Astronomy & Astrophysics manuscript no. output
April 20, 2026

Panchromatic view of the frigid Jovian exoplanet COCONUTS-2 b

Matthieu Ravet[1,2,3], Mickaël Bonnefoy[2], Gaël Chauvin[3], Zhoujian Zhang[4], Jacqueline K. Faherty[5], Maël Voyer[6], Mark W. Phillips[7], Pascal Tremblin[6], Rocio Kiman[8], Jessica Copeland[9], James J. Mang[10], Caroline V. Morley[10], Helena Kühnle[11], Benjamin Charnay[12], Sam de Regt[13], Paul Mollière[3], Simon Petrus[14,15,16], Allan Denis[17], Alice Radcliffe[12], Paulina Palma-Bifani[12], Arthur Vigan[17], Mathilde Mâlin[18,19,12], Gabriel-Dominique Marleau[20,21,3], Elena Manjavacas[18,22], Kevin Hoy[15,16,23], Elisabeth C. Matthews[3], and Thomas K. Henning[3]

(Affiliations can be found after the references)

Received XXXX; accepted YYYY

ABSTRACT

Context. Cold exo-Jovian planets are beginning to be imaged and characterized using the James Webb Space Telescope (JWST) instruments. These observations often reveal new molecular species (CO_2, NH_3, PH_3), challenge atmospheric models, and raise questions about the formation pathways and evolution of these objects.

Aims. We revisited the atmosphere of the cold (T_{eff} = 483^{+44}_{-53} K), mature (414 ± 23 Myr), and large-separation (> 5000 au) Jovian exoplanet COCONUTS-2 b (WISEPA J075108.79-763449.6), adding new spectral information beyond 5 μm and combining them with existing spectrophotometry to consolidate the constraints on the object properties and identify disagreements from self-consistent atmospheric models.

Methods. We used a high signal-to-noise MIRI-LRS spectrum (5.45–11 μm, $R_\lambda \sim 100$) of COCONUTS-2 b revealing prominent molecular features of H_2O, CH_4, and NH_3. This dataset is combined with spectra from Gemini/FLAMINGOS-2 and JWST/NIRSpec (G395H), as well as photometry from WISE and Spitzer, resulting in almost continuous wavelength coverage from 1 to 15 μm. We analyzed the data using five grids of self-consistent atmospheric models, spanning a wide range of T_{eff}, log(g), and [M/H]. We also investigated the use of Gaussian processes to account for correlated noise either caused by the spectrograph or by systematic departures of models in the inversion framework.

Results. All models manage to fit the overall combined observations, but predict fainter flux in Y- and N-bands. Classical model comparison suggests that the ATMO2020++ synthetic specra (with and without PH_3) are statistically preferred. However, when accounting for correlated noise using Gaussian processes, Sonora Elf Owl models are favored, although they still provide a comparatively poor fit to the data with bulk properties inconsistent with cooling model predictions. Fitting for the correlated noise of the three spectroscopic instruments, the ATMO2020++ model yields constraints consistent with previous studies and evolutionary model predictions: T_{eff} = 496^{+5}_{-3} K, log(g) = $4.30^{+0.04}_{-0.02}$ dex, