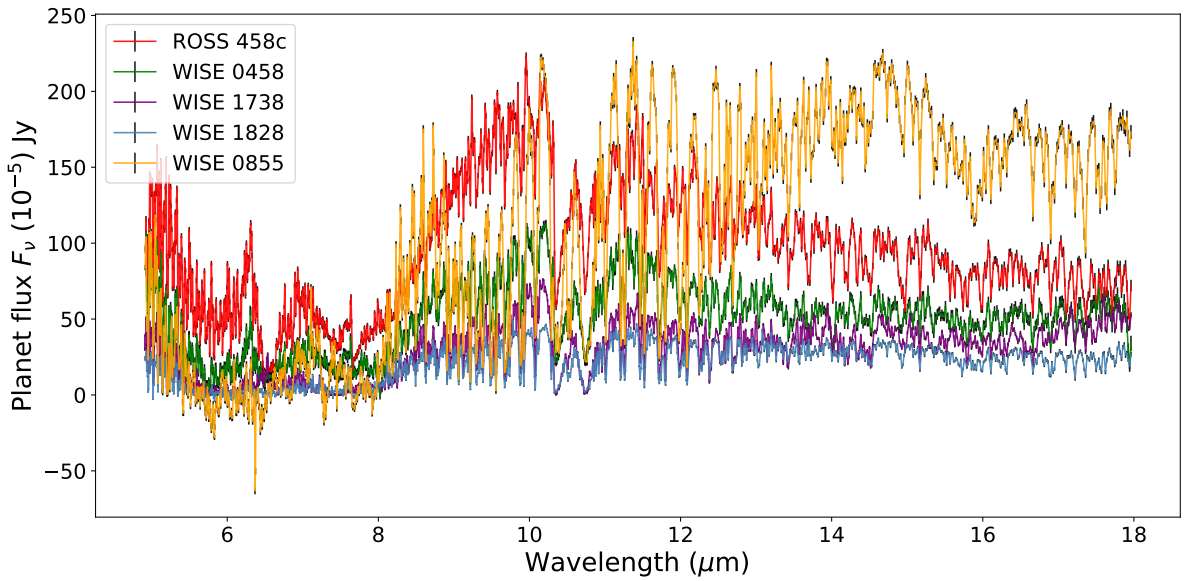


Figure 7.5: MIRI spectral flux of five brown dwarfs within the GTO program.

7.2 Characterizing other brown dwarfs from the GTO

One of the main objectives of studying exoplanet and brown dwarf spectra is that it not only unveils fundamental properties such as mass, radius, composition, and age but also provides insights into their dispersion. This variability helps identify broader trends, offering valuable constraints on theoretical models of formation and evolution. For instance, the C/O ratio has been shown to be influenced by the location of an exoplanet's formation and migration relative to the ice-lines in its proto-star disk (Öberg et al., 2011; Madhusudhan et al., 2014; Chachan et al., 2023). Further, formation models have suggested that atmospheric enrichment influences both metallicity along with the C/O ratio (Schneider and Bitsch, 2021). There are other tracers such as the nitrogen over oxygen ratio along with sodium and magnesium, that inform the formation pathway of exoplanets (Cridland et al., 2020; Notsu et al., 2022; Bosman et al., 2019; Turrini et al., 2021b; Oberg et al., 2021). Such studies motivate finding trends present in brown dwarf atmospheres. Apart from formation, abundances of primary opacity species such as H_2O, CH_4 and NH_3 inform us about the dynamics of the atmosphere. For instance, while the former two species are expected to be consistent with equilibrium chemistry in T/Y spectral type brown dwarfs, in contrast, NH_3 is thermochemically expected to show a sharp decrease in abundance with an increase in effective temperature (Burrows and Sharp, 1999), due to the strong influence from disequilibrum chemistry (Saumon et al., 2006; Marley and Robinson, 2015).

Table 7.4: Retrieved molecular abundances for 5 Y dwarf atmospheres.

Source	Data	H_2O	CH_4	NH_3
ROSS 458C	M	$-3.42^{+0.22}_{-0.15}$	$-3.21^{+0.19}_{-0.19}$	$-6.29^{+0.27}_{-0.22}$
WISE 0458	M	$-3.45^{+0.10}_{-0.10}$	$-3.06^{+0.12}_{-0.12}$	$-5.54^{+0.13}_{-0.13}$
	M[a]	$-2.72^{+0.06}_{-0.05}$	$-3.17^{+0.07}_{-0.07}$	$-4.52^{+0.07}_{-0.06}$
WISE J1738	M	$-2.40^{+0.31}_{-0.23}$	$-2.19^{+0.28}_{-0.22}$	$-3.60^{+0.32}_{-0.24}$
	W+G+M[b]	$-2.86^{+0.11}_{-0.11}$	$-2.72^{+0.14}_{-0.15}$	$-4.2^{+0.12}_{-0.12}$
WISE J1828	M	$-3.06^{+0.05}_{-0.03}$	$-3.58^{+0.04}_{-0.04}$	$-4.90^{+0.04}_{-0.05}$
	W+M	$-3.24^{+0.11}_{-0.12}$	$-3.93^{+0.13}_{-0.13}$	$-5.16^{+0.10}_{-0.10}$
	W+M[c]	$-3.03^{+0.02}_{-0.02}$	$-3.72^{+0.03}_{-0.03}$	$-4.93^{+0.02}_{-0.02}$
	W+M[d]	$-3.03^{+0.18}_{-0.21}$	$-3.65^{+0.21}_{-0.21}$	$-4.79^{+0.15}_{-0.25}$
	P+M[e]	$-2.95^{+0.03}_{-0.03}$	$-3.53^{+0.04}_{-0.04}$	$-4.75^{+0.03}_{-0.04}$
WISE 0855	M	$-3.19^{+0.14}_{-0.14}$	$-3.20^{+0.12}_{-0.10}$	$-4.64^{+0.08}_{-0.08}$
	GN+P+M[f]	$-3.18^{+0.03}_{-0.02}$	$-3.32^{+0.04}_{-0.03}$	$-4.27^{+0.03}_{-0.03}$

Notes. This table enlists the retrieved molecular abundances for water (H_2O), methane (CH_4), and ammonia (NH_3) in five Y dwarfs using different data sets and models.
[a] Matthews et al. (2025) [b] Vasist et al. 2025 (submitted) [c] PT-free model retrieval from Barrado et al. (2023) Barrado et al. (2023) [d] averaged over all the models from Barrado et al. (2023) [e] Ardévol Martínez (2024) [f] Kühnle et al. (2024)

After comprehensive studies on the atmospheres of WISE 1738 and WISE 1828, we conduct a short systematic study on the MIRI spectra of five brown dwarfs (including latter two) within the GTO program "MIRI Spectroscopic Observations of Brown Dwarfs". In the decreasing order of effective temperature they are, ROSS 458c (T8, 700-800K), WISE 0458 (T8.5-T9 565K), WISE 1738 (Y0 400K), WISE 1828 (Y2 300-400K) and WISE 0855 (Y4 285K) at distances 11.5 pc, 14 pc, 7.34 pc, 9.9 pc and 2.28 pc respectively. Their corresponding MIRI spectra are plotted in Figure 7.5.

Since late-T and Y-type brown dwarfs, with effective temperatures between 300–700 K, share similar atmospheric compositions, they can typically be modeled using a cloud-free framework dominated by H_2O, CH_4, and NH_3. Leveraging this similarity, we reuse the trained models of previous WISE 1738 and WISE 1828 retrievals to perform only MIRI retrievals which is presented in Figure 7.6. Such retrievals not only enable a systematic comparison of the different characterizations to identify potential trends in their properties, but also rapidly detect unique features that can be further followed up by deeper independent analysis. For instance, we see that NPE retrieval detects $^{15}NH_3$ in WISE 1828 and HCN in WISE 0458, which were originally detected by Barrado et al. (2023) and Matthews et al. (2025) respectively.

The chemical and physical properties from these retrievals, along with independent studies that incorporate various datasets and forward models are tabulated in Tables 7.4 and 7.5 respectively. The MIRI retrieved values are then plotted in Figures 7.7, 7.8 and 7.9 to identify any emerging trends.

Table 7.5: Retrieved parameters for 5 Y dwarf atmospheres.

Source	Data	C/O	[M/H]	Radius ($R_{\rm Jup}$)	Mass ($M_{\rm Jup}$)	log g cm s^{-2}	$^{14/15}NH_3$	log(L/L$_\odot$)	$T_{\rm eff}$ (K)
ROSS 458C	M	$1.51^{+0.78}_{-0.53}$	$-0.17^{+0.18}_{-0.15}$	$1.76^{+0.06}_{-0.05}$	9.5^{+21}_{-7}	$3.92^{+0.51}_{-0.49}$	-	$-5.56^{+0.57}_{-0.06}$	556^{+213}_{-22}
WISE 0458	M	$2.39^{+0.79}_{-0.57}$	$-0.08^{+0.09}_{-0.10}$	$1.49^{+0.05}_{-0.04}$[g]	$17^{+1.07}_{-0.97}$[g]	$4.29^{+0.13}_{-0.29}$	-	$-5.58^{+0.69}_{-0.08}$	502^{+243}_{-26}
	M[a]	$0.35^{+0.03}_{-0.03}$	$0.13^{+0.06}_{-0.05}$	$0.81^{+0.008}_{-0.007}$[g]	66^{+19}_{-15}[g]	$5.4^{+0.11}_{-0.12}$	-	-	566^{+7}_{-6}
WISE J1738	M	$1.57^{+0.32}_{-0.25}$	$0.83^{+0.28}_{-0.21}$	$1.06^{+0.02}_{-0.02}$	31^{+13}_{-15}	$4.85^{+0.16}_{-0.29}$	-	$-6.59^{+0.52}_{-0.04}$	395^{+137}_{-11}
	W+G+M[b]	$1.35^{+0.39}_{-0.31}$	$0.34^{+0.12}_{-0.11}$	$1.14^{+0.03}_{-0.03}$	13^{+11}_{-7}	$4.43^{+0.26}_{-0.34}$	-	$-6.52^{+0.05}_{-0.04}$	402^{+12}_{-9}
WISE J1828	M	$0.30^{+0.03}_{-0.03}$	$-0.12^{+0.03}_{-0.04}$	$1.07^{+0.03}_{-0.03}$[g]	$16^{+0.95}_{-0.83}$[g]	$4.55^{+0.16}_{-0.22}$	427^{+210}_{-118}	$-6.46^{+0.03}_{-0.03}$	355^{+8}_{-7}
	W+M	$0.20^{+0.06}_{-0.04}$	$-0.33^{+0.12}_{-0.11}$	$1.17^{+0.06}_{-0.06}$[g]	$0.56^{+0.5}_{-0.21}$[g]	$3.02^{+0.31}_{-0.24}$	968^{+1895}_{-489}	$-6.51^{+0.06}_{-0.06}$	330^{+15}_{-15}
	W+M[c]	$0.21^{+0.01}_{-0.01}$	$-0.07^{+0.02}_{-0.02}$	$1.02^{+0.02}_{-0.02}$[g]	$2.88^{+0.52}_{-0.45}$[g]	$3.83^{+0.08}_{-0.08}$	560^{+165}_{-115}	-	364^{+3}_{-3}
	W+M[d]	$0.21^{+0.45}_{-0.03}$	$-0.05^{+0.15}_{-0.27}$	$0.97^{+0.18}_{-0.09}$[g]	$7.91^{+11.33}_{-6.34}$[g]	$4.34^{+0.42}_{-0.88}$	673^{+393}_{-212}	-	378^{+13}_{-18}
	P+M[e]	$0.26^{+0.01}_{-0.01}$	$-0.06^{+0.03}_{-0.03}$	$0.94^{+0.02}_{-0.02}$[g]	-	$5.14^{+0.08}_{-0.08}$	291^{+108}_{-72}	-	404^{+2}_{-2}
WISE 0855	M	$1.00^{+0.38}_{-0.27}$	$-0.06^{+0.10}_{-0.10}$	$1.45^{+0.09}_{-0.08}$	$49.11^{+0.68}_{-3.41}$	$4.77^{+0.05}_{-0.06}$	-	$-7.16^{+0.04}_{-0.03}$	244^{+8}_{-9}
	GN+P+M[f]	-	-	$0.83^{+0.008}_{-0.008}$	$14.14^{+1.77}_{-1.87}$	$4.70^{+0.06}_{-0.06}$	332^{+63}_{-43}	$-7.29^{+0.01}_{-0.01}$	298^{+2}_{-2}

Notes. This table encapsulates the results from different retrievals performed on 5 shown brown dwarfs using different instrumental data and forward models. The data column indicates the source of observational data used for each retrieval. The data used are spectra from MIRI/JWST (M), WFC3/HST (W), GNIRS/Gemini (G), G395M/NIRSpec (GN) and PRISM/NIRSpec (P). [(a)] Matthews et al. (2025) [(b)] Vasist et al 2025 (submitted) [(c)] PT-free model retrieval Barrado et al. (2023) [(d)] averaged over all the models from Barrado et al. (2023) [(e)] Ardévol Martínez (2024) [(f)] Kühnle et al. (2024) [(g)] Binned Radius on multipying by $1/\sqrt{2}$ assuming an equal property binary. Mass is accordingly calculated from the inferred gravity.

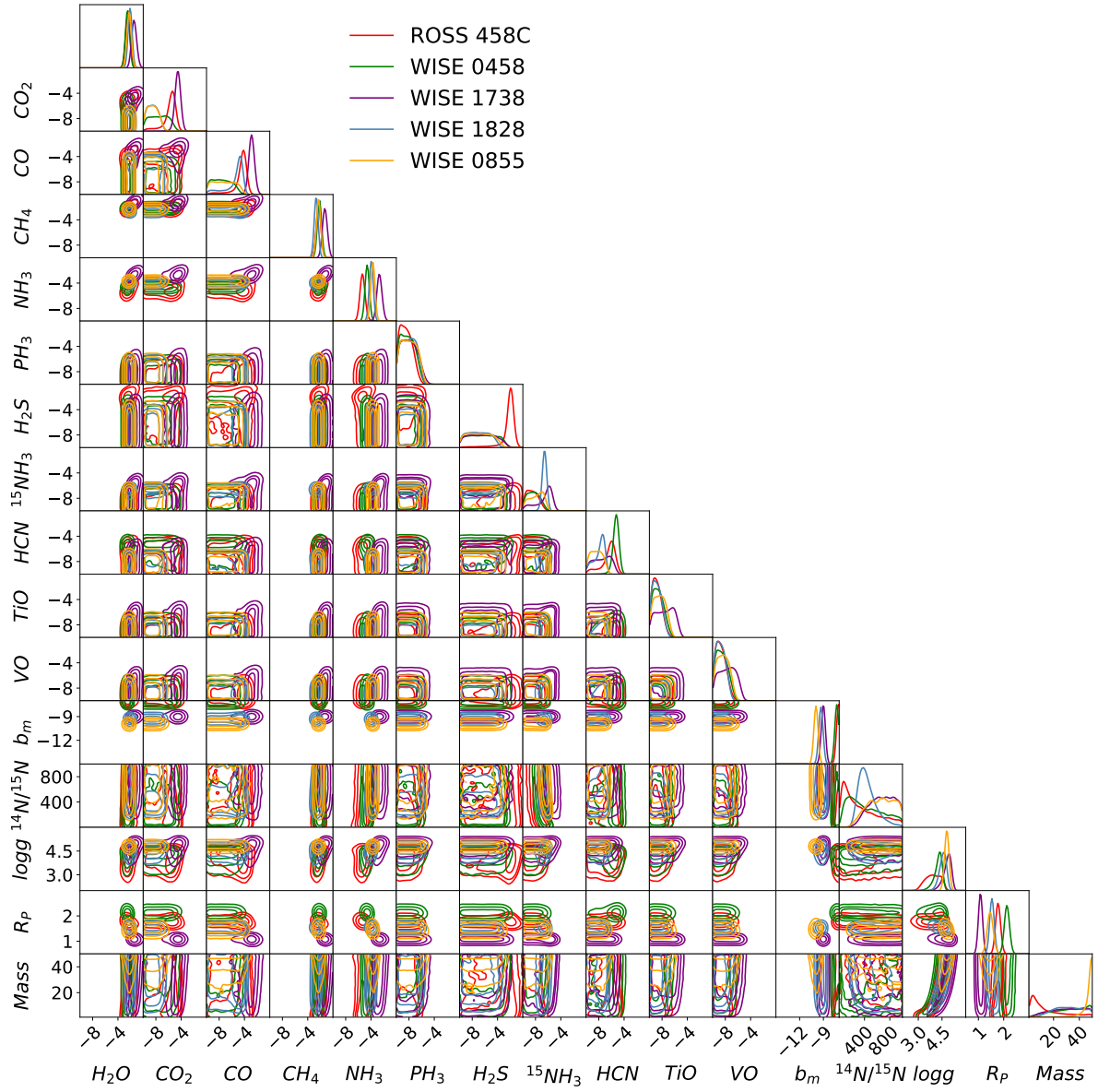


Figure 7.6: A systematic cloud-free retrieval comparison of several late-T and Y dwarf spectra, using the amortization feature of NPE. The sources compared are ROSS 458c (T8, 700-800K in red), WISE 0458 (T8.5-T9 565K, in green), WISE 1738 (Y0 400K, in purple), WISE 1828 (Y2 300-400K, in steelblue) and WISE 0855 (Y4 285K, in orange). The corner plot shows the full 1D and 2D marginal posterior distributions obtained for each of their MIRI spectral observations x_{obs}, leveraging amortization.

Figure 7.7 plots the well-constrained molecular abundances of water, ammonia, and methane across the five brown dwarfs. Other molecules such as CO, CO_2, H_2S and HCN although chemically interesting, aren't constrained in retrievals across all the brown dwarfs considered for comparison. Hence, they are not plotted here. Between temperatures 250-600 K considered, one would expect H_2O and CH_4 to be thermally constant since neither of them is sensitive to

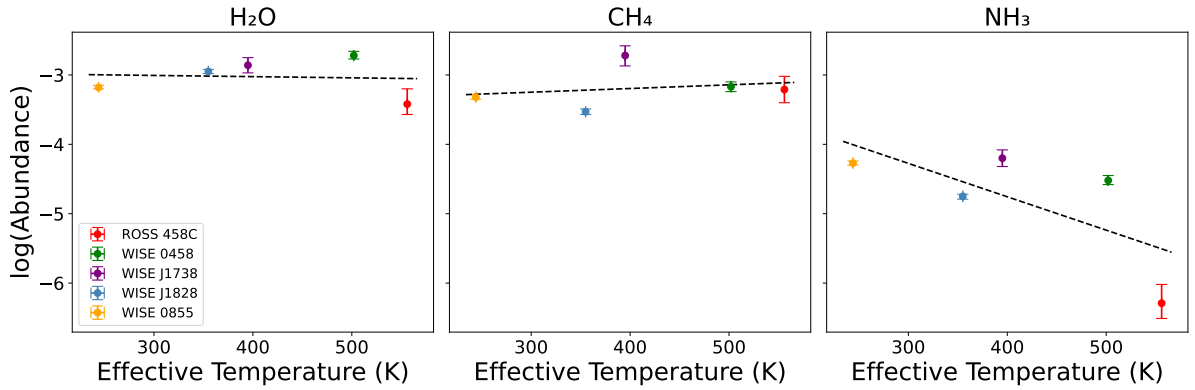


Figure 7.7: Variation of the retrieved molecular abundances over five brown dwarfs across molecules water (H_2O), methane (CH_4), and ammonia (NH_3) in units of volume mixing ratio.

chemical processes such as vertical disequilibrium mixing (Burrows and Sharp, 1999; Sharp and Burrows, 2007), and that is what we see here. However, in contrast, NH_3 does show a decreasing trend in abundance with increasing temperature. We see that the log abundances (in units of volume mixing ratio) decrease from -4 to -6 over increasing effective temperatures across the five brown dwarfs. However, we note that this trend is mostly driven by the data point from ROSS 458c. Interestingly Zalesky et al. (2019) and Line et al. (2017) do not find a trend in NH_3 in their retrieval analysis over WFC3/HST data of around 25 T and Y dwarfs. Although Zalesky et al. (2019) finds a slightly increasing trend, it is largely consistent with chemical equilibrium at a wide range of metallicities and gravities. This is attributed to disequilirbium chemistry in their atmospheres.

Further analysis on their equilibrium chemistry values along the height of each of their atmospheres (not shown here) show that all the retrieved abundances are less than the equilibrium values by a significant order of magnitude as seen before in Saumon et al. (2006), except WISE 1738. The Y0 spectral standard retrieves a value that agrees with equilibrium chemistry at the lower photosphere (see Figure 5.10). However, further evaluation on CO is necessary to comment on the relative efficiency of mixing in these atmospheres, which cannot be observed in the mid-infrared.

Figure 7.8 and Figure 7.9 present key atmospheric properties such as metallicity, C/O ratio, radius, mass, $\log g$, and effective temperature, to explore potential trends across the sample. As expected, $^{15}NH_3$ is only constrained in the cooler atmospheres of WISE 1828 and WISE 0855. We also observe a linear decrease in luminosity with decreasing $T_{\rm eff}$, as anticipated. The C/O ratios are generally aligned, with the exception of WISE 1828, which shows a notably lower value.

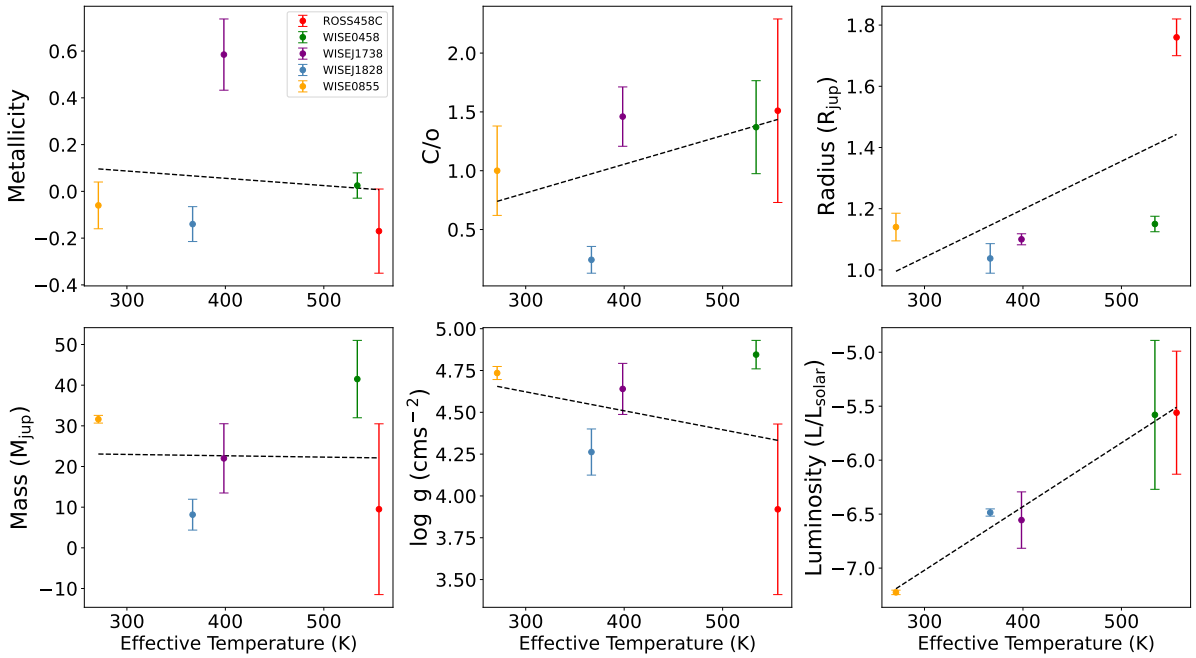


Figure 7.8: Bulk properties plotted against the effective temperature across five brown dwarfs.

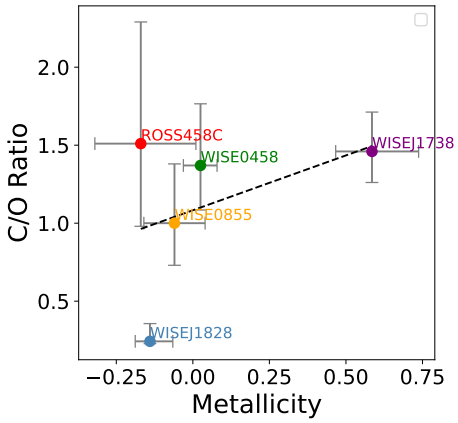


Figure 7.9: C/O across metallicity

Additionally, we perform a cloudy retrieval on these spectra, and find a correlation between degeneracies in radius and cloud parameters. We see degeneracies in the late-T dwarfs ROSS 458c and WISE 0458 along with WISE 1738 (see Figure 7.10), which interestingly coincides with higher value constraints on their respective b factors (see Figures 7.6 and 7.10). The b factors of these objects, constrained in units of the measurement noise of WISE 1738, are significantly higher than those of the other Y dwarfs, implying that they are noisier than their counterparts, i.e if we assume no significant model-data mismatch. This is consistent with highlighting the non-Gaussian nature of the real observational noise, which is seen through corresponding degeneracies. Additionally, there is an epistemic factor to the noise modeling, which could be

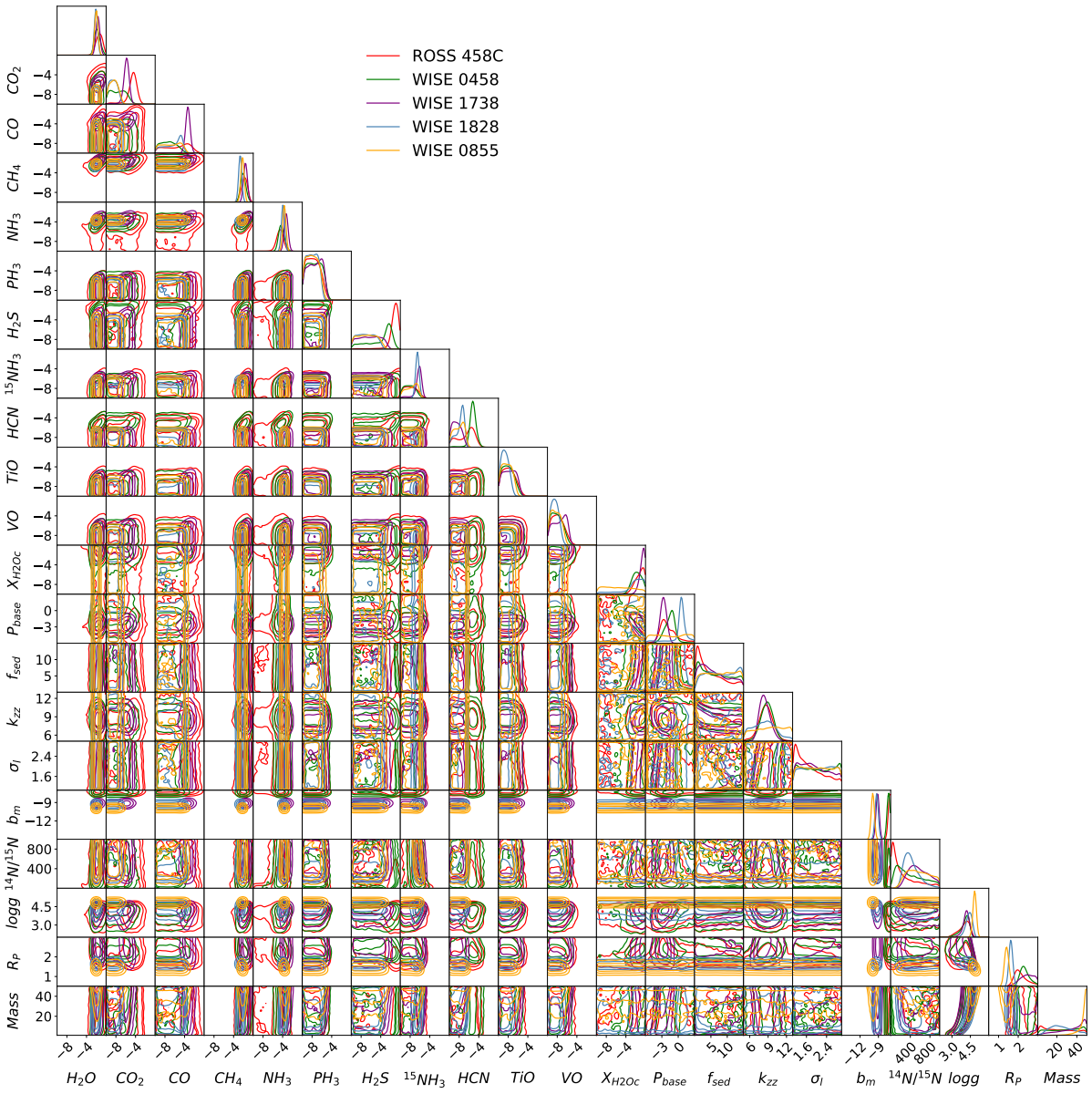


Figure 7.10: Cloudy retrievals using neural posterior estimation on five brown dwarfs from the GTO. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the MIRI spectra x_{obs} of each brown dwarf.

further refined by considering the noise correlation across wavelength bins, since the scaling b factors are inadequate for accounting systematic offsets in the data (Nasedkin et al., 2023). We also find a positive trend of the upper limits on HCN with temperature, which is expected. No other strong trends are apparent from either sets of retrievals.

A similar effort has been made by Beiler et al. (2024a), who analyze (but not retrieve) the PRISM/NIRSpec and MIRI/LRS (low-resolution spectrometer) spectral energy distribution (SED) covering 1-12 μm of 23 late-T and Y dwarfs, to observe their chemical (such as prominent molecular absorption bands) and identify and calculate physical properties (such as radius, effective temperature, luminosity). However, this sample set does not include the five brown dwarfs from our study. The study finds a negative correlation of the NH_3 abundance with effective temperature, presence of CO and CO_2 features, and a lack of PH_3 in these atmospheres; and calculate luminosities that are correlated with effective temperature. Retrievals on these sources along with the brown dwarfs from our study, hence expanding the sample size to span a wider temperature range with diverse atmospheres will further help inform any existing patterns.

7.3 Conclusion

In this chapter, we perform combined NPE retrievals over HST and MIRI data of WISE 1828 using both cloud-free and cloudy models. The results yield well-constrained values for H_2O, CH_4, NH_3, and $^{15}NH_3$, while providing upper limits for other molecular species. A comparison with the `MultiNest` free-PT retrieval from (Barrado et al., 2023) reveals discrepancies in surface gravity, CO and H_2S abundances, and $^{14}N/^{15}N$ ratios, with NPE suggesting a more extended, cooler atmosphere with lower metallicity. The $^{14}N/^{15}N$ ratio, which peaks around 500, remains consistent with both stellar-like and planet-like formation scenarios for WISE 1828, making its formation history ambiguous. This uncertainty arises from overlapping P-T profiles between the models in Barrado et al. (2023) and Ardévol Martínez (2024), leading to a broad range of retrieved $^{15}NH_3$ abundances. We observe an interplay between surface gravity, composition, and atmospheric extent, which helps maintain similar thermal structures between the NPE and the `MultiNest` free-P-T model in Barrado et al. (2023). Additionally, the variation in $\log g$ reported across different studies, along with its confirmed correlation with molecular abundances, raises concerns about the reliability of the retrieved abundance values. This suggests that an essential aspect may be missing from the model, with other parameters compensating for these deficiencies. It also suggests that there could be unaddressed systematics. The need for more refined models becomes evident in the broad posterior predictive distributions, which struggle to simultaneously fit both the near-infrared and MIRI spectral regions. However, no direct constraints on water clouds are obtained.

We also leverage amortization to perform rapid retrievals of the five brown dwarfs in the GTO program. This demonstrates the potential of amortization for future rapid characterizations, paving the way for efficient large scale population studies. In this study, we find a decreasing trend in the NH_3 abundance, driven mostly by ROSS 458c potentially due to vertical mixing. However, further analysis shows that that NH_3 deviates from equilibrium chemistry in all brown dwarfs but it is less pronounced in WISE 1738. Additionally, we find the expected trend of decreasing luminosity and upper limits on HCN, with temperature. Furthermore, we find no particular trends in the other chemical or physical properties of the five brown dwarfs, motivating more such systematic bulk studies over a bigger sample size.

Chapter 8

Characterizing DENIS J0255 from its CRIRES+ spectrum

Another important avenue for studying the characteristics of exoplanets and brown dwarfs apart from medium-resolution spectroscopy that is detailed in the previous chapter, is high resolution Doppler spectroscopy. In this chapter, for the first time we perform NPE retrievals on an L9 brown dwarf named DENIS J0255 using its high resolution spectra and obtain consistent results.

8.1 Context

Planets are believed to be formed either via core accretion (CA; (Pollack et al., 1996)) or gravitational instability (GI; (Boss, 1998)). Core accretion is a bottom-up approach where the dust grains in the disk of the protostar clump together due to collision, to form pebbles which further accrete to form the planet core. These chunks of rocks are also called planetesimals. Alternatively, gravitational instability describes a top-down approach where large masses of gas become unstable within the disk around the protostar due to structural irregularities. This instability ultimately leads to fragmentation and inward collapse giving rise to giant planets.

Either of these formation pathways is expected to leave a trace on the associated planetary spectra. Several of these tracers have been proposed to potentially link the planet to its formation pathway. Any measured deviation in these tracers from the interstellar value is hypothesised to inform the host's formation history. The proposed tracers include several chemical abundance ratios such as the carbon-to-oxygen ratio (C/O, Öberg et al., 2011; Madhusudhan, 2012; Mordasini et al., 2016) and nitrogen-to-carbon or nitrogen-to-oxygen ratio (N/C, N/O, Cridland et al., 2016; Turrini et al., 2021a).

Additionally, isotope ratios (Mollière and Snellen, 2019; Zhang et al., 2021) have been suggested to trace formation pathways due to isotope fractionation. If a planet formed via core accretion, the composition of its ice core depends on how far it formed from its host star, since different molecules condense at different temperatures, and therefore different distances from the host

star. The location where each molecule condenses is called its ice-line. Additionally, while the main isotopologues of these molecules are less susceptible to UV-driven photo-dissociation because of self-shielding, the rarer isotopologues are more readily dissociated and become more available in the gas phase to enrich the ice cores. This isotope fractionation can cause the ice cores to be more abundant in the isotopologue-bearing molecules rather than the main isotopes. Assuming that some amount of out-gassing occurs from these ice-cores into the planet's atmosphere, atmospheric isotopologue abundances relative to their main isotope abundances can be used as a tracer to interpret formation pathways.

Specifically, isotopologue ratios such as Deuterium/Hydrogen (D/H), $^{12}CO/^{13}CO$ and $^{14}N/^{15}N$ isotope ratios have been found to trace formation. A recent study on a super Jupiter TYC 8998-760-1 b (Zhang et al., 2021) reported that the $^{12}CO/^{13}CO$ in its atmosphere was half the interstellar standard. While a lower value suggests planet-like formation, a higher value suggests star-like formation. The observed ^{13}CO-enrichment in Zhang et al. (2021) was proposed to be analogous to the D-enrichment of Neptune and Uranus (Mordasini et al., 2016) in our own solar system and hypothesized that the enrichment occurs due to the icy ^{13}CO planetesimals accreting into the planetary atmospheres beyond the CO ice line, something that is not observed in our solar system since all planets are formed within it. This suggests that $^{12}CO/^{13}CO$ ratio could be a reliable diagnostic to ascertain the formation pathway of exoplanets in systems bigger than our own.

However, to evaluate this hypothesis, it is necessary to conduct a population-wise study on brown dwarfs and exoplanets, to analyze their spectra and characterize their atmospheric chemical abundance and isotope ratios, which can help us establish connections to their respective formation pathways.

8.2 CRIRES+ SupJup survey

Towards this effect, the SupJup survey obtained high-resolution spectra of several brown dwarfs and low-mass companions on the CRIRES+ spectrograph. CRIRES+ is recent upgrade on CRIRES, a cross-dispersed near-infrared spectrograph spanning between 0.95-5.3μm in the near-infrared range with a significantly enhanced performance compared to its predecessor. CRIRES+ offers up to ten times greater instantaneous spectral coverage while combining the substantial light-gathering power of the Very Large Telescope (VLT) with high spectral resolution ($R > 80{,}000$ using a 0.2″ slit). The overall throughput has been increased by approximately 15%. Additionally, the upgraded instrument features enhanced detector arrays for higher quality data and improved wavelength calibration capabilities (see Holmberg and Madhusudhan (2022)).

Of the three primary science goals of CRIRES+, two focus on exoplanet research (Follert et al., 2014): (1) the search for terrestrial planets within the habitable zones of low-mass stars, and (2) the atmospheric characterization of exoplanets. In line with these objectives, numerous ongoing and planned surveys are contributing to these efforts. One such survey is the SupJup Survey, which aims to investigate the formation pathways of exoplanets and brown dwarfs by constraining key chemical tracers in their atmospheres. This is facilitated by the detection of a wide range of molecules, including CO, H_2O, CH_4, NH_3, and HCN (Madhusudhan et al., 2016), as well as various chemical abundance ratios in the infrared.

The SupJup survey obtained high-resolution spectra of 19 isolated objects (brown dwarfs from mid-M dwarfs to mid-T dwarfs), 19 low-mass companions (mid-M dwarfs to early T-dwarfs), and 11 hosts (mostly M dwarfs) on the CRIRES+ spectrograph totaling 49 objects. These objects were selected to ensure diverse spectral types as illustrated in Figure 8.1. It also included a wide range of orbital separations in the case of planetary-mass companions. The main objective of this survey is to constrain their $^{12}C/^{13}C$ isotope ratio, in combination with the C/O ratio and metallicity.

Independent analyses such as in de Regt et al. (2024); González Picos et al. (2024) have been conducted on some of these sources. González Picos et al. (2024) retrieved the high-resolution spectra of the three young brown dwarfs, 2MASS J12003792-7845082, TWA 28, and 2MASS J08561384-1342242 and found that their carbon isotope ratio $^{12}CO/^{13}CO$ shows no significant deviation from the local ISM, suggesting a fragmentation-based formation mechanism similar to star formation. Similarly, de Regt et al. (2024) examined the isolated brown dwarf DENIS J0255 that also suggests fragmentation. Both these results align with the proposed hypothesis.

In contrast to these independent studies, one of the goals of my PhD thesis is to establish a foundation for a systematic large scale study of exoplanets and brown dwarfs based on high-resolution spectra using SBI. Here I perform a pilot study to demonstrate the potential of NPE for such a study by characterizing the SupJup survey source DENIS J0255, similar to the study in de Regt et al. 2024. We also compare the results obtained with a similar (but not identical) *MultiNest* retrieval.

8.3 NPE retrieval on DENIS J0255

DENIS J0255 is a field L9 brown dwarf with an effective temperature of around 1400K. It lies at the edge of the L-T transition zone between the hotter L dwarfs and the cooler T dwarfs. This transition is marked by a shifting dominance of CH_4 as the carbon-bearing molecule in the T dwarf spectra, in contrast to the dominating presence of CO in the L-type dwarfs (Cushing et al., 2005). DENIS J0255 is expected to have a high log gravity of $\gtrapprox 5$ and be of a relatively old age of 2-4 Gyr. This object has been previously studied by Cushing et al. (2008);

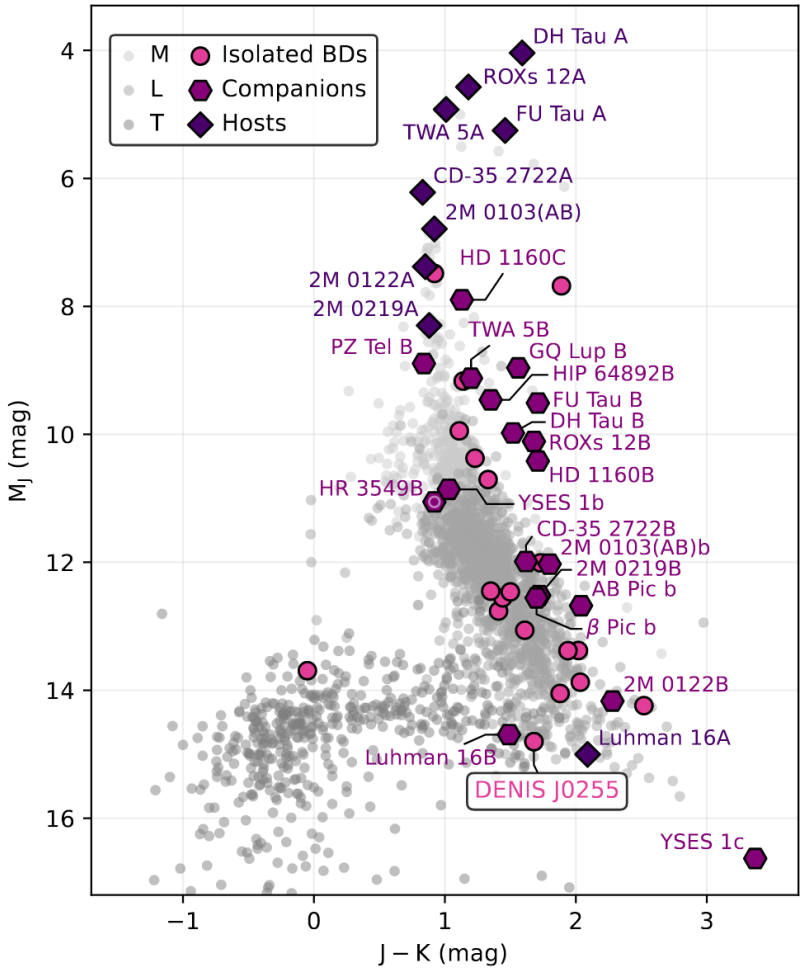


Figure 8.1: The pink, circular markers indicate the observed isolated brown dwarfs. The purple hexagons and dark purple diamonds depict the observed companions and their hosts, respectively. As a reference, the photometry of isolated brown dwarfs was obtained from the UltracoolSheet and is used to display late M, L and T dwarfs with increasingly darker marker shades. Image from de Regt et al. (2024).

Tremblin et al. (2016a); Charnay et al. (2018); Lueber et al. (2022); Cushing et al. (2005); Creech-Eakman et al. (2004); Roellig et al. (2004). These studies found the brown dwarf to have a higher surface gravity $\log g \geq 5$ making it a relatively old brown dwarf (2-4 Gyrs), which is consistent with a high rotational velocity of $v \sin i \sim 40$ km s^{-1}. Lower resolution spectra of the source reveal potential CH_4, NH_3/silicate cloud absorption features, bringing into question its spectral classification as an L-type dwarf (Roellig et al., 2004; Cushing et al., 2005).

DENIS J0255 was observed with CRIRES+ on November 2nd, 2022 as part of the ESO SupJup survey (Program ID: 110.23RW.001, PI: Snellen). The data was reduced using `excalibuhr`*, a Python data reduction pipeline that largely follows the steps outlined by Holmberg and Madhusudhan (2022). The spectrum used here was obtained from de Regt et al. (2024) in seven orders encompassing three detectors each. For this proof of concept retrieval, we only consider one order covering the wavelength range between 2.32 to 2.37 μm (x_{obs}).

8.3.1 Setup

Radiative transfer simulator

Similar to the previous retrievals, the atmospheric model used in this study consists of the deterministic atmospheric forward model implemented with `petitRADTRANS`, together with a noise model accounting for measurement noise. The model includes a freely parameterized pressure-temperature profile and an equilibrium chemistry emission model with a simple quenching implementation at higher altitudes. The model is defined by 16 parameters in total.

The pressure-temperature structure is identical to the one used for the retrieval of HR 8799 e described in section 4.4.2, except that here there are 80 pressure layers defined in the atmospheric model lying between -6 to 2 $\log_{10}$ bars, instead of 100 layers between -6 to 3 $\log_{10}$ bars defined there. The atmosphere is set up to contain the main absorbers within the observed wavelength region such as H_2O and CO with its main isotope $^{12}C^{16}O$ and its rarer isotopologue $^{13}C^{16}O$, where the abundance of the $^{13}C^{16}O$ is set as a fraction of the abundance of $^{12}C^{16}O$ with the free parameter $\log r_{\rm iso}$ such that the abundance of $^{13}C^{16}O = 10^{\log r_{\rm iso}} \times {}^{12}C^{16}O$. The abundances of the absorber species along the height of the atmosphere are interpolated from the chemical equilibrium table calculated with `easyCHEM` (Mollière et al., 2017) as a function of T-P-[Fe/H]-C/O, where [Fe/H] and C/O are free model parameters.

The free parameter $\log P_{\rm quench}$ ($\log P_{\rm q}$ in corner plot) is used to account for disequilibrium chemistry through atmospheric mixing. For pressures less than $P_{\rm quench}$ in the upper atmosphere, the mass fractions of H_2O and COs are held constant at their values at $P = P_{\rm quench}$. This assumption of disequilibrium at low pressures can be justified because the vertical mixing timescale could become shorter than the chemical reaction timescales of the CO-CH_4 and N_2-NH_3 conversion processes (see, Visscher and Moses, 2011; Zahnle and Marley, 2014). Further, collision-induced absorption from H_2-H_2 and H_2-He along with Rayleigh scattering of H_2 and He are taken into account. The surface gravity $\log g$ is an additional free model parameter to calculate the emission flux.

*`https://github.com/yapenzhang/excalibuhr`

The linelists for the absorber species are obtained internally in `petitRADTRANS` from the HITEMP spectral database Rothman et al. (2010) which provides opacity as a function of frequency. For the spectral flux calculations, we configure `petitRADTRANS` to consider every 2nd data point from the high-resolution opacities by setting $lbl\_opacity\_sampling = 2$. This reduces the resolution by a factor of 2 (i.e $R = 10^6/2$), hence decreasing the computational load and accelerating the calculations.

The spectrum generated by `petitRADTRANS` is first scaled by the square of the source radius, R_p, to account for the increase in flux proportional to the source's surface area. Next, the spectrum undergoes rotational broadening, incorporating the projected rotational velocity, $v \sin i$, and a linear limb-darkening coefficient, ϵ_{limb}, using the `fastRotBroad` routine from `PyAstronomy` (Czesla et al., 2019). Rotational broadening occurs because the planet's rotation causes parts of it to move toward or away from the observer, resulting in broadened spectral lines. After rotational broadening, a radial velocity shift, v_{rad}, is applied to the spectrum. This Doppler shift adjusts the spectrum to account for the relative motion of the source along the line of sight. Finally, the spectrum is interpolated to the wavelength grid of the observational data, ensuring alignment with that of the observed spectrum. Parameters R_p, $v \sin i$, ϵ_{limb} and v_{rad} are considered free model parameters. The final output spectrum generated by the simulator, which incorporates both the radiative model along with corrections for high-resolution spectral effects, is expressed as $f(\theta)$, where θ represents all the input 16 model parameters.

We consider a Gaussian noise model such that the generated spectra are randomly perturbed with an additive noise $\epsilon \sim \mathcal{N}(0, s^2)$, where $\epsilon \in \mathbb{R}$ is a vector of random noise instances in each wavelength bin from the observed spectrum of DENIS J0255. The standard deviation of the measurement noise σ is scaled by a multiplicative scaling factor denoted as b which is a free parameter such that, $s = \sigma \times b$. Here we assume that the noise has no covariance for the purpose of simplicity. However, more complex models are possible to be used. The final simulator output is given by $x = f(\theta) + \epsilon$ similar to previous experiments.

Prior

The NPE model prior is defined as a 16-dimensional distribution with physically motivated ranges in each dimension, highlighted in Table 8.1 (Left). The prior distribution is uniform in all dimensions except on the noise parameter b, which is defined as a log-normal distribution with a mean of 1.5 and a standard deviation of 0.5. This distribution keeps the value of b strictly positive where $\log b$ is a normal distribution.

Table 8.1: Prior distribution over the model parameters.

Parameter	Prior	Parameter	Prior
$T_{\rm int}$	$\mathcal{U}(300, 3500)$ K	$\log P_{\rm q}$	$\mathcal{U}(-6, 2)$
T_1	$\mathcal{U}(300, 3500)$	$\log\ r_{\rm iso}$	$\mathcal{U}(-11, -1)$
T_2	$\mathcal{U}(300, 3500)$	$vsini$	$\mathcal{U}(0, 50)$
T_3	$\mathcal{U}(300, 3500)$	$v_{\rm rad}$	$\mathcal{U}(10, 30)$
$\log \delta$	$P_{\rm phot} \in \mathcal{U}(10^3, 10^8)$ [a]	$\epsilon_{\rm limb}$	$\mathcal{U}(0, 1)$
α	$\mathcal{U}(1, 2)$	R_P	$\mathcal{U}(0.8, 2)$
C/O	$\mathcal{U}(0.1, 1.6)$	$\log\ g$	$\mathcal{U}(2.5, 5.5)$
Fe/H	$\mathcal{U}(-1.5, 1.5)$	log b	$\mathcal{N}(1.5, 0.5)$

Parameter	Prior	Parameter	Prior
T_0	$\mathcal{U}(0, 5000)$	$\log P_{\rm q}$	$\mathcal{U}(-6, 2)$
T_1	$\mathcal{U}(0, 1)$	$vsini$	$\mathcal{U}(35, 50)$
T_2	$\mathcal{U}(0, 1)$	rv	$\mathcal{U}(20, 25)$
T_3	$\mathcal{U}(0, 1)$	R_P	$\mathcal{U}(0.8, 2)$
$\Delta \log P_{PT}$	$\mathcal{U}(0, 4)$	$\log\ g$	$\mathcal{U}(4.5, 6)$
C/O	$\mathcal{U}(0, 1)$		
Fe/H	$\mathcal{U}(-1.5, 1.5)$		

Notes. *Left.* Prior on the NPE model. *Right.* Prior on the *MultiNest* model.
[(a)] $P_{\rm phot}$ is the photospheric pressure defined as the pressure where the optical depth $\tau = 1$. The parameter δ dictates it accordingly.

Training set

Our training data set is composed of around 3.5 million parameters-spectrum pairs $(\theta, f(\theta))$.

Technical details on NPE

For the flow, we explored various architectures, including NAF (which resulted in NaN losses), CNF, neural circular spline flow (NCSF), NSF (all of which struggled to constrain the posterior), and MAF. Additionally, we experimented with retrieving 2D data without interpolating the spectrum to the wavelength grid of the observational data, incorporating both spectra and their wavelengths to improve accuracy.

For both 1D and 2D data, we tested several embedding strategies. These included bypassing the Softclip and normalizing the spectra with the mean and variance of the training set, employing a ResMLP by horizontally stacking the flux and wavelength into a 12288-dimensional vector (which proved too memory-intensive to train with MAF), using a 2D ResMLP, and incorporating causal convolution layers to preserve the sequential order of the data. We also explored a CNN with attention layers, leveraging multi-headed attention. The attention mechanism was particularly sensitive to the choice of word embeddings, requiring careful consideration, which we decided was suitable as a future goal. Ultimately, the ResMLP with only the flux emerged as the most effective option for an initial proof of concept, balancing performance and feasibility. Therefore, we adopted it as the baseline architecture.

We tuned each hyper-parameter in the setup by randomly searching over a grid within a certain range, and studied their impact over $\sim$80 runs in parallel. We selected those that led to lower validation loss and/or more stable training. The technical details on the architecture of the flow and other hyper-parameters are provided in Table 8.2. For more details, we refer to the source code of the experiments[†]. For more details on the hyper-parameters, refer to Section 4.3.

[†] `https://github.com/MalAstronomy/Highres`

The flow is trained for a total of 800 epochs during which 1024 and 256 random batches of 2048 pairs $(\theta, f(\theta))$ were taken from the training and validation sets respectively. The architectural hyperparameters were optimized using validation data, with extensive tuning of both the flow and embedding network parameters.

Table 8.2: Technical details of the posterior estimator and training for high-resolution spectroscopic retrievals.

Flow and embedding		Tuned	
Neural architecture	Details	Hyperparameter	Value
Normalizing flow	MAF	Optimizer	AdamW
Flow transforms	3	Initial learning rate	0.001
Transform	MLP	Weight decay	0
Hidden features (transform)	$5 \times [512]$	Scheduler	ReduceLROnPlateau
Activation (transform)	ELU	LR reduction factor	0.5
Embedding architecture	Residual MLP	LR stagnation patience	16
Embedding depth	$[512] \times 2 + [256] \times 3 + [128] \times 5$	Epochs	800
Embedding output	$6144 \to 64$	Batch size	2048

8.3.2 Results

The corner plot of the NPE retrieval is shown in Figure 8.2, and the posterior estimates are tabulated in Table 8.3 (Left). The $\log\, g$ and the radius are constrained at $5.53^{+0.23}_{-0.29}$ and $0.89^{+0.06}_{-0.06}$ respectively, which are consistent with previous analyses (Lueber et al., 2022; Tremblin et al., 2016b; Charnay et al., 2018), including the most recent detailed study by de Regt et al. (2024). Furthermore, the kinematic parameters such as the rotational and radial velocities and their impact on the spectrum via the limb darkening coefficient $\epsilon_{\rm limb}$ are also found to be consistent with that work.

A notable difference with de Regt et al. (2024) is that NPE does not convincingly detect ^{13}CO, as indicated by our constraint on $\log r_{\rm iso}$ at $-5.43^{+2.32}_{-2.25}$. In contrast, de Regt et al. (2024) reports a value of $-2.27^{+0.11}_{-0.13}$. However, they also note that the evidence for ^{13}CO in the study is tentative based on a weak cross-correlation detection. This suggests that NPE's lack of detection of ^{13}CO could be valid. Further, in our proof of concept study, we do not include the presence of NH_3 and CH_4 in the atmosphere since it is not expected to be dominant in L-dwarfs; however, interestingly de Regt et al. (2024) finds robust evidence for their presence, challenging DENIS J0255's classification type as an L dwarf (Roellig et al., 2004; Cushing et al., 2005). Additionally, the lower resolution retrievals also suspect silicate clouds, which is hinted at by the MnS and $MgSiO_3$ condensation curves intersecting the *P-T* profile within the photosphere. However, given that we obtain a good fit to the cloud-free model (see Figure 8.3), and following Occam's razor, we suggest that its effect must be negligible, thus agreeing with de Regt et al. (2024). Furthermore, the noise factor is quantified as $b = 1.88^{+0.75}_{-0.58}$ times the measurement noise.

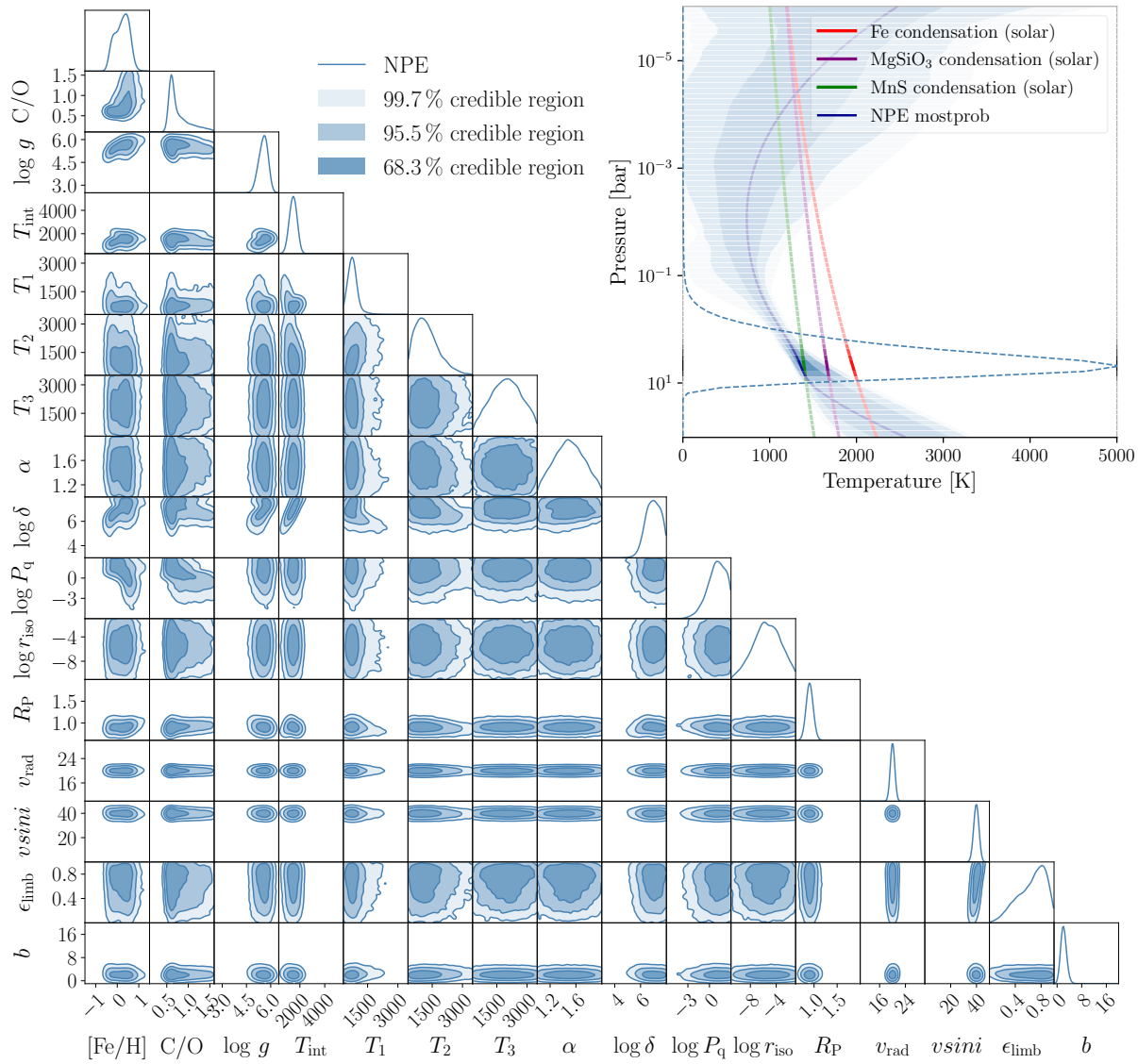


Figure 8.2: *Left.* Retrieval using neural posterior estimation on the DENIS J0255-4700 spectrum. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the high-resolution observations x_{obs}. *Right.* The top right figure illustrates the posterior distribution of the *P-T* profiles, that has the contribution function overlaid on top of it in shades of white highlighting visibility. The equilibrium state $MgSiO_3$, MnS and Fe condensation curves are plotted along the profile in purple, green and red respectively. The profiles of the most probable sample from NPE (dark blue) is also plotted.

Table 8.3: Posterior estimates of the model parameters using NPE and *MultiNest* algorithms.

Parameter	Median	Parameter	Median
$T_{\rm int}$	1425^{+257}_{-273}	$\log P_{\rm q}$	$1.23^{+1.46}_{-1.48}$
T_1	710^{+209}_{-213}	$\log\ r_{\rm iso}$	$-5.43^{+2.32}_{-2.25}$
T_2	1046^{+707}_{-595}	$vsini$	$39.48^{+1.26}_{-1.29}$
T_3	1973^{+956}_{-941}	$v_{\rm rad}$	$19.84^{+0.27}_{-0.25}$
$\log\delta$	$6.97^{+0.60}_{-0.60}$	$\epsilon_{\rm limb}$	$0.69^{+0.17}_{-0.28}$
α	$1.49^{+0.28}_{-0.27}$	R_P	$0.89^{+0.06}_{-0.06}$
C/O	$0.65^{+0.39}_{-0.08}$	$\log\ g$	$5.53^{+0.23}_{-0.29}$
Fe/H	$0.23^{+0.3}_{-0.41}$	log b	$1.88^{+0.75}_{-0.58}$

Parameter	Median	Parameter	Median
T_0	4955^{+31}_{-64}	$\log P_{\rm q}$	$1.09^{+0.01}_{-0.02}$
T_1	896^{+25}_{-20}	$vsini$	$40.39^{+0.12}_{-0.12}$
T_2	698^{+30}_{-31}	$v_{\rm rad}$	$22.64^{+0.06}_{-0.06}$
T_3	563^{+87}_{-116}	R_P	$0.76^{+0.01}_{-0.01}$
$\Delta\log P_{PT}$	$2.36^{+0.06}_{-0.08}$	$\log\ g$	$5.38^{+0.02}_{-0.03}$
C/O	$0.59^{+0.01}_{-0.01}$		
Fe/H	$-0.25^{+0.01}_{-0.01}$		

8.3.3 Validation

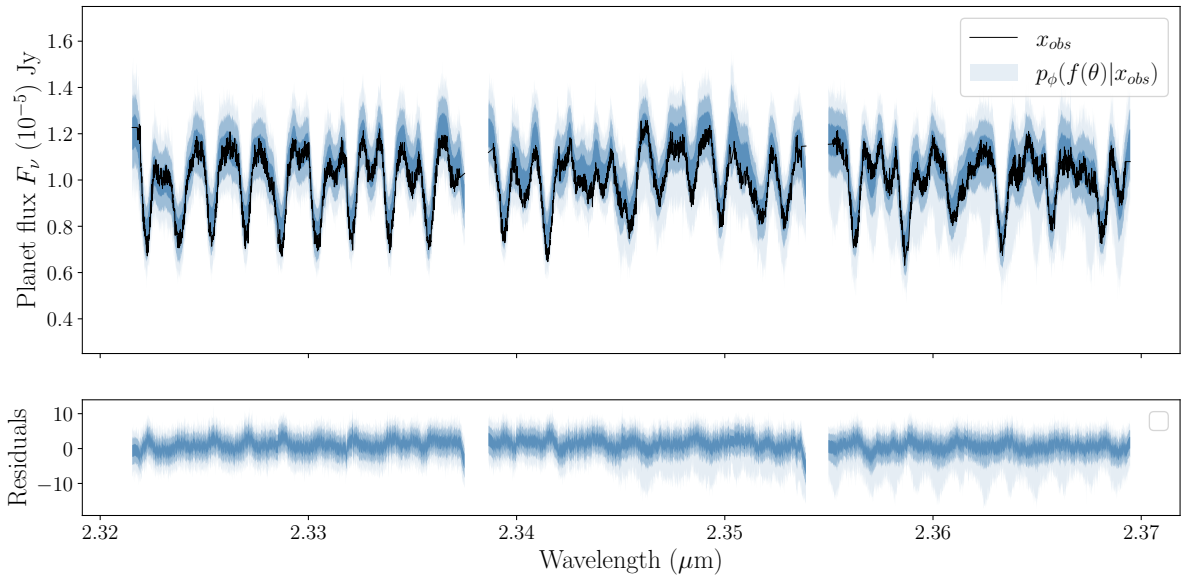


Figure 8.3: Consistency plots for the brown dwarf DENIS J0255. The posterior predictive distributions $p(f(\theta)+\epsilon|x_{\rm obs})$ obtained from the NPE retrieval plotted with the high-resolution observation spectrum $x_{\rm obs}$.

To validate these results, we use a consistency plot (see Figure 8.3) and a coverage plot (see Figure 8.4). The PPD seems consistent with the observations in all wavelengths, and the coverage plot indicates that, on average, the estimator is appropriately dispersed.

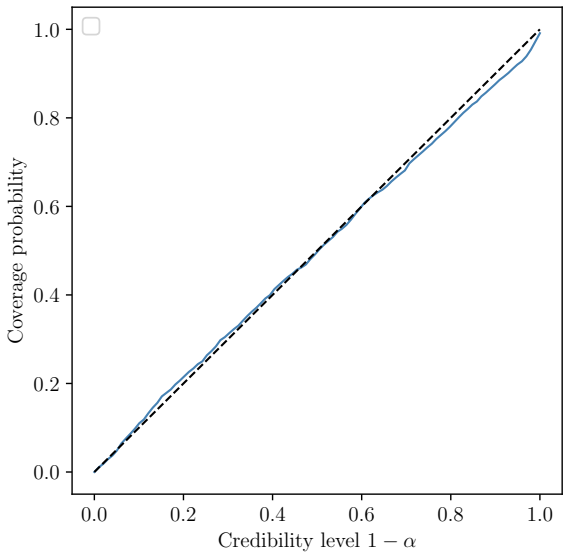


Figure 8.4: Coverage plot for the trained NPE posterior estimator.

8.4 Comparison with *MultiNest*

We also compare these results with a *MultiNest* retrieval. These results are tabulated in Table 8.3. However, we note that although the radiative transfer model for the *MultiNest* retrieval is implemented similarly to the NPE model, there are some differences. These differences include the (i) inclusion of the CH_4 molecule and exclusion of the isotopologue of CO, (ii) using the Pokazatel line list for the H_2O molecule from the ExoMol database (Polyansky et al., 2018) instead of the HITEMP database (Rothman et al., 2010; Barber et al., 2006) used by NPE, (iii) using a P-spline parameterization of the PT profile (Li and Cao, 2022) and (iv) setting the ϵ_{limb} coefficient to a constant value (of 0.65).

The *MultiNest* model prior is defined as a 12-dimensional multivariate uniform distribution with physically motivated ranges as tabulated in Table 8.1 (Right).

The corner plot in Figure 8.5 displays the 1D and 2D marginal posterior distributions of common parameters between NPE and *MultiNest* retrievals. The NPE distributions are obtained by drawing 2991 samples from the joint posterior distribution using the normalizing flow, which equals the number of samples drawn using the *MultiNest* algorithm.

While the marginal posteriors from the two algorithms have huge overlaps, they do differ in some ways. In general, the NPE marginals seem to spread out more than those from *MultiNest*. The coverage plot in Figure. 8.4 indicates that on average the NPE posterior estimator is ideally dispersed, implying that most probably *MultiNest* generates an under-dispersed posterior. One obvious reason for this could be because the model used for *MultiNest* retrieval has 4 less number of parameters, that significantly shrinks the prior volume. Additionally, narrower

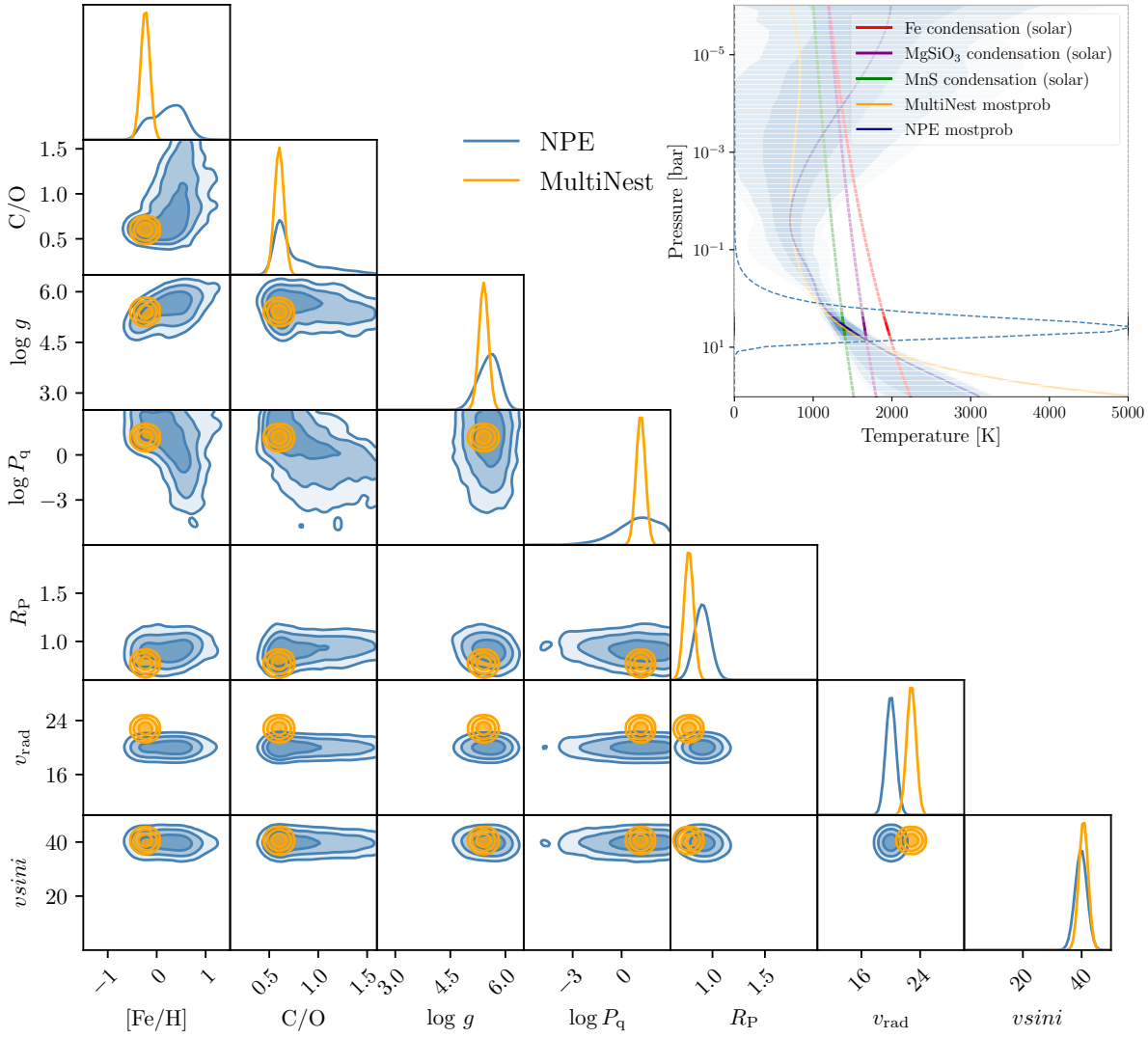


Figure 8.5: *Left.* Retrieval using neural posterior estimation and *MultiNest* on the DENIS J0255-4700 spectra. The corner plot shows the full 1D and 2D marginal posterior distributions obtained for the high-resolution observations x_{obs}. *Right.* The top right figure illustrates the posterior distribution of the P-T profiles, that has the contribution function overlaid on top of it in shades of white highlighting visibility. The equilibrium state $MgSiO_3$, MnS and Fe condensation curves are plotted along the profile in purple, green and red respectively. The profiles of the most probable sample from NPE (dark blue) and *MultiNest* (orange) are also plotted.

posterior *MultiNest* widths underscores the trend observed regarding the under-dispersion of *MultiNest* described in the previous chapter (see Section 6.5.1), and also in works by Ardévol Martínez et al. (2022b); Vasist et al. (2023); Latouf et al. (2023). Specifically, Chubb and Min (2022) suggests that a low sampling efficiency of 5% used here, as opposed to the recommended 80% for parameter estimation, and a constant efficiency mode can lead to such under-dispersion. Further, variations in the median values can also be attributed to the different forward models used to conduct these retrievals (see Table 8.1).

The figure also illustrates the spread in the NPE posterior *P-T* profiles, along with the most probable sample from the NPE and *MultiNest* retrievals. The different parameterizations of the profile are apparent here. However, within the photosphere, the two profiles agree.

8.5 Conclusion

Based on the proof of concept established in this study, we demonstrate that NPE is a viable retrieval method for high-resolution spectra. The results show that NPE provides physically consistent constraints on key atmospheric and kinematic parameters, aligning well with previous studies, while offering computational efficiency.

Given its robust performance, NPE can be extended to retrieve parameters across all spectral orders in high-resolution datasets. This capability opens the door for large-scale, fast population studies of brown dwarfs and exoplanets, enabling a more comprehensive exploration of atmospheric properties and their formation pathways across diverse objects. Future applications of NPE could significantly accelerate the characterization of planetary atmospheres in high-resolution spectroscopic surveys such as CRIRES+.

Chapter 9

Challenges and future of SBI in atmospheric retrievals

9.1 Main takeaways

In this thesis, we aimed to answer two main questions, namely:

1. How to address the shortcomings of the conventional algorithms?
2. What can we say about cold brown dwarf atmospheres from their mid-infrared spectra?

and we addressed them in the following way.

We explored the concept of simulation-based inference (SBI), which side-steps the explicit computation of the likelihood function, to make it faster, scalable and testable. Originally developed to handle intractable likelihoods, SBI, especially the neural inference algorithms within this family, improve parallelization and scalability. Among these algorithms, we use the neural posterior estimation (NPE), that directly estimates the posterior of the model parameters considered. The main benefit of NPE is amortization, which allows the network to be trained once and then used to retrieve multiple observations almost instantaneously. This makes NPE incredibly time efficient for bulk retrievals, especially for systematic studies of similarly modeled objects. A comparison of NPE with `MultiNest` over 1000 retrievals of synthetic spectra reveals a boost of wall-clock time by a factor 4000 for a similar hardware setup. Amortization is incredibly useful for future population studies in the wake of JWST and future missions such as ARIEL. Although by excessively parallelizing `MultiNest` one can achieve comparable times at lower model parameter dimensions, each retrieval needs to start from scratch, which is not time efficient for multiple observations. In contrast, NPE offers amortization, which not only provides a boost to bulk retrievals but also allows for reusing the training set to perform retrievals on more statistically complex models (such as mixture and patchy clouds, etc) and various noise models. Additionally, amortization makes the inference testable. Since many retrievals can be performed almost quasi-instantaneously, it enables validity checks such

as posterior predictive checks and coverage tests. This ensures statistical rigor to the results obtained, which is generally unfeasible with conventional algorithms. We benchmark NPE with the `Multinest` algorithm by performing a retrieval over a synthetic spectrum of exoplanet HR 8799 e and found consistent results. This work is described in Chapter 4.

We performed NPE retrievals on brown dwarf spectra obtained from the MIRI GTO program. We built a comprehensive understanding of the Y dwarf WISE 1738 by retrieving over its near and mid-infrared spectra obtained from HST/WFC3, GNIRS/Gemini and JWST/MIRI. Based on comparing independent retrievals on each of these datasets with a combined retrieval on all of them, we cautioned against potential biases in posterior estimates when only a narrow wavelength range is accessible. We constrained bulk physical properties of WISE 1738 that are consistent with evolutionary models. We also constrained its chemical properties by estimating the abundances of major opacity species and identifying traces of vertical mixing in its atmosphere. This work is described in Chapter 5.

To investigate theoretical claims of finding water clouds in the atmosphere of WISE 1738, we performed fully and patchy (water ice) cloudy retrievals on its spectrum. This resulted in posteriors with very broad PPD. To explain the source of such predictive uncertainty, we performed various tests, which reveal that they are mainly epistemic in nature. We showed that this is due to broad prior assumptions, with insufficient information in the wavelength range to uniquely and confidently estimate the cloud parameters, and an overly simplistic assumption of gaussian noise. This latter of which leads to degeneracies between the cloud parameters and radius, further broadening the posteriors. We improved epistemic uncertainty by using patchy clouds in our models instead of thick clouds, and performed a Bayes factor test to discover that it is statistically preferred over cloud-free and the cloudy model. However, we concluded that the cloud-free model is a better predictive model due to its narrow posterior predictive distribution. This lead us to prefer the cloud-free model as the most suitable one to describe the combined spectrum of WISE 1738. Additionally, we apply importance sampling to remove any bias from the NPE posterior estimates and find that it is hard to implement, possibly due to noise model mismatch in a retrieval at higher dimensions. This work is described in Chapter 6.

We performed a comprehensive analysis on the combined HST/WFC3 and JWST/MIRI spectra of WISE 1828, which is a brown dwarf very close in spectral type to WISE 1738. We further benchmark it with a `MultiNest` retrieval. We constrained both the physical and chemical properties of WISE 1828 via cloud-free and cloudy model retrievals. To shed light on its formation, we evaluated the $^{14}N/^{15}N$ obtained from the associated different isotopologues of ammonia. We found that the ratio overlaps with values cannot be interpreted to link to any specific formation scenario. Interestingly, the physical parameters constrained are consistent with the evolutionary models under the assumption of binarity (i.e for binary mass and radius). Further, we found that although CO and H_2S are constrained in other retrievals, NPE does not constrain it. In general, we found the positive correlation that is hypothesized to exist between the

molecular abundances and constrained surface gravity, when we compared our retrieval with all other retrievals of this source. NPE retrieval for this object adds on to the variation seen in constrained values for its physical and chemical properties of WISE 1828, insinuating that some key aspect might be missing in the forward model, or some systematic noise could be leading to such compensations in the retrieval. This is also implied by the poor predictive strength of both cloudy and cloud-free retrieved models, although, suggesting better predictive power for the latter compared to the former. Furthermore, we conducted a preliminary comparative study on all five isolated brown dwarfs observed within the MIRI GTO program, that include not only WISE 1738 and WISE 1828 but also ROSS 458c, WISE 0458 and WISE 0855, to find any trends in their physical and chemical properties that can inform us about their formation scenarios or current atmospheric mechanisms operating in their atmospheres. This experiment was setup as a precedent for future population studies using NPE/SBI. We found that for all atmospheres, there is a deviation of ammonia from its chemical equilibrium value, which suggests vertical mixing. We did not find any other significant trends in the atmospheric properties, leading us to conclude that a systematic study on a larger number of objects with diverse properties is necessary to discover such trends. This work is described in Chapter 7.

The last study involved a pilot NPE spectral retrieval of a single channel of the high-resolution spectrum of DENIS J0255, obtained using the CRIRES+ instrument. Interestingly, we did not find evidence for ^{13}CO in its atmosphere, whose presence is tentative based on previous works. We benchmark our results against `MultiNest` and found significant overlaps between the two posterior estimates, albeit with slight difference in constraints. The cause of this difference is difficult to ascertain since the forward models are significantly different. However, NPE provided a posterior estimate consistent with the observations, hence establishing it as a reliable tool for high-resolution retrievals. This work is described in Chapter 8.

9.2 Challenges of SBI

Although NPE has proven to offer significant advantages, it also presents some challenges. Here we identify them and find ways to mitigate them.

First, in NPE, the upfront simulation of a large training set is costly both in terms of computation and time. It also requires a much larger training set than its sequential counterpart SNPE (Ardévol Martínez et al., 2024) or the variational inference targeted at a single spectrum (Yip et al., 2022). Therefore, the main challenge with amortized inference is to *identify a simulation model that is general enough* to be applicable and valid in many situations, so that the whole training process does not need to be repeated for each individual case. This may be possible for studies focusing on specific classes of brown dwarfs such as in this thesis, or hot Jupiters observed in transit, or self-luminous giant planets observed with direct imaging.

Second, NPE does not provide a direct estimate of the evidence (marginal likelihood), in contrast to NS algorithms, which give the evidence as a natural byproduct of the inference process. Since evidence is crucial for performing model comparison between competing models, such comparisons are not as straightforward with NPE. Additional procedures or experiments are therefore required to estimate the evidence when using NPE. Instead, it is easier to compute a quantity that is proportional to the marginal likelihoods between two competing models (the Bayes factor) by posing it as a neural classification problem leveraging the likelihood ratio trick.

Third, the current set-up of NPE has an embedding network which requires the input observation to have a fixed length. This limits the flexibility of observations that can be retrieved.

Fourth, the current SBI package LAMPE allows users the freedom to modify or replace any neural architecture as they see fit. Although this is extremely useful, it is an interdisciplinary task that requires domain knowledge of the field of machine learning and profound insight on SBI. These stringent requirements risk limiting the use of SBI in the community of exoplanet studies to isolated initiatives, potentially making SBI less popular than the conventional algorithms like MCMC and NS, which have far fewer parameters to tune.

Since the first two challenges are inherent to the nature of NPE, they cannot be fully eliminated. However they can be significantly improved upon. For example, generating large data sets in advance that can be readily used for training (Zorzan et al., 2025), or pre-training Bayes factor classifiers across several potential models, can significantly reduce the overhead time of the retrievals. In contrast, the third and the fourth challenges can be fully mitigated. For instance, the embedding network can be made flexible using networks such as transformers. This would adjust the positional encoding dynamically based on the input length, irrespective of the length of the training dataset. While this was attempted during the course of the thesis, it needs further exploration. Furthermore, the fourth challenge can be mitigated by building a simplified SBI package in the future that can mask the underlying complexity and enhance its accessibility to exoplanet scientists. Accounting for these improvements would enhance the reach and usability of NPE for exoplanet retrievals. Broadly, NPE and SBI algorithms in general have a wide scope of application in exoplanet atmospheric inference in the future.

9.3 Future of SBI

In this thesis, we took the first step of applying the SBI algorithm NPE to analyze the spectra from several brown dwarfs obtained from direct imaging using JWST. However, in the next ten to twenty years, instruments such as the ELT, NGRST, LIFE, HWO are either operational or planned to contribute to characterizing exoplanets via direct imaging. These will directly image tens or hundreds of exoplanets, constituting not only planets more massive than Jupiter, but also a few super-Earths. These missions will immensely extend our understanding of exoplanets and their diversity since all the exoplanets imaged so far are young (mostly $<$100 million

years old) and self-luminous (hence hot) gas giants, that lie at a distance $>$0.3 arc seconds away from the host star. However, with the technological advancements of coronagraphs, it will be possible to image dimmer (and therefore cooler) and mature gas giants that are over several billion years old, located just beyond the habitable zone of nearly sun-like stars ($>$0.15 arcseconds) and in some special cases even inside it.

Looking beyond direct imaging, ARIEL is the European Space Agency's space mission dedicated to the characterization of exoplanets with infrared spectroscopy, due to be launched in 2029. This mission aims to observe at least 1000 known exoplanets using the transit method, and study and characterize their chemical composition, cloud coverage, and thermal structures. Given this vast number of planet candidates in several regions of planetary systems, it is crucial to adopt a systematic approach for characterizing their atmospheres. To identify the best way to interpret the bulk of data we will receive from the ARIEL mission, the Ariel Data Challenge Organizing team have been conducting a Data Challenge each year for the last three years. Each year, the difficulty is increased. In 2023, the training set was sparser, the planetary observations were more difficult, and the chemistry was more non-linear than in 2022. During this year, an SBI algorithm implemented by the AstroAI team (Aubin et al., 2023) secured the top position among the 293 competitors who participated. The SBI's win in the ARIEL data challenge does not come as a surprise, owing to the advantages that are identified and verified in this work. Therefore, we conclude that SBI is a powerful tool to have in the characterization toolkit to perform rapid large-scale systematic retrievals on exoplanet spectra.

Bibliography

A. S. Ackerman and M. S. Marley. Precipitating condensation clouds in substellar atmospheres. *The Astrophysical Journal*, 556(2):872, 2001.

D. B. F. Agakov. The im algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

F. Allard, P. H. Hauschildt, I. Baraffe, and G. Chabrier. Synthetic Spectra and Mass Determination of the Brown Dwarf GI 229B. *ApJ*, 465:L123, July 1996. doi: 10.1086/310143.

J. Alsing, B. Wandelt, and S. Feeney. Massive optimal data compression and density estimation for scalable, likelihood-free inference in cosmology. *Monthly Notices of the Royal Astronomical Society*, 477(3):2874–2885, 2018.

D. S. Amundsen, P. Tremblin, J. Manners, I. Baraffe, and N. J. Mayne. Treatment of overlapping gaseous absorption with the correlated-k method in hot Jupiter and brown dwarf atmosphere models. *A&A*, 598:A97, Feb. 2017. doi: 10.1051/0004-6361/201629322.

R. Andrae. Error estimation in astronomy: A guide. *arXiv e-prints*, art. arXiv:1009.2755, Sept. 2010. doi: 10.48550/arXiv.1009.2755.

R. Andrae, T. Schulze-Hartung, and P. Melchior. Dos and don'ts of reduced chi-squared. *arXiv e-prints*, art. arXiv:1012.3754, Dec. 2010. doi: 10.48550/arXiv.1012.3754.

A. M. Arabhavi, I. Kamp, T. Henning, E. F. van Dishoeck, V. Christiaens, D. Gasman, A. Perrin, M. Güdel, B. Tabone, J. Kanwar, L. B. F. M. Waters, I. Pascucci, M. Samland, G. Perotti, G. Bettoni, S. L. Grant, P. O. Lagage, T. P. Ray, B. Vandenbussche, O. Absil, I. Argyriou, D. Barrado, A. Boccaletti, J. Bouwman, A. Caratti o Garatti, A. M. Glauser, F. Lahuis, M. Mueller, G. Olofsson, E. Pantin, S. Scheithauer, M. Morales-Calderón, R. Franceschi, H. Jang, N. Pawellek, D. Rodgers-Lee, J. Schreiber, K. Schwarz, M. Temmink, M. Vlasblom, G. Wright, L. Colina, and G. Östlin. Abundant hydrocarbons in the disk around a very-low-mass star. Science, 384(6700):1086–1090, June 2024. doi: 10.1126/science.adi8147.

F. Ardévol Martínez. *Machine learning for exoplanet characterisation in the JWST era*. PhD thesis, University of Groningen, 2024.

F. Ardévol Martínez, M. Min, I. Kamp, and P. I. Palmer. Convolutional neural networks as an alternative to Bayesian retrievals for interpreting exoplanet transmission spectra. *A&A*, 662: A108, June 2022a. doi: 10.1051/0004-6361/202142976.

F. Ardévol Martínez, M. Min, I. Kamp, and P. I. Palmer. Convolutional neural networks as an alternative to Bayesian retrievals for interpreting exoplanet transmission spectra. *A&A*, 662: A108, June 2022b. doi: 10.1051/0004-6361/202142976.

F. Ardévol Martínez, M. Min, D. Huppenkothen, I. Kamp, and P. I. Palmer. FlopPITy: Enabling self-consistent exoplanet atmospheric retrievals with machine learning. *A&A*, 681:L14, Jan. 2024. doi: 10.1051/0004-6361/202348367.

I. Argyriou, A. Glasse, D. R. Law, A. Labiano, J. Álvarez-Márquez, P. Patapis, P. J. Kavanagh, D. Gasman, M. Mueller, K. Larson, B. Vandenbussche, A. M. Glauser, P. Royer, D. Dicken, J. Harkett, B. A. Sargent, M. Engesser, O. C. Jones, S. Kendrew, A. Noriega-Crespo, B. Brandl, G. H. Rieke, G. S. Wright, D. Lee, and M. Wells. JWST MIRI flight performance: The Medium-Resolution Spectrometer. *A&A*, 675:A111, July 2023. doi: 10.1051/0004-6361/202346489.

M. Aubin, C. Cuesta-Lazaro, E. Tregidga, J. Viaña, C. Garraffo, I. E. Gordon, M. López-Morales, R. J. Hargreaves, V. Y. Makhnev, J. J. Drake, D. P. Finkbeiner, and P. Cargile. Simulation-based Inference for Exoplanet Atmospheric Retrieval: Insights from winning the Ariel Data Challenge 2023 using Normalizing Flows. *arXiv e-prints*, art. arXiv:2309.09337, Sept. 2023. doi: 10.48550/arXiv.2309.09337.

S. Ballard, D. Fabrycky, F. Fressin, D. Charbonneau, J.-M. Desert, G. Torres, G. Marcy, C. J. Burke, H. Isaacson, C. Henze, J. H. Steffen, D. R. Ciardi, S. B. Howell, W. D. Cochran, M. Endl, S. T. Bryson, J. F. Rowe, M. J. Holman, J. J. Lissauer, J. M. Jenkins, M. Still, E. B. Ford, J. L. Christiansen, C. K. Middour, M. R. Haas, J. Li, J. R. Hall, S. McCauliff, N. M. Batalha, D. G. Koch, and W. J. Borucki. The Kepler-19 System: A Transiting 2.2 R $_\oplus$ Planet and a Second Planet Detected via Transit Timing Variations. *ApJ*, 743(2):200, Dec. 2011. doi: 10.1088/0004-637X/743/2/200.

R. J. Barber, J. Tennyson, G. J. Harris, and R. N. Tolchenov. A high-accuracy computed water line list. *Monthly Notices of the Royal Astronomical Society*, 368(3):1087–1094, 04 2006. ISSN 0035-8711. doi: 10.1111/j.1365-2966.2006.10184.x. URL `https://doi.org/10.1111/j.1365-2966.2006.10184.x`.

D. Barrado, P. Mollière, P. Patapis, M. Min, P. Tremblin, F. Ardevol Martinez, N. Whiteford, M. Vasist, I. Argyriou, M. Samland, et al. 15nh3 in the atmosphere of a cool brown dwarf. *Nature*, 624(7991):263–266, 2023.

C. Beichman, C. R. Gelino, J. D. Kirkpatrick, T. S. Barman, K. A. Marsh, M. C. Cushing, and E. L. Wright. The Coldest Brown Dwarf (or Free-floating Planet)?: The Y Dwarf WISE 1828+2650. *ApJ*, 764(1):101, Feb. 2013. doi: 10.1088/0004-637X/764/1/101.

C. Beichman, C. R. Gelino, J. D. Kirkpatrick, M. C. Cushing, S. Dodson-Robinson, M. S. Marley, C. V. Morley, and E. L. Wright. Wise y dwarfs as probes of the brown dwarf–exoplanet connection. *ApJ*, 783(2):68, feb 2014. doi: 10.1088/0004-637X/783/2/68. URL `https://dx.doi.org/10.1088/0004-637X/783/2/68`.

S. A. Beiler, M. C. Cushing, J. D. Kirkpatrick, A. C. Schneider, S. Mukherjee, M. S. Marley, F. Marocco, and R. L. Smart. Precise Bolometric Luminosities and Effective Temperatures of 23 Late-T and Y Dwarfs Obtained with JWST. *ApJ*, 973(2):107, Oct. 2024a. doi: 10.3847/1538-4357/ad6301.

S. A. Beiler, S. Mukherjee, M. C. Cushing, J. D. Kirkpatrick, A. C. Schneider, H. Kothari, M. S. Marley, and C. Visscher. A Tale of Two Molecules: The Underprediction of CO_2 and Overprediction of PH_3 in Late T and Y Dwarf Atmospheric Models. *ApJ*, 973(1):60, Sept. 2024b. doi: 10.3847/1538-4357/ad6759.

B. Benneke and S. Seager. How to distinguish between cloudy mini-neptunes and water/volatile-dominated super-earths. *The Astrophysical Journal*, 778(2):153, 2013.

J. Birkby, R. de Kok, M. Brogi, H. Schwarz, S. Albrecht, E. de Mooij, and I. Snellen. Characterising Exoplanet Atmospheres with High-resolution Spectroscopy. *The Messenger*, 154:57–61, Dec. 2013.

J. L. Birkby. Exoplanet Atmospheres at High Spectral Resolution. *arXiv e-prints*, art. arXiv:1806.04617, June 2018. doi: 10.48550/arXiv.1806.04617.

M. Blanc, G. J. Herczeg, V. Sterken, H. Lammer, W. Benz, S. Udry, R. Rodrigo, and M. Falanga. *From Disks to Planets: The Making of Planets and Their Early Atmospheres*. 2018. doi: 10.1007/978-94-024-1518-6.

J. Blecic. Observations, thermochemical calculations, and modeling of exoplanetary atmospheres. *arXiv preprint arXiv:1604.02692*, 2016.

R. A. Booth, C. J. Clarke, N. Madhusudhan, and J. D. Ilee. Chemical enrichment of giant planets and discs due to pebble drift. *MNRAS*, 469(4):3994–4011, Aug. 2017. doi: 10.1093/mnras/stx1103.

A. D. Bosman, A. J. Cridland, and Y. Miguel. Jupiter formed as a pebble pile around the N_2 ice line. *A&A*, 632:L11, Dec. 2019. doi: 10.1051/0004-6361/201936827.

A. P. Boss. Formation of Extrasolar Giant Planets: Core Accretion or Disk Instability? *Earth Moon and Planets*, 81(1):19–26, Jan. 1998. doi: 10.1023/A:1006370021545.

J. Bouvier, S. P. Matt, S. Mohanty, A. Scholz, K. G. Stassun, and C. Zanni. Angular Momentum Evolution of Young Low-Mass Stars and Brown Dwarfs: Observations and Theory. In H. Beuther, R. S. Klessen, C. P. Dullemond, and T. Henning, editors, *Protostars and Planets VI*, pages 433–450, Jan. 2014. doi: 10.2458/azu_uapress_9780816531240-ch019.

B. P. Bowler, M. C. Liu, T. J. Dupuy, and M. C. Cushing. Near-infrared Spectroscopy of the Extrasolar Planet HR 8799 b. *ApJ*, 723(1):850–868, Nov. 2010. doi: 10.1088/0004-637X/723/1/850.

M. Brogi and M. R. Line. Retrieving Temperatures and Abundances of Exoplanet Atmospheres with High-resolution Cross-correlation Spectroscopy. *AJ*, 157(3):114, Mar. 2019. doi: 10.3847/1538-3881/aaffd3.

M. Brogi, I. A. G. Snellen, R. J. de Kok, S. Albrecht, J. Birkby, and E. J. W. de Mooij. The signature of orbital motion from the dayside of the planet τ Boötis b. *Nature*, 486(7404):502–504, June 2012. doi: 10.1038/nature11161.

M. Brogi, R. J. de Kok, S. Albrecht, I. A. G. Snellen, J. L. Birkby, and H. Schwarz. Rotation and winds of exoplanet hd 189733 b measured with high-dispersion transmission spectroscopy. *The Astrophysical Journal*, 817(2):106, jan 2016. doi: 10.3847/0004-637X/817/2/106. URL `https://dx.doi.org/10.3847/0004-637X/817/2/106`.

J. Buchner, A. Georgakakis, K. Nandra, L. Hsu, C. Rangel, M. Brightman, A. Merloni, M. Salvato, J. Donley, and D. Kocevski. X-ray spectral modelling of the AGN obscuring region in the CDFS: Bayesian model selection and catalogue. *A&A*, 564:A125, Apr. 2014. doi: 10.1051/0004-6361/201322971.

B. Burningham, M. S. Marley, M. R. Line, R. Lupu, C. Visscher, C. V. Morley, D. Saumon, and R. Freedman. Retrieval of atmospheric properties of cloudy l dwarfs. *Monthly Notices of the Royal Astronomical Society*, 470(1):1177–1197, 2017.

A. Burrows and C. M. Sharp. Chemical Equilibrium Abundances in Brown Dwarf and Extrasolar Giant Planet Atmospheres. *ApJ*, 512(2):843–863, Feb. 1999. doi: 10.1086/306811.

A. Burrows, W. B. Hubbard, J. I. Lunine, and J. Liebert. The theory of brown dwarfs and extrasolar giant planets. *Rev. Mod. Phys.*, 73(3):719–765, July 2001. doi: 10.1103/RevModPhys.73.719.

A. Burrows, L. Ibgui, and I. Hubeny. Optical Albedo Theory of Strongly Irradiated Giant Planets: The Case of HD 209458b. *ApJ*, 682(2):1277–1282, Aug. 2008. doi: 10.1086/589824.

H. Bushouse, J. Eisenhamer, N. Dencheva, J. Davies, P. Greenfield, J. Morrison, P. Hodge, B. Simon, D. Grumm, M. Droettboom, E. Slavich, M. Sosey, T. Pauly, T. Miller, R. Jedrzejewski, W. Hack, D. Davis, S. Crawford, D. Law, K. Gordon, M. Regan, M. Cara, K. MacDonald, L. Bradley, C. Shanahan, W. Jamieson, M. Teodoro, and T. Williams. JWST Calibration Pipeline, Jan. 2023.

S. H. C. Cabot, N. Madhusudhan, G. A. Hawker, and S. Gandhi. On the robustness of analysis techniques for molecular detections using high-resolution exoplanet spectroscopy. *MNRAS*, 482(4):4422–4436, Feb. 2019. doi: 10.1093/mnras/sty2994.

Y. Chachan, H. A. Knutson, J. Lothringer, and G. A. Blake. Breaking Degeneracies in Formation Histories by Measuring Refractory Content in Gas Giants. *ApJ*, 943(2):112, Feb. 2023. doi: 10.3847/1538-4357/aca614.

D. Charbonneau, T. M. Brown, D. W. Latham, and M. Mayor. Detection of planetary transits across a sun-like star. *The Astrophysical Journal*, 529(1):L45, dec 1999. doi: 10.1086/312457. URL https://dx.doi.org/10.1086/312457.

D. Charbonneau, T. M. Brown, R. W. Noyes, and R. L. Gilliland. Detection of an Extrasolar Planet Atmosphere. *ApJ*, 568(1):377–384, Mar. 2002. doi: 10.1086/338770.

D. Charbonneau, L. Allen, T. Barman, F. Bouchy, T. Brown, M. Mayor, T. Megeath, C. Moutou, D. Queloz, and S. Udry. Thermal Emission from the Newest, Closest, and Brightest Transiting Planet. Spitzer Proposal ID 261, Oct. 2005.

B. Charnay, B. Bézard, J. L. Baudino, M. Bonnefoy, A. Boccaletti, and R. Galicher. A Self-consistent Cloud Model for Brown Dwarfs and Young Giant Exoplanets: Comparison with Photometric and Spectroscopic Observations. *ApJ*, 854(2):172, Feb. 2018. doi: 10.3847/1538-4357/aaac7d.

B. Charnay, B. Bézard, J.-L. Baudino, M. Bonnefoy, A. Boccaletti, and R. Galicher. A self-consistent cloud model for brown dwarfs and young giant exoplanets: Comparison with photometric and spectroscopic observations. *The Astrophysical Journal*, 854(2):172, feb 2018. doi: 10.3847/1538-4357/aaac7d. URL https://dx.doi.org/10.3847/1538-4357/aaac7d.

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural Ordinary Differential Equations. *arXiv e-prints*, art. arXiv:1806.07366, June 2018. doi: 10.48550/arXiv.1806.07366.

K. L. Chubb and M. Min. Exoplanet Atmosphere Retrievals in 3D Using Phase Curve Data with ARCiS: Application to WASP-43b. *A&A*, 665:A2, Sept. 2022. doi: 10.1051/0004-6361/202142800.

D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

A. D. Cobb, M. D. Himes, F. Soboczenski, S. Zorzan, M. D. O'Beirne, A. G. Baydin, Y. Gal, S. D. Domagal-Goldman, G. N. Arney, D. Angerhausen, and . N. F. A. T. II. An ensemble of bayesian neural networks for exoplanetary atmospheric retrieval. *The Astronomical Journal*, 158(1):33, jun 2019. doi: 10.3847/1538-3881/ab2390. URL https://dx.doi.org/10.3847/1538-3881/ab2390.

A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino, and C. Weniger. Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation. *Journal of Cosmology and Astroparticle Physics*, 2022(09):004, sep 2022. doi: 10.1088/1475-7516/2022/09/004. URL https://dx.doi.org/10.1088/1475-7516/2022/09/004.

K. Cranmer, J. Brehmer, and G. Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020. doi: 10.1073/pnas.1912789117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1912789117.

M. J. Creech-Eakman, G. S. Orton, E. Serabyn, and T. L. Hayward. Mid-Infrared Detection of the L Dwarf DENISP J0255-4700. *ApJ*, 602(2):L129–L132, Feb. 2004. doi: 10.1086/382589.

A. J. Cridland, R. E. Pudritz, and M. Alessi. Composition of early planetary atmospheres - I. Connecting disc astrochemistry to the formation of planetary atmospheres. *MNRAS*, 461(3): 3274–3295, Sept. 2016. doi: 10.1093/mnras/stw1511.

A. J. Cridland, E. F. van Dishoeck, M. Alessi, and R. E. Pudritz. Connecting planet formation and astrochemistry. C/Os and N/Os of warm giant planets and Jupiter analogues. *A&A*, 642:A229, Oct. 2020. doi: 10.1051/0004-6361/202038767.

I. J. M. Crossfield. Volatile-to-sulfur Ratios Can Recover a Gas Giant's Accretion History. *ApJ*, 952(1):L18, July 2023. doi: 10.3847/2041-8213/ace35f.

T. Currie, A. Burrows, Y. Itoh, S. Matsumura, M. Fukagawa, D. Apai, N. Madhusudhan, P. M. Hinz, T. J. Rodigas, M. Kasper, T. S. Pyo, and S. Ogino. A Combined Subaru/VLT/MMT 1-5 μm Study of Planets Orbiting HR 8799: Implications for Atmospheric Properties, Masses, and Formation. *ApJ*, 729(2):128, Mar. 2011. doi: 10.1088/0004-637X/729/2/128.

T. Currie, B. Biller, A. Lagrange, C. Marois, O. Guyon, E. L. Nielsen, M. Bonnefoy, and R. J. De Rosa. Direct Imaging and Spectroscopy of Extrasolar Planets. In S. Inutsuka, Y. Aikawa, T. Muto, K. Tomida, and M. Tamura, editors, *Protostars and Planets VII*, volume 534 of *Astronomical Society of the Pacific Conference Series*, page 799, July 2023. doi: 10.48550/arXiv.2205.05696.

M. C. Cushing, J. T. Rayner, and W. D. Vacca. An Infrared Spectroscopic Sequence of M, L, and T Dwarfs. *ApJ*, 623(2):1115–1140, Apr. 2005. doi: 10.1086/428040.

M. C. Cushing, M. S. Marley, D. Saumon, B. C. Kelly, W. D. Vacca, J. T. Rayner, R. S. Freedman, K. Lodders, and T. L. Roellig. Atmospheric Parameters of Field L and T Dwarfs. *ApJ*, 678(2): 1372–1395, May 2008. doi: 10.1086/526489.

M. C. Cushing, J. D. Kirkpatrick, C. R. Gelino, R. L. Griffith, M. F. Skrutskie, A. Mainzer, K. A. Marsh, C. A. Beichman, A. J. Burgasser, L. A. Prato, et al. The discovery of y dwarfs using data from the wide-field infrared survey explorer (wise). *ApJ*, 743(1):50, 2011.

S. Czesla, S. Schröter, C. P. Schneider, K. F. Huber, F. Pfeifer, D. T. Andreasen, and M. Zechmeister. PyA: Python astronomy-related packages. Astrophysics Source Code Library, record ascl:1906.010, June 2019.

M. De Furio, B. Lew, C. Beichman, T. Roellig, G. Bryden, D. Ciardi, M. Meyer, M. Rieke, A. Greenbaum, J. Leisenring, J. Llop-Sayson, M. Ygouf, L. Albert, M. Boyer, D. Eisenstein, K. Hodapp, S. Horner, D. Johnstone, D. Kelly, K. Misselt, G. Rieke, J. Stansberry, and E. Young. Jwst observations of the enigmatic y-dwarf wise 1828+2650. i. limits to a binary companion. *The Astrophysical Journal*, 948(2):92, may 2023. doi: 10.3847/1538-4357/acbf1e. URL `https://dx.doi.org/10.3847/1538-4357/acbf1e`.

S. de Regt, S. Gandhi, I. A. G. Snellen, Y. Zhang, C. Ginski, D. González Picos, A. Y. Kesseli, R. Landman, P. Mollière, E. Nasedkin, A. Sánchez-López, and T. Stolker. The ESO SupJup Survey. I. Chemical and isotopic characterisation of the late L-dwarf DENIS J0255-4700 with CRIRES$^+$. *A&A*, 688:A116, Aug. 2024. doi: 10.1051/0004-6361/202348508.

A. Delaunoy, M. de la Brassinne Bonardeaux, S. Mishra-Sharma, and G. Louppe. Low-Budget Simulation-Based Inference with Bayesian Neural Networks. *arXiv e-prints*, art. arXiv:2408.15136, Aug. 2024. doi: 10.48550/arXiv.2408.15136.

C. Durkan, I. Murray, and G. Papamakarios. On contrastive learning for likelihood-free inference. In *International Conference on Machine Learning*, pages 2771–2781. PMLR, 2020.

D. Ehrenreich, C. Lovis, R. Allart, M. R. Zapatero Osorio, F. Pepe, S. Cristiani, R. Rebolo, N. C. Santos, F. Borsa, O. Demangeon, X. Dumusque, J. I. González Hernández, N. Casasayas-Barris, D. Ségransan, S. Sousa, M. Abreu, V. Adibekyan, M. Affolter, C. Allende Prieto, Y. Alibert, M. Aliverti, D. Alves, M. Amate, G. Avila, V. Baldini, T. Bandy, W. Benz, A. Bianco, É. Bolmont, F. Bouchy, V. Bourrier, C. Broeg, A. Cabral, G. Calderone, E. Pallé, H. M. Cegla, R. Cirami, J. M. P. Coelho, P. Conconi, I. Coretti, C. Cumani, G. Cupani, H. Dekker, B. Delabre, S. Deiries, V. D'Odorico, P. Di Marcantonio, P. Figueira, A. Fragoso, L. Genolet, M. Genoni, R. Génova Santos, N. Hara, I. Hughes, O. Iwert, F. Kerber, J. Knudstrup, M. Landoni, B. Lavie, J.-L. Lizon, M. Lendl, G. Lo Curto, C. Maire, A. Manescau, C. J. A. P. Martins, D. Mégevand, A. Mehner, G. Micela, A. Modigliani, P. Molaro, M. Monteiro, M. Monteiro, M. Moschetti, E. Müller, N. Nunes, L. Oggioni, A. Oliveira, G. Pariani, L. Pasquini, E. Poretti, J. L. Rasilla, E. Redaelli, M. Riva, S. Santana Tschudi, P. Santin, P. Santos, A. Segovia Milla, J. V. Seidel, D. Sosnowska, A. Sozzetti, P. Spanò, A. Suárez Mascareño, H. Tabernero, F. Tenegi, S. Udry, A. Zanutta, and F. Zerbi. Nightside condensation of iron in an ultrahot giant exoplanet. *Nature*, 580(7805):597–601, Apr. 2020. doi: 10.1038/s41586-020-2107-1.

J. H. Elias, R. R. Joyce, M. Liang, G. P. Muller, E. A. Hileman, and J. R. George. Design of the Gemini near-infrared spectrograph. In I. S. McLean and M. Iye, editors, *Ground-based and Airborne Instrumentation for Astronomy*, volume 6269, page 62694C. International Society for Optics and Photonics, SPIE, 2006a. doi: 10.1117/12.671817. URL `https://doi.org/10.1117/12.671817`.

J. H. Elias, B. Rodgers, R. R. Joyce, M. Lazo, G. Doppmann, C. Winge, and A. Rodríguez-Ardila. Performance of the gemini near-infrared spectrograph. In *Ground-based and Airborne Instrumentation for Astronomy*, volume 6269, pages 374–385. SPIE, 2006b.

T. M. Evans, D. K. Sing, T. Kataria, J. Goyal, N. Nikolov, H. R. Wakeford, D. Deming, M. S. Marley, D. S. Amundsen, G. E. Ballester, et al. An ultrahot gas-giant exoplanet with a stratosphere. *Nature*, 548(7665):58–61, 2017.

P. Feautrier. Sur la resolution numerique de l'equation de transfert. *Comptes Rendus Academie des Sciences (serie non specifiee)*, 258:3189, 1964.

B. Fegley, Jr. and K. Lodders. Chemical Models of the Deep Atmospheres of Jupiter and Saturn. *Icarus*, 110(1):117–154, July 1994. doi: 10.1006/icar.1994.1111.

F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *MNRAS*, 384(2): 449–463, Feb 2008. doi: 10.1111/j.1365-2966.2007.12353.x.

F. Feroz, M. P. Hobson, and M. Bridges. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. *MNRAS*, 398(4):1601–1614, Oct 2009. doi: 10.1111/j.1365-2966.2009.14548.x.

F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt. Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, 2(1):10, Nov. 2019a. doi: 10.21105/astro.1306.2144.

F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt. Importance Nested Sampling and the MultiNest Algorithm. *The Open Journal of Astrophysics*, 2(1):10, Nov. 2019b. doi: 10.21105/astro.1306.2144.

E. Flowers, M. Brogi, E. Rauscher, E. M.-R. Kempton, and A. Chiavassa. The high-resolution transmission spectrum of hd 189733b interpreted with atmospheric doppler shifts from three-dimensional general circulation models. *The Astronomical Journal*, 157(5):209, may 2019. doi: 10.3847/1538-3881/ab164c. URL `https://dx.doi.org/10.3847/1538-3881/ab164c`.

R. Follert, R. J. Dorn, E. Oliva, J. L. Lizon, A. Hatzes, N. Piskunov, A. Reiners, U. Seemann, E. Stempels, U. Heiter, T. Marquart, M. Lockhart, G. Anglada-Escude, T. Löwinger, D. Baade, J. Grunhut, P. Bristow, B. Klein, Y. Jung, D. J. Ives, F. Kerber, E. Pozna, J. Paufique, H. U. Kaeufl, L. Origlia, E. Valenti, D. Gojak, M. Hilker, L. Pasquini, A. Smette, and J. Smoker. CRIRES+: a cross-dispersed high-resolution infrared spectrograph for the ESO VLT. In S. K. Ramsay, I. S. McLean, and H. Takami, editors, *Ground-based and Airborne Instrumentation for Astronomy V*, volume 9147, page 914719. International Society for Optics and Photonics, SPIE, 2014. doi: 10.1117/12.2054197. URL `https://doi.org/10.1117/12.2054197`.

J. Gaarn, B. Burningham, J. K. Faherty, C. Visscher, M. S. Marley, E. C. Gonzales, E. Calamari, D. Bardalez Gagliuffi, R. Lupu, and R. Freedman. The puzzle of the formation of T8 dwarf Ross 458c. *MNRAS*, 521(4):5761–5775, June 2023. doi: 10.1093/mnras/stad753.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/gal16.html`.

S. Gandhi and N. Madhusudhan. Retrieval of exoplanet emission spectra with hydra. *Monthly Notices of the Royal Astronomical Society*, 474(1):271–288, 2018.

S. Gandhi, N. Madhusudhan, G. Hawker, and A. Piette. HyDRA-H: Simultaneous Hybrid Retrieval of Exoplanetary Emission Spectra. *AJ*, 158(6):228, Dec. 2019. doi: 10.3847/1538-3881/ab4efc.

R. Gansch and A. Adee. System Theoretic View on Uncertainties. *arXiv e-prints*, art. arXiv:2303.04042, Mar. 2023. doi: 10.48550/arXiv.2303.04042.

T. D. Gebhard, J. Wildberger, M. Dax, A. Kofler, D. Angerhausen, S. P. Quanz, and B. Schölkopf. Flow matching for atmospheric retrieval of exoplanets: Where reliability meets adaptive noise levels. *A&A*, 693:A42, Jan. 2025. doi: 10.1051/0004-6361/202451861.

A. Gelman and D. B. Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7:457–472, Jan. 1992. doi: 10.1214/ss/1177011136.

D. A. Golimowski, S. K. Leggett, M. S. Marley, X. Fan, T. R. Geballe, G. R. Knapp, F. J. Vrba, A. A. Henden, C. B. Luginbuhl, H. H. Guetter, J. A. Munn, B. Canzian, W. Zheng, Z. I. Tsvetanov, K. Chiu, K. Glazebrook, E. A. Hoversten, D. P. Schneider, and J. Brinkmann. L' and M' Photometry of Ultracool Dwarfs. *AJ*, 127(6):3516–3536, June 2004. doi: 10.1086/420709.

D. González Picos, I. A. G. Snellen, S. de Regt, R. Landman, Y. Zhang, S. Gandhi, C. Ginski, A. Y. Kesseli, P. Mollière, and T. Stolker. The ESO SupJup Survey II: The $^{12}C/^{13}C$ ratios of three young brown dwarfs with CRIRES^{+}. *arXiv e-prints*, art. arXiv:2407.07678, July 2024. doi: 10.48550/arXiv.2407.07678.

T. Guillot. On the radiative equilibrium of irradiated planetary atmospheres. *A&A*, 520:A27, Sept. 2010. doi: 10.1051/0004-6361/200913396.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

C. Helling and P. Woitke. Dust in brown dwarfs-v. growth and evaporation of dirty dust grains. *Astronomy & Astrophysics*, 455(1):325–338, 2006.

C. Helling, N. Iro, L. Corrales, D. Samra, K. Ohno, M. K. Alam, M. Steinrueck, B. Lew, K. Molaverdikhani, R. J. MacDonald, O. Herbort, P. Woitke, and V. Parmentier. Understanding the atmospheric properties and chemical composition of the ultra-hot Jupiter HAT-P-7b. I. Cloud and chemistry mapping. *A&A*, 631:A79, Nov. 2019. doi: 10.1051/0004-6361/201935771.

J. Hermans, V. Begy, and G. Louppe. Likelihood-free mcmc with amortized approximate ratio estimators. In *International Conference on Machine Learning*, pages 4239–4248. PMLR, 2020.

N. R. Hinkel, F. X. Timmes, P. A. Young, M. D. Pagano, and M. C. Turnbull. Stellar Abundances in the Solar Neighborhood: The Hypatia Catalog. *AJ*, 148(3):54, Sept. 2014. doi: 10.1088/0004-6256/148/3/54.

M. Holmberg and N. Madhusudhan. A First Look at CRIRES+: Performance Assessment and Exoplanet Spectroscopy. *AJ*, 164(3):79, Sept. 2022. doi: 10.3847/1538-3881/ac77eb.

C.-W. Huang, D. Krueger, A. Lacoste, and A. Courville. Neural autoregressive flows. In *International Conference on Machine Learning*, pages 2078–2087. PMLR, 2018.

I. Hubeny. Model atmospheres of sub-stellar mass objects. *MNRAS*, 469(1):841–869, July 2017. doi: 10.1093/mnras/stx758.

I. Hubeny and T. Lanz. Non-LTE Line-blanketed Model Atmospheres of Hot Stars. I. Hybrid Complete Linearization/Accelerated Lambda Iteration Method. *ApJ*, 439:875, Feb. 1995. doi: 10.1086/175226.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An Introduction to Variational Methods for Graphical Models. *Mach. Learn.*, 37(2):183–233, 1999.

H.-U. Kaeufl, P. Ballester, P. Biereichel, B. Delabre, R. Donaldson, R. Dorn, E. Fedrigo, G. Finger, G. Fischer, F. Franza, D. Gojak, G. Huster, Y. Jung, J.-L. Lizon, L. Mehrgan, M. Meyer, A. Moorwood, J.-F. Pirard, J. Paufique, E. Pozna, R. Siebenmorgen, A. Silber, J. Stegmeier, and S. Wegerer. CRIRES: a high-resolution infrared spectrograph for ESO's VLT. In A. F. M. Moorwood and M. Iye, editors, *Ground-based Instrumentation for Astronomy*, volume 5492 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 1218–1227, Sept. 2004. doi: 10.1117/12.551480.

T. Karalidi, M. Marley, J. J. Fortney, C. Morley, D. Saumon, R. Lupu, C. Visscher, and R. Freedman. The Sonora Substellar Atmosphere Models. II. Cholla: A Grid of Cloud-free, Solar Metallicity Models in Chemical Disequilibrium for the JWST Era. *ApJ*, 923(2):269, Dec. 2021. doi: 10.3847/1538-4357/ac3140.

A. Y. Kesseli and I. A. G. Snellen. Confirmation of Asymmetric Iron Absorption in WASP-76b with HARPS. *ApJ*, 908(1):L17, Feb. 2021. doi: 10.3847/2041-8213/abe047.

N. Khorshid, M. Min, and J. M. Désert. Retrieving planet formation scenarios from the observations of hot Jupiters. In *44th COSPAR Scientific Assembly. Held 16-24 July*, volume 44, page 268, July 2022.

J. Kim and V. Rockova. Deep Bayes Factors. *arXiv e-prints*, art. arXiv:2312.05411, Dec. 2023. doi: 10.48550/arXiv.2312.05411.

R. A. Kimble, J. W. MacKenty, R. W. O'Connell, and J. A. Townsend. Wide Field Camera 3: a powerful new imager for the Hubble Space Telescope. In J. M. Oschmann, Jr., M. W. M. de Graauw, and H. A. MacEwen, editors, *Space Telescopes and Instrumentation 2008: Optical, Infrared, and Millimeter*, volume 7010, page 70101E. International Society for Optics and Photonics, SPIE, 2008. doi: 10.1117/12.789581. URL `https://doi.org/10.1117/12.789581`.

J. D. Kirkpatrick. New spectral types l and t. *Annu. Rev. Astron. Astrophys.*, 43:195–245, 2005.

J. D. Kirkpatrick, I. N. Reid, J. Liebert, R. M. Cutri, B. Nelson, C. A. Beichman, C. C. Dahn, D. G. Monet, J. E. Gizis, and M. F. Skrutskie. Dwarfs Cooler than "M": The Definition of Spectral Type "L" Using Discoveries from the 2 Micron All-Sky Survey (2MASS). *ApJ*, 519(2):802–833, July 1999. doi: 10.1086/307414.

J. D. Kirkpatrick, M. C. Cushing, C. R. Gelino, R. L. Griffith, M. F. Skrutskie, K. A. Marsh, E. L. Wright, A. Mainzer, P. R. Eisenhardt, I. S. McLean, et al. The first hundred brown dwarfs discovered by the wide-field infrared survey explorer (wise). *ApJS*, 197(2):19, 2011.

J. D. Kirkpatrick, C. R. Gelino, M. C. Cushing, G. N. Mace, R. L. Griffith, M. F. Skrutskie, K. A. Marsh, E. L. Wright, P. R. Eisenhardt, I. S. McLean, A. K. Mainzer, A. J. Burgasser, C. G. Tinney, S. Parker, and G. Salter. Further Defining Spectral Type "Y" and Exploring the Low-mass End of the Field Brown Dwarf Mass Function. *ApJ*, 753(2):156, July 2012. doi: 10.1088/0004-637X/753/2/156.

J. D. Kirkpatrick, F. Abdurrahman, W. M. Best, T. J. Dupuy, J. K. Faherty, C. B. Henderson, F. Marocco, P. Mroz, J. Sahlmann, R. L. Smart, C. A. Theissen, and E. L. Wright. The Need for Infrared Astrometry of Brown Dwarfs in the Post-Gaia Era. *BAAS*, 51(3):105, May 2019a.

J. D. Kirkpatrick, E. C. Martin, R. L. Smart, A. J. Cayago, C. A. Beichman, F. Marocco, C. R. Gelino, J. K. Faherty, M. C. Cushing, A. C. Schneider, G. N. Mace, C. G. Tinney, E. L. Wright, P. J. Lowrance, J. G. Ingalls, F. J. Vrba, J. A. Munn, S. E. Dahm, and I. S. McLean. Preliminary Trigonometric Parallaxes of 184 Late-T and Y Dwarfs and an Analysis of the Field Substellar Mass Function into the "Planetary" Mass Regime. *ApJS*, 240(2):19, Feb. 2019b. doi: 10.3847/1538-4365/aaf6af.

H. Kühnle, P. Patapis, P. Mollière, P. Tremblin, E. Matthews, A. M. Glauser, N. Whiteford, M. Vasist, O. Absil, D. Barrado, M. Min, P. O. Lagage, L. B. F. M. Waters, M. Guedel, T. Henning, B. Vandenbussche, P. Baudoz, L. Decin, J. P. Pye, P. Royer, E. F. van Dishoeck, G. Östlin, T. P. Ray, and G. Wright. Water depletion and 15NH3 in the atmosphere of the coldest brown dwarf observed with JWST/MIRI. *arXiv e-prints*, art. arXiv:2410.10933, Oct. 2024. doi: 10.48550/arXiv.2410.10933.

H. Kühnle, P. Patapis, P. Mollière, P. Tremblin, E. Matthews, A. Glauser, N. Whiteford, M. Vasist, O. Absil, D. Barrado, M. Min, P. Lagage, L. Waters, M. Guedel, T. Henning, B. Vandenbussche, P. Baudoz, L. Decin, J. Pye, and G. Wright. Water depletion and 15nh3 in the atmosphere of the coldest brown dwarf observed with jwst/miri. In Astronomy & Astrophysics, 2024.

B. Lacy and A. Burrows. Self-consistent Models of Y Dwarf Atmospheres with Water Clouds and Disequilibrium Chemistry. *ApJ*, 950(1):8, June 2023. doi: 10.3847/1538-4357/acc8cb.

J. U. Lange. NAUTILUS: boosting Bayesian importance nested sampling with deep learning. *MNRAS*, 525(2):3181–3194, Oct. 2023. doi: 10.1093/mnras/stad2441.

N. Latouf, A. M. Mandell, G. L. Villanueva, M. D. Himes, M. D. Moore, N. Susemiehl, J. Crouse, S. Domagal-Goldman, G. Arney, V. Kofman, and A. V. Young. Bayesian analysis for remote biosignature identification on exoearths (barbie). ii. using grid-based nested sampling in coronagraphy observation simulations for o2 and o3. *The Astronomical Journal*, 167(1):27, dec 2023. doi: 10.3847/1538-3881/ad0fde. URL `https://dx.doi.org/10.3847/1538-3881/ad0fde`.

B. Lavie, J. M. Mendonça, C. Mordasini, M. Malik, M. Bonnefoy, B.-O. Demory, M. Oreshenko, S. L. Grimm, D. Ehrenreich, and K. Heng. Helios–retrieval: an open-source, nested sampling atmospheric retrieval code; application to the hr 8799 exoplanets and inferred constraints for planet formation. *The Astronomical Journal*, 154(3):91, 2017.

B. Lavie, J. M. Mendonça, C. Mordasini, M. Malik, M. Bonnefoy, B.-O. Demory, M. Oreshenko, S. L. Grimm, D. Ehrenreich, and K. Heng. HELIOS-RETRIEVAL: An Open-source, Nested Sampling Atmospheric Retrieval Code; Application to the HR 8799 Exoplanets and Inferred Constraints for Planet Formation. *AJ*, 154(3):91, Sept. 2017. doi: 10.3847/1538-3881/aa7ed8.

E. K. H. Lee, J. Blecic, and C. Helling. Dust in brown dwarfs and extra-solar planets. *Astronomy & Astrophysics*, 614:A126, jun 2018. doi: 10.1051/0004-6361/201731977. URL `https://doi.org/10.1051%2F0004-6361%2F201731977`.

S. Leggett, C. V. Morley, M. Marley, and D. Saumon. Near-infrared photometry of y dwarfs: Low ammonia abundance and the onset of water clouds. *ApJ*, 799(1):37, 2015.

S. Leggett, M. C. Cushing, K. K. Hardegree-Ullman, J. L. Trucks, M. Marley, C. V. Morley, D. Saumon, S. Carey, J. Fortney, C. Gelino, et al. Observed variability at 1 and 4 μm in the y0 brown dwarf wisep j173835. 52+ 273258.9. *ApJ*, 830(2):141, 2016a.

S. Leggett, P. Tremblin, T. Esplin, K. Luhman, and C. V. Morley. The y-type brown dwarfs: estimates of mass and age from new astrometry, homogenized photometry, and near-infrared spectroscopy. *ApJ*, 842(2):118, 2017.

S. K. Leggett, D. Saumon, M. S. Marley, T. R. Geballe, D. A. Golimowski, D. Stephens, and X. Fan. 3.6-7.9 μm Photometry of L and T Dwarfs and the Prevalence of Vertical Mixing in their Atmospheres. *ApJ*, 655(2):1079–1094, Feb. 2007. doi: 10.1086/510014.

S. K. Leggett, C. V. Morley, M. S. Marley, D. Saumon, J. J. Fortney, and C. Visscher. A Comparison of Near-infrared Photometry and Spectra for Y Dwarfs with a New Generation of Cool Cloudy Models. *ApJ*, 763(2):130, Feb. 2013. doi: 10.1088/0004-637X/763/2/130.

S. K. Leggett, P. Tremblin, D. Saumon, M. S. Marley, C. V. Morley, D. S. Amundsen, I. Baraffe, and G. Chabrier. Near-infrared spectroscopy of the y0 wisep j173835. 52+ 273258.9 and the y1 wise j035000. 32–565830.2: the importance of non-equilibrium chemistry. *ApJ*, 824(1):2, 2016b.

E. Lei and P. Mollière. easyCHEM: A Python package for calculating chemical equilibrium abundances in exoplanet atmospheres. *arXiv e-prints*, art. arXiv:2410.21364, Oct. 2024. doi: 10.48550/arXiv.2410.21364.

B. W. P. Lew, T. Roellig, N. E. Batalha, M. Line, T. Greene, S. Murkherjee, R. Freedman, M. Meyer, C. Beichman, C. Alves de Oliveira, M. De Furio, D. Johnstone, A. Z. Greenbaum, M. Marley, J. J. Fortney, E. T. Young, J. Leisenring, M. Boyer, K. Hodapp, K. Misselt, J. Stansberry, and M. Rieke. High-precision Atmospheric Characterization of a Y Dwarf with JWST NIRSpec G395H Spectroscopy: Isotopologue, C/O Ratio, Metallicity, and the Abundances of Six Molecular Species. *AJ*, 167(5):237, May 2024. doi: 10.3847/1538-3881/ad3425.

Z. Li and J. Cao. General P-Splines for Non-Uniform B-Splines. *arXiv e-prints*, art. arXiv:2201.06808, Jan. 2022. doi: 10.48550/arXiv.2201.06808.

M. R. Line, A. S. Wolf, X. Zhang, H. Knutson, J. A. Kammer, E. Ellison, P. Deroo, D. Crisp, and Y. L. Yung. A systematic retrieval analysis of secondary eclipse spectra. i. a comparison of atmospheric retrieval techniques. *The Astrophysical Journal*, 775(2):137, 2013.

M. R. Line, H. Knutson, A. S. Wolf, and Y. L. Yung. A systematic retrieval analysis of secondary eclipse spectra. ii. a uniform analysis of nine planets and their c to o ratios. *The Astrophysical Journal*, 783(2):70, 2014.

M. R. Line, J. Teske, B. Burningham, J. J. Fortney, and M. S. Marley. Uniform atmospheric retrieval analysis of ultracool dwarfs. i. characterizing benchmarks, gl 570d and hd 3651b. *ApJ*, 807(2):183, jul 2015. doi: 10.1088/0004-637X/807/2/183. URL `https://dx.doi.org/10.1088/0004-637X/807/2/183`.

M. R. Line, M. S. Marley, M. C. Liu, B. Burningham, C. V. Morley, N. R. Hinkel, J. Teske, J. J. Fortney, R. Freedman, and R. Lupu. Uniform atmospheric retrieval analysis of ultracool dwarfs. ii. properties of 11 t dwarfs. *ApJ*, 848(2):83, oct 2017. doi: 10.3847/1538-4357/aa7ff0. URL `https://dx.doi.org/10.3847/1538-4357/aa7ff0`.

J. Linhart, A. Gramfort, and P. Rodrigues. L-c2st: Local diagnostics for posterior approximations in simulation-based inference. *Adv. Neural Inf. Process. Syst.*, 36, 2024.

K. Lodders and B. Fegley, Jr. Chemistry of Low Mass Substellar Objects. In J. W. Mason, editor, *Astrophysics Update 2*, page 1. Springer, 2006. doi: 10.1007/3-540-30313-8_1.

K. Lodders and B. Fegley Jr. Chemistry of low mass substellar objects. In *Astrophysics Update 2*, pages 1–28. Springer, 2006.

I. Loshchilov and F. Hutter. Decoupled Weight Decay Regularization. *arXiv e-prints*, art. arXiv:1711.05101, Nov. 2017. doi: 10.48550/arXiv.1711.05101.

A. Lueber, D. Kitzmann, B. P. Bowler, A. J. Burgasser, and K. Heng. Retrieval Study of Brown Dwarfs across the L-T Sequence. *ApJ*, 930(2):136, May 2022. doi: 10.3847/1538-4357/ac63b9.

A. Lueber, D. Kitzmann, C. E. Fisher, B. P. Bowler, A. J. Burgasser, M. Marley, and K. Heng. Intercomparison of Brown Dwarf Model Grids and Atmospheric Retrieval Using Machine Learning. *ApJ*, 954(1):22, Sept. 2023. doi: 10.3847/1538-4357/ace530.

A. Lueber, K. Karchev, C. Fisher, M. Heim, R. Trotta, and K. Heng. Near-instantaneous Atmospheric Retrievals and Model Comparison with FASTER. *arXiv e-prints*, art. arXiv:2502.18045, Feb. 2025. doi: 10.48550/arXiv.2502.18045.

R. J. MacDonald and N. Madhusudhan. Hd 209458b in new light: evidence of nitrogen chemistry, patchy clouds and sub-solar water. *Monthly Notices of the Royal Astronomical Society*, 469(2):1979–1996, 2017.

N. Madhusudhan. C/O Ratio as a Dimension for Characterizing Exoplanetary Atmospheres. *ApJ*, 758(1):36, Oct. 2012. doi: 10.1088/0004-637X/758/1/36.

N. Madhusudhan. *Atmospheric Retrieval of Exoplanets*, pages 1–30. Springer International Publishing, Cham, 2018. ISBN 978-3-319-30648-3. doi: 10.1007/978-3-319-30648-3_104-1. URL https://doi.org/10.1007/978-3-319-30648-3_104-1.

N. Madhusudhan. Exoplanetary Atmospheres: Key Insights, Challenges, and Prospects. *ARA&A*, 57:617–663, Aug. 2019. doi: 10.1146/annurev-astro-081817-051846.

N. Madhusudhan, A. Burrows, and T. Currie. Model Atmospheres for Massive Gas Giants with Thick Clouds: Application to the HR 8799 Planets and Predictions for Future Detections. *ApJ*, 737(1):34, Aug. 2011. doi: 10.1088/0004-637X/737/1/34.

N. Madhusudhan, J. Harrington, K. B. Stevenson, S. Nymeyer, C. J. Campo, P. J. Wheatley, D. Deming, J. Blecic, R. A. Hardy, N. B. Lust, et al. A high c/o ratio and weak thermal inversion in the atmosphere of exoplanet wasp-12b. *Nature*, 469(7328):64–67, 2011.

N. Madhusudhan, N. Crouzet, P. R. McCullough, D. Deming, and C. Hedges. H2o abundances in the atmospheres of three hot jupiters. *The Astrophysical Journal Letters*, 791(1):L9, 2014.

N. Madhusudhan, M. Agúndez, J. I. Moses, and Y. Hu. Exoplanetary Atmospheres—Chemistry, Formation Conditions, and Habitability. *Space Sci. Rev.*, 205(1-4):285–348, Dec. 2016. doi: 10.1007/s11214-016-0254-3.

J. Mang, P. Gao, C. E. Hood, J. J. Fortney, N. Batalha, X. Yu, and I. de Pater. Microphysics of Water Clouds in the Atmospheres of Y Dwarfs and Temperate Giant Planets. *ApJ*, 927(2):184, Mar. 2022. doi: 10.3847/1538-4357/ac51d3.

M. S. Marley and T. D. Robinson. On the cool side: modeling the atmospheres of brown dwarfs and giant planets. *Annual Review of Astronomy and Astrophysics*, 53:279–323, 2015.

M. S. Marley, D. Saumon, T. Guillot, R. S. Freedman, W. B. Hubbard, A. Burrows, and J. I. Lunine. Atmospheric, Evolutionary, and Spectral Models of the Brown Dwarf Gliese 229 B. *Science*, 272(5270):1919–1921, June 1996. doi: 10.1126/science.272.5270.1919.

M. S. Marley, D. Saumon, and C. Goldblatt. A patchy cloud model for the l to t dwarf transition. *ApJ*, 723(1):L117, oct 2010. doi: 10.1088/2041-8205/723/1/L117. URL `https://dx.doi.org/10.1088/2041-8205/723/1/L117`.

M. S. Marley, D. Saumon, M. Cushing, A. S. Ackerman, J. J. Fortney, and R. Freedman. Masses, radii, and cloud properties of the hr 8799 planets. *The Astrophysical Journal*, 754(2):135, jul 2012. doi: 10.1088/0004-637X/754/2/135. URL `https://dx.doi.org/10.1088/0004-637X/754/2/135`.

M. S. Marley, D. Saumon, C. Visscher, R. Lupu, R. Freedman, C. Morley, J. J. Fortney, C. Seay, A. J. R. W. Smith, D. J. Teal, and R. Wang. The Sonora Brown Dwarf Atmosphere and Evolution Models. I. Model Description and Application to Cloudless Atmospheres in Rainout Chemical Equilibrium. *ApJ*, 920(2):85, Oct. 2021. doi: 10.3847/1538-4357/ac141d.

C. Marois, B. Macintosh, T. Barman, B. Zuckerman, I. Song, J. Patience, D. Lafrenière, and R. Doyon. Direct Imaging of Multiple Planets Orbiting the Star HR 8799. *Science*, 322(5906): 1348, Nov. 2008. doi: 10.1126/science.1166585.

C. Marois, B. Zuckerman, Q. M. Konopacky, B. Macintosh, and T. Barman. Images of a fourth planet orbiting HR 8799. *Nature*, 468(7327):1080–1083, Dec. 2010. doi: 10.1038/nature09684.

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

E. C. Martin, J. D. Kirkpatrick, C. A. Beichman, R. L. Smart, J. K. Faherty, C. R. Gelino, M. C. Cushing, A. C. Schneider, E. L. Wright, P. Lowrance, J. Ingalls, C. G. Tinney, I. S. McLean, S. E. Logsdon, and J. Lebreton. Y Dwarf Trigonometric Parallaxes from the Spitzer Space Telescope. *ApJ*, 867(2):109, Nov. 2018. doi: 10.3847/1538-4357/aae1af.

E. C. Martin, J. D. Kirkpatrick, C. A. Beichman, R. L. Smart, J. K. Faherty, C. R. Gelino, M. C. Cushing, A. C. Schneider, E. L. Wright, P. Lowrance, J. Ingalls, C. G. Tinney, I. S. McLean, S. E. Logsdon, and J. Lebreton. Y dwarf trigonometric parallaxes from the spitzer space telescope. *ApJ*, 867(2):109, nov 2018. doi: 10.3847/1538-4357/aae1af. URL `https://dx.doi.org/10.3847/1538-4357/aae1af`.

E. C. Matthews, A. L. Carter, P. Pathak, C. V. Morley, M. W. Phillips, S. K. P. M., F. Feng, M. J. Bonse, L. A. Boogaard, J. A. Burt, I. J. M. Crossfield, E. S. Douglas, T. Henning, J. Hom, C. L. Ko, M. Kasper, A. M. Lagrange, D. Petit dit de la Roche, and F. Philipot. A temperate super-Jupiter imaged with JWST in the mid-infrared. *Nature*, 633(8031):789–792, Sept. 2024. doi: 10.1038/s41586-024-07837-8.

E. C. Matthews, P. Mollière, H. Kühnle, P. Patapis, N. Whiteford, M. Samland, P.-O. Lagage, R. Waters, S.-M. Tsai, K. Zahnle, M. Guedel, T. Henning, B. Vandenbussche, O. Absil, I. Argyriou, D. Barrado, A. Coulais, A. M. Glauser, G. Olofsson, J. P. Pye, D. Rouan, P. Royer, E. F. van Dishoeck, T. P. Ray, and G. Östlin. HCN and C2H2 in the atmosphere of a T8.5+T9 brown dwarf binary. *arXiv e-prints*, art. arXiv:2502.13610, Feb. 2025. doi: 10.48550/arXiv.2502.13610.

M. Mayor and D. Queloz. A jupiter-mass companion to a solar-type star. *Nature*, 378(6555): 355–359, November 1995. ISSN 1476-4687. doi: 10.1038/378355a0. URL `https://doi.org/10.1038/378355a0`.

P. Mollière and I. A. G. Snellen. Detecting isotopologues in exoplanet atmospheres using ground-based high-dispersion spectroscopy. *A&A*, 622:A139, Feb. 2019. doi: 10.1051/0004-6361/201834169.

P. Mollière, R. van Boekel, C. Dullemond, T. Henning, and C. Mordasini. Model atmospheres of irradiated exoplanets: the influence of stellar parameters, metallicity, and the c/o ratio. *The Astrophysical Journal*, 813(1):47, 2015.

P. Mollière, R. van Boekel, J. Bouwman, T. Henning, P.-O. Lagage, and M. Min. Observing transiting planets with jwst-prime targets and their synthetic spectral observations. *Astronomy & Astrophysics*, 600:A10, 2017.

P. Mollière, J. Wardenier, R. van Boekel, T. Henning, K. Molaverdikhani, and I. Snellen. petitradtrans-a python radiative transfer package for exoplanet characterization and retrieval. *Astronomy & Astrophysics*, 627:A67, 2019.

P. Mollière, T. Stolker, S. Lacour, G. P. P. L. Otten, J. Shangguan, B. Charnay, T. Molyarova, M. Nowak, T. Henning, G.-D. Marleau, D. A. Semenov, E. van Dishoeck, F. Eisenhauer, P. Garcia, R. G. Lopez, J. H. Girard, A. Z. Greenbaum, S. Hinkley, P. Kervella, L. Kreidberg, A.-L. Maire, E. Nasedkin, L. Pueyo, I. A. G. Snellen, A. Vigan, J. Wang, P. T. de Zeeuw, and A. Zurlo. Retrieving scattering clouds and disequilibrium chemistry in the atmosphere of HR 8799e. *Astronomy & Astrophysics*, 640:A131, aug 2020. doi: 10.1051/0004-6361/202038325. URL `https://doi.org/10.1051%2F0004-6361%2F202038325`.

C. Mordasini, R. van Boekel, P. Mollière, T. Henning, and B. Benneke. The Imprint of Exoplanet Formation History on Observable Present-day Spectra of Hot Jupiters. *ApJ*, 832(1):41, Nov. 2016. doi: 10.3847/0004-637X/832/1/41.

C. V. Morley, J. J. Fortney, M. S. Marley, C. Visscher, D. Saumon, and S. K. Leggett. Neglected clouds in t and y dwarf atmospheres. *ApJ*, 756(2):172, aug 2012. doi: 10.1088/0004-637X/756/2/172. URL https://dx.doi.org/10.1088/0004-637X/756/2/172.

C. V. Morley, M. S. Marley, J. J. Fortney, R. Lupu, D. Saumon, T. Greene, and K. Lodders. Water clouds in y dwarfs and exoplanets. *ApJ*, 787(1):78, 2014.

C. V. Morley, A. J. Skemer, K. N. Allers, M. S. Marley, J. K. Faherty, C. Visscher, S. A. Beiler, B. E. Miles, R. Lupu, R. S. Freedman, J. J. Fortney, T. R. Geballe, and G. L. Bjoraker. An l band spectrum of the coldest brown dwarf. *ApJ*, 858(2):97, may 2018. doi: 10.3847/1538-4357/aabe8b. URL https://dx.doi.org/10.3847/1538-4357/aabe8b.

S. Mukherjee, J. J. Fortney, N. E. Batalha, T. Karalidi, M. S. Marley, C. Visscher, B. E. Miles, and A. J. I. Skemer. Probing the Extent of Vertical Mixing in Brown Dwarf Atmospheres with Disequilibrium Chemistry. *ApJ*, 938(2):107, Oct. 2022. doi: 10.3847/1538-4357/ac8dfb.

S. Mukherjee, J. J. Fortney, C. V. Morley, N. E. Batalha, M. S. Marley, T. Karalidi, C. Visscher, R. Lupu, R. Freedman, and E. Gharib-Nezhad. The Sonora Substellar Atmosphere Models. IV. Elf Owl: Atmospheric Mixing and Chemical Disequilibrium with Varying Metallicity and C/O Ratios. *ApJ*, 963(1):73, Mar. 2024a. doi: 10.3847/1538-4357/ad18c2.

S. Mukherjee, J. J. Fortney, N. F. Wogan, D. K. Sing, and K. Ohno. Effects of Planetary Parameters on Disequilibrium Chemistry in Irradiated Planetary Atmospheres: From Gas Giants to Sub-Neptunes. *arXiv e-prints*, art. arXiv:2410.17169, Oct. 2024b. doi: 10.48550/arXiv.2410.17169.

E. Nasedkin, P. Mollière, J. Wang, F. Cantalloube, L. Kreidberg, L. Pueyo, T. Stolker, and A. Vigan. Impacts of high-contrast image processing on atmospheric retrievals. *A&A*, 678:A41, Oct. 2023. doi: 10.1051/0004-6361/202346585.

M. C. Nixon, L. Welbanks, P. McGill, and E. M. R. Kempton. Methods for Incorporating Model Uncertainty into Exoplanet Atmospheric Analysis. *ApJ*, 966(2):156, May 2024. doi: 10.3847/1538-4357/ad354e.

K. S. Noll, T. R. Geballe, and M. S. Marley. Detection of Abundant Carbon Monoxide in the Brown Dwarf Gliese 229B. *ApJ*, 489(1):L87–L90, Nov. 1997. doi: 10.1086/310954.

S. Notsu, K. Ohno, T. Ueda, C. Walsh, C. Eistrup, and H. Nomura. The Molecular Composition of Shadowed Proto-solar Disk Midplanes Beyond the Water Snowline. *ApJ*, 936(2):188, Sept. 2022. doi: 10.3847/1538-4357/ac87fa.

A. Novais, C. Fisher, L. Ghezzi, D. Kitzmann, B. Thorsbro, and K. Heng. Parameter degeneracies associated with interpreting HST WFC3 transmission spectra of exoplanetary atmospheres. *MNRAS*, 538(4):2521–2547, Apr. 2025. doi: 10.1093/mnras/staf397.

K. I. Öberg, R. Murray-Clay, and E. A. Bergin. The Effects of Snowlines on C/O in Planetary Atmospheres. *ApJ*, 743(1):L16, Dec. 2011. doi: 10.1088/2041-8205/743/1/L16.

N. Oberg, I. Kamp, and S. Cazaux. The Chemical Inheritance of Icy Moons. In *European Planetary Science Congress*, pages EPSC2021–608, Sept. 2021. doi: 10.5194/epsc2021-608.

M. Oreshenko, B. Lavie, S. L. Grimm, S.-M. Tsai, M. Malik, B.-O. Demory, C. Mordasini, Y. Alibert, W. Benz, S. P. Quanz, et al. Retrieval analysis of the emission spectrum of wasp-12b: sensitivity of outcomes to prior assumptions and implications for formation history. *The Astrophysical Journal Letters*, 847(1):L3, 2017.

G. Papamakarios and I. Murray. Fast ε-free inference of simulation models with bayesian conditional density estimation. *Advances in neural information processing systems*, 29, 2016.

G. Papamakarios, T. Pavlakou, and I. Murray. Masked Autoregressive Flow for Density Estimation. *arXiv e-prints*, art. arXiv:1705.07057, May 2017. doi: 10.48550/arXiv.1705.07057.

G. Papamakarios, D. Sterratt, and I. Murray. Sequential neural likelihood: Fast likelihood-free inference with autoregressive flows. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 837–848. PMLR, 2019.

G. Papamakarios, E. T. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(57): 1–64, 2021.

I. Pascucci, G. Herczeg, J. S. Carr, and S. Bruderer. The Atomic and Molecular Content of Disks around Very Low-mass Stars and Brown Dwarfs. *ApJ*, 779(2):178, Dec. 2013. doi: 10.1088/0004-637X/779/2/178.

C. L. Phillips, J. K. Faherty, B. Burningham, J. M. Vos, E. C. Gonzales, E. J. Griffith, S. Alejandro Merchan, E. Calamari, C. Visscher, C. V. Morley, N. Whiteford, J. Gaarn, I. Ilyin, K. Strassmeier, and J. Wang. Retrieving young cloudy l dwarfs: A nearby planetary-mass companion bd+60 1417b and its isolated red twin w0047. *The Astrophysical Journal*, 972 (2):172, sep 2024. doi: 10.3847/1538-4357/ad5d57. URL `https://dx.doi.org/10.3847/1538-4357/ad5d57`.

M. W. Phillips, P. Tremblin, I. Baraffe, G. Chabrier, N. F. Allard, F. Spiegelman, J. M. Goyal, B. Drummond, and E. Hébrard. A new set of atmosphere and evolution models for cool T-Y brown dwarfs and giant exoplanets. *A&A*, 637:A38, May 2020. doi: 10.1051/0004-6361/201937381.

J. B. Pollack, O. Hubickyj, P. Bodenheimer, J. J. Lissauer, M. Podolak, and Y. Greenzweig. Formation of the Giant Planets by Concurrent Accretion of Solids and Gas. *Icarus*, 124(1): 62–85, Nov. 1996. doi: 10.1006/icar.1996.0190.

O. L. Polyansky, A. A. Kyuberis, N. F. Zobov, J. Tennyson, S. N. Yurchenko, and L. Lodi. Exomol molecular line lists xxx: a complete high-accuracy line list for water. *Monthly Notices of the Royal Astronomical Society*, 480(2):2597–2608, 08 2018. ISSN 0035-8711. doi: 10.1093/mnras/sty1877. URL https://doi.org/10.1093/mnras/sty1877.

A. M. Ritchey, S. R. Federman, and D. L. Lambert. The $C^{14}N/C^{15}N$ Ratio in Diffuse Molecular Clouds. *ApJ*, 804(1):L3, May 2015. doi: 10.1088/2041-8205/804/1/L3.

T. L. Roellig, J. E. Van Cleve, G. C. Sloan, J. C. Wilson, D. Saumon, S. K. Leggett, M. S. Marley, M. C. Cushing, J. D. Kirkpatrick, A. K. Mainzer, and J. R. Houck. Spitzer Infrared Spectrograph (IRS) Observations of M, L, and T Dwarfs. *ApJS*, 154(1):418–421, Sept. 2004. doi: 10.1086/421978.

L. Rothman, I. Gordon, R. Barber, H. Dothe, R. Gamache, A. Goldman, V. Perevalov, S. Tashkun, and J. Tennyson. Hitemp, the high-temperature molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 111(15):2139–2150, 2010. ISSN 0022-4073. doi: https://doi.org/10.1016/j.jqsrt.2010.05.001. URL https://www.sciencedirect.com/science/article/pii/S002240731000169X. XVIth Symposium on High Resolution Molecular Spectroscopy (HighRus-2009).

L. S. Rothman, I. E. Gordon, R. J. Barber, H. Dothe, R. R. Gamache, A. Goldman, V. I. Perevalov, S. A. Tashkun, and J. Tennyson. HITEMP, the high-temperature molecular spectroscopic database. *J. Quant. Spectr. Rad. Transf.*, 111:2139–2150, Oct. 2010. doi: 10.1016/j.jqsrt.2010.05.001.

M. J. Rowland, C. V. Morley, and M. R. Line. Toward Robust Atmospheric Retrieval on Cloudy L Dwarfs: the Impact of Thermal and Abundance Profile Assumptions. *ApJ*, 947(1):6, Apr. 2023. doi: 10.3847/1538-4357/acbb07.

F. Rozet. Arbitrary marginal neural ratio estimation for likelihood-free inference. Master's thesis, Université de Liège, 2022. URL https://matheo.uliege.be/bitstream/2268.2/12993/6/report.pdf.

D. Saumon and M. S. Marley. The evolution of l and t dwarfs in color-magnitude diagrams. *ApJ*, 689(2):1327, 2008.

D. Saumon, T. R. Geballe, S. K. Leggett, M. S. Marley, R. S. Freedman, K. Lodders, B. Fegley, Jr., and S. K. Sengupta. Molecular Abundances in the Atmosphere of the T Dwarf GL 229B. *ApJ*, 541(1):374–389, Sept. 2000. doi: 10.1086/309410.

D. Saumon, M. S. Marley, M. C. Cushing, S. K. Leggett, T. L. Roellig, K. Lodders, and R. S. Freedman. Ammonia as a Tracer of Chemical Equilibrium in the T7.5 Dwarf Gliese 570D. *ApJ*, 647(1):552–557, Aug. 2006. doi: 10.1086/505419.

D. Saumon, M. S. Marley, M. Abel, L. Frommhold, and R. S. Freedman. New h2 collision-induced absorption and nh3 opacity and the spectra of the coolest brown dwarfs. *ApJ*, 750 (1):74, 2012.

A. C. Schneider, M. C. Cushing, J. D. Kirkpatrick, C. R. Gelino, G. N. Mace, E. L. Wright, P. R. Eisenhardt, M. Skrutskie, R. L. Griffith, and K. A. Marsh. Hubble space telescope spectroscopy of brown dwarfs discovered with the wide-field infrared survey explorer. *ApJ*, 804(2):92, 2015.

A. D. Schneider and B. Bitsch. How drifting and evaporating pebbles shape giant planets. I. Heavy element content and atmospheric C/O. *A&A*, 654:A71, Oct. 2021. doi: 10.1051/0004-6361/202039640.

H. Schwarz, M. Brogi, R. de Kok, J. Birkby, and I. Snellen. Evidence against a strong thermal inversion in HD 209458b from high-dispersion spectroscopy. *A&A*, 576:A111, Apr. 2015. doi: 10.1051/0004-6361/201425170.

C. M. Sharp and A. Burrows. Atomic and Molecular Opacities for Brown Dwarf and Giant Planet Atmospheres. *ApJS*, 168(1):140–166, Jan. 2007. doi: 10.1086/508708.

S. A. Sisson, Y. Fan, and M. A. Beaumont. Overview of Approximate Bayesian Computation. *arXiv e-prints*, art. arXiv:1802.09720, Feb. 2018. doi: 10.48550/arXiv.1802.09720.

J. Skilling. Nested Sampling. In R. Fischer, R. Preuss, and U. V. Toussaint, editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735 of *American Institute of Physics Conference Series*, pages 395–405. AIP, Nov. 2004. doi: 10.1063/1.1835238.

I. Snellen, B. Brandl, R. de Kok, M. Brogi, J. Birkby, and H. Schwarz. The fast spin-rotation of a young extrasolar planet. *arXiv e-prints*, art. arXiv:1404.7506, Apr. 2014. doi: 10.48550/arXiv.1404.7506.

I. A. G. Snellen, R. J. de Kok, E. J. W. de Mooij, and S. Albrecht. The orbital motion, absolute mass and high-altitude winds of exoplanet HD209458b. *Nature*, 465(7301):1049–1051, June 2010. doi: 10.1038/nature09111.

F. Soboczenski, M. D. Himes, M. D. O'Beirne, S. Zorzan, A. Gunes Baydin, A. D. Cobb, Y. Gal, D. Angerhausen, M. Mascaro, G. N. Arney, and S. D. Domagal-Goldman. Bayesian Deep Learning for Exoplanet Atmospheric Retrieval. *arXiv e-prints*, art. arXiv:1811.03390, Nov. 2018. doi: 10.48550/arXiv.1811.03390.

K. Y. L. Su, G. H. Rieke, K. R. Stapelfeldt, R. Malhotra, G. Bryden, P. S. Smith, K. A. Misselt, A. Moro-Martin, and J. P. Williams. The Debris Disk Around HR 8799. *ApJ*, 705(1):314–327, Nov. 2009. doi: 10.1088/0004-637X/705/1/314.

D. Sudarsky, A. Burrows, I. Hubeny, and A. Li. Phase Functions and Light Curves of Wide-Separation Extrasolar Giant Planets. *ApJ*, 627(1):520–533, July 2005. doi: 10.1086/430206.

B. Tabone, G. Bettoni, E. F. van Dishoeck, A. M. Arabhavi, S. Grant, D. Gasman, T. Henning, I. Kamp, M. Güdel, P. O. Lagage, T. Ray, B. Vandenbussche, A. Abergel, O. Absil, I. Argyriou, D. Barrado, A. Boccaletti, J. Bouwman, A. Caratti o Garatti, V. Geers, A. M. Glauser, K. Justannont, F. Lahuis, M. Mueller, C. Nehmé, G. Olofsson, E. Pantin, S. Scheithauer, C. Waelkens, L. B. F. M. Waters, J. H. Black, V. Christiaens, R. Guadarrama, M. Morales-Calderón, H. Jang, J. Kanwar, N. Pawellek, G. Perotti, A. Perrin, D. Rodgers-Lee, M. Samland, J. Schreiber, K. Schwarz, L. Colina, G. Östlin, and G. Wright. A rich hydrocarbon chemistry and high C to O ratio in the inner disk around a very low-mass star. *Nature*, 7:805–814, July 2023. doi: 10.1038/s41550-023-01965-3.

S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman. Validating Bayesian Inference Algorithms with Simulation-Based Calibration. *arXiv e-prints*, art. arXiv:1804.06788, Apr. 2018. doi: 10.48550/arXiv.1804.06788.

M. Tian and K. Heng. Atmospheric Chemistry of Secondary and Hybrid Atmospheres of Super Earths and Sub-Neptunes. *ApJ*, 963(2):157, Mar. 2024. doi: 10.3847/1538-4357/ad217c.

K. O. Todorov, M. R. Line, J. E. Pineda, M. R. Meyer, S. P. Quanz, S. Hinkley, and J. J. Fortney. The water abundance of the directly imaged substellar companion κ and b retrieved from a near infrared spectrum. *The Astrophysical Journal*, 823(1):14, 2016.

S. Tokdar and R. Kass. Importance sampling: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:54 – 60, 01 2010. doi: 10.1002/wics.56.

P. Tremblin, D. S. Amundsen, P. Mourier, I. Baraffe, G. Chabrier, B. Drummond, D. Homeier, and O. Venot. Fingering Convection and Cloudless Models for Cool Brown Dwarf Atmospheres. *ApJ*, 804(1):L17, May 2015. doi: 10.1088/2041-8205/804/1/L17.

P. Tremblin, D. S. Amundsen, G. Chabrier, I. Baraffe, B. Drummond, S. Hinkley, P. Mourier, and O. Venot. Cloudless atmospheres for l/t dwarfs and extrasolar giant planets. *The Astrophysical Journal Letters*, 817(2):L19, jan 2016a. doi: 10.3847/2041-8205/817/2/L19. URL `https://dx.doi.org/10.3847/2041-8205/817/2/L19`.

P. Tremblin, D. S. Amundsen, G. Chabrier, I. Baraffe, B. Drummond, S. Hinkley, P. Mourier, and O. Venot. Cloudless atmospheres for l/t dwarfs and extrasolar giant planets. *The Astrophysical journal letters*, 817(2):L19, 2016b.

T. Tsuji, K. Ohnaka, W. Aoki, and T. Nakajima. Evolution of dusty photospheres through red to brown dwarfs: how dust forms in very low mass objects. *A&A*, 308:L29–L32, Apr. 1996.

D. Turrini, S. Molinari, S. Fonte, and E. Schisano. Exploring the compositional signatures of formation and migration on giant planets. In *43rd COSPAR Scientific Assembly. Held 28 January - 4 February*, volume 43, page 510, Jan. 2021a.

D. Turrini, E. Schisano, S. Fonte, S. Molinari, R. Politi, D. Fedele, O. Panić, M. Kama, Q. Changeat, and G. Tinetti. Tracing the Formation History of Giant Planets in Protoplanetary Disks with Carbon, Oxygen, Nitrogen, and Sulfur. *ApJ*, 909(1):40, Mar. 2021b. doi: 10.3847/1538-4357/abd6e5.

M. Vasist, F. Rozet, O. Absil, P. Mollière, E. Nasedkin, and G. Louppe. Neural posterior estimation for exoplanetary atmospheric retrieval. *A&A*, 672:A147, Apr. 2023. doi: 10.1051/0004-6361/202245263.

S. Vaughan and J. Birkby. Reflection spectroscopy: a pathway to the atmospheric characterization of Earth-like exoplanets. In *AAS/Division for Extreme Solar Systems Abstracts*, volume 56 of *AAS/Division for Extreme Solar Systems Abstracts*, page 609.04, Apr. 2024.

C. Visscher and J. I. Moses. Quenching of Carbon Monoxide and Methane in the Atmospheres of Cool Brown Dwarfs and Hot Jupiters. *ApJ*, 738(1):72, Sept. 2011. doi: 10.1088/0004-637X/738/1/72.

C. Visscher, K. Lodders, and J. Bruce Fegley. Atmospheric chemistry in giant planets, brown dwarfs, and low-mass dwarf stars. ii. sulfur and phosphorus. *ApJ*, 648(2):1181, sep 2006. doi: 10.1086/506245. URL `https://dx.doi.org/10.1086/506245`.

H. R. Wakeford, D. K. Sing, T. Kataria, D. Deming, N. Nikolov, E. D. Lopez, P. Tremblin, D. S. Amundsen, N. K. Lewis, A. M. Mandell, J. J. Fortney, H. Knutson, B. Benneke, and T. M. Evans. HAT-P-26b: A Neptune-mass exoplanet with a well-constrained heavy element abundance. *Science*, 356(6338):628–631, May 2017. doi: 10.1126/science.aah4668.

H. R. Wakeford, D. K. Sing, T. Kataria, D. Deming, N. Nikolov, E. D. Lopez, P. Tremblin, D. S. Amundsen, N. K. Lewis, A. M. Mandell, et al. Hat-p-26b: A neptune-mass exoplanet with a well-constrained heavy element abundance. *Science*, 356(6338):628–631, 2017.

I. P. Waldmann, M. Rocchetto, G. Tinetti, E. J. Barton, S. N. Yurchenko, and J. Tennyson. -rex. ii. retrieval of emission spectra. *The Astrophysical Journal*, 813(1):13, 2015a.

I. P. Waldmann, G. Tinetti, M. Rocchetto, E. J. Barton, S. N. Yurchenko, and J. Tennyson. Tau-rex i: A next generation retrieval code for exoplanetary atmospheres. *The Astrophysical Journal*, 802(2):107, 2015b.

J. J. Wang, J. R. Graham, R. Dawson, D. Fabrycky, R. J. De Rosa, L. Pueyo, Q. Konopacky, B. Macintosh, C. Marois, E. Chiang, S. M. Ammons, P. Arriaga, V. P. Bailey, T. Barman, J. Bulger, J. Chilcote, T. Cotten, R. Doyon, G. Duchêne, T. M. Esposito, M. P. Fitzgerald, K. B. Follette, B. L. Gerard, S. J. Goodsell, A. Z. Greenbaum, P. Hibon, L.-W. Hung, P. Ingraham,

P. Kalas, J. E. Larkin, J. Maire, F. Marchis, M. S. Marley, S. Metchev, M. A. Millar-Blanchaer, E. L. Nielsen, R. Oppenheimer, D. Palmer, J. Patience, M. Perrin, L. Poyneer, A. Rajan, J. Rameau, F. T. Rantakyrö, J.-B. Ruffio, D. Savransky, A. C. Schneider, A. Sivaramakrishnan, I. Song, R. Soummer, S. Thomas, J. K. Wallace, K. Ward-Duong, S. Wiktorowicz, and S. Wolff. Dynamical Constraints on the HR 8799 Planets with GPI. *AJ*, 156(5):192, Nov. 2018. doi: 10.3847/1538-3881/aae150.

L. Welbanks, M. C. Nixon, P. McGill, L. J. Tilke, L. S. Wiser, Y. Rotman, S. Mukherjee, A. Feinstein, M. R. Line, S. Seager, T. G. Beatty, D. Z. Seligman, V. Parmentier, and D. Sing. The Challenges of Detecting Gases in Exoplanet Atmospheres. *arXiv e-prints*, art. arXiv:2504.21788, Apr. 2025. doi: 10.48550/arXiv.2504.21788.

J. B. Wildberger, M. Dax, S. Buchholz, S. R. Green, J. Macke, and B. Schölkopf. Flow Matching for Scalable Simulation-Based Inference. In *Machine Learning for Astrophysics*, page 34, July 2023. doi: 10.48550/arXiv.2305.17161.

P. Woitke, C. Helling, and O. Gunn. Dust in brown dwarfs and extra-solar planets-vii. cloud formation in diffusive atmospheres. *Astronomy & Astrophysics*, 634:A23, 2020.

A. Wolszczan and D. A. Frail. A planetary system around the millisecond pulsar PSR1257 + 12. *Nature*, 355(6356):145–147, Jan. 1992. doi: 10.1038/355145a0.

G. S. Wright, D. Wright, G. B. Goodson, G. H. Rieke, G. Aitink-Kroes, J. Amiaux, A. Aricha-Yanguas, R. Azzollini, K. Banks, D. Barrado-Navascues, T. Belenguer-Davila, J. A. D. L. Blommaert, P. Bouchet, B. R. Brandl, L. Colina, Ö. Detre, E. Diaz-Catala, P. Eccleston, S. D. Friedman, M. García-Marín, M. Güdel, A. Glasse, A. M. Glauser, T. P. Greene, U. Groezinger, T. Grundy, P. Hastings, T. Henning, R. Hofferbert, F. Hunter, N. C. Jessen, K. Justtanont, A. R. Karnik, M. A. Khorrami, O. Krause, A. Labiano, P. O. Lagage, U. Langer, D. Lemke, T. Lim, J. Lorenzo-Alvarez, E. Mazy, N. McGowan, M. E. Meixner, N. Morris, J. E. Morrison, F. Müller, H. U. N. rgaard-Nielson, G. Olofsson, B. O'Sullivan, J. W. Pel, K. Penanen, M. B. Petach, J. P. Pye, T. P. Ray, E. Renotte, I. Renouf, M. E. Ressler, P. Samara-Ratna, S. Scheithauer, A. Schneider, B. Shaughnessy, T. Stevenson, K. Sukhatme, B. Swinyard, J. Sykes, J. Thatcher, T. Tikkanen, E. F. van Dishoeck, C. Waelkens, H. Walker, M. Wells, and A. Zhender. The Mid-Infrared Instrument for the James Webb Space Telescope, II: Design and Build. *PASP*, 127(953):595, July 2015. doi: 10.1086/682253.

G. S. Wright, G. H. Rieke, A. Glasse, M. Ressler, M. García Marín, J. Aguilar, S. Alberts, J. Álvarez-Márquez, I. Argyriou, K. Banks, P. Baudoz, A. Boccaletti, P. Bouchet, J. Bouwman, B. R. Brandl, D. Breda, S. Bright, S. Cale, L. Colina, C. Cossou, A. Coulais, M. Cracraft, W. De Meester, D. Dicken, M. Engesser, M. Etxaluze, O. D. Fox, S. Friedman, H. Fu, D. Gasman, A. Gáspár, R. Gastaud, V. Geers, A. M. Glauser, K. D. Gordon, T. Greene, T. R. Greve, T. Grundy, M. Güdel, P. Guillard, P. Haderlein, R. Hashimoto, T. Henning,

D. Hines, B. Holler, Ö. H. Detre, A. Jahromi, B. James, O. C. Jones, K. Justtanont, P. Kavanagh, S. Kendrew, P. Klaassen, O. Krause, A. Labiano, P.-O. Lagage, S. Lambros, K. Larson, D. Law, D. Lee, M. Libralato, J. Lorenzo Alverez, M. Meixner, J. Morrison, M. Mueller, K. Murray, M. Mycroft, R. Myers, O. Nayak, B. Naylor, B. Nickson, A. Noriega-Crespo, G. Östlin, B. O'Sullivan, R. Ottens, P. Patapis, K. Penanen, M. Pietraszkiewicz, T. Ray, M. Regan, A. Roteliuk, P. Royer, P. Samara-Ratna, B. Samuelson, B. A. Sargent, S. Scheithauer, A. Schneider, J. Schreiber, B. Shaughnessy, E. Sheehan, I. Shivaei, G. C. Sloan, L. Tamas, K. Teague, T. Temim, T. Tikkanen, S. Tustain, E. F. van Dishoeck, B. Vandenbussche, M. Weilert, P. Whitehouse, and S. Wolff. The Mid-infrared Instrument for JWST and Its In-flight Performance. *PASP*, 135(1046):048003, Apr. 2023. doi: 10.1088/1538-3873/acbe66.

F. Yan, E. Pallé, A. Reiners, K. Molaverdikhani, N. Casasayas-Barris, L. Nortmann, G. Chen, P. Mollière, and M. Stangret. A temperature inversion with atomic iron in the ultra-hot dayside atmosphere of WASP-189b. *A&A*, 640:L5, Aug. 2020. doi: 10.1051/0004-6361/202038294.

K. H. Yip, Q. Changeat, A. Al-Refaie, and I. Waldmann. To sample or not to sample: Retrieving exoplanetary spectra with variational inference and normalising flows. *arXiv preprint arXiv:2205.07037*, 2022.

K. J. Zahnle and M. S. Marley. Methane, Carbon Monoxide, and Ammonia in Brown Dwarfs and Self-Luminous Giant Planets. *ApJ*, 797(1):41, Dec. 2014. doi: 10.1088/0004-637X/797/1/41.

K. J. Zahnle and M. S. Marley. Methane, carbon monoxide, and ammonia in brown dwarfs and self-luminous giant planets. *ApJ*, 797(1):41, nov 2014. doi: 10.1088/0004-637X/797/1/41. URL `https://dx.doi.org/10.1088/0004-637X/797/1/41`.

J. A. Zalesky, M. R. Line, A. C. Schneider, and J. Patience. A uniform retrieval analysis of ultracool dwarfs. iii. properties of y dwarfs. *ApJ*, 877(1):24, 2019.

A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021.

H. Zhang, J. Wang, and M. K. Plummer. Detecting Biosignatures in Nearby Rocky Exoplanets Using High-contrast Imaging and Medium-resolution Spectroscopy with the Extremely Large Telescope. *AJ*, 167(1):37, Jan. 2024. doi: 10.3847/1538-3881/ad109e.

K. Zhang, J. Bloom, and N. Hernitschek. nbi: the Astronomer's Package for Neural Posterior Estimation. In *Machine Learning for Astrophysics*, page 38, July 2023. doi: 10.48550/arXiv.2312.03824.

Y. Zhang, I. A. G. Snellen, and P. Mollière. The ^{12}CO/^{13}CO isotopologue ratio of a young, isolated brown dwarf. Possibly distinct formation pathways of super-Jupiters and brown dwarfs. *A&A*, 656:A76, Dec. 2021. doi: 10.1051/0004-6361/202141502.

T. Zingales and I. P. Waldmann. ExoGAN: Retrieving Exoplanetary Atmospheres Using Deep Convolutional Generative Adversarial Networks. *AJ*, 156(6):268, Dec. 2018. doi: 10.3847/1538-3881/aae77c.

S. Zorzan, F. Soboczenski, M. D. O'Beirne, M. D. Himes, M. B. Lund, J. C. van Eyken, G. N. Arney, G. L. Villanueva, M. Mascaro, S. D. Domagal-Goldman, and A. G. Baydin. A Machine Learning–ready Data Set for Exoplanet Atmospheric Retrieval. *ApJS*, 277(2):38, Apr. 2025. doi: 10.3847/1538-4365/adb03a.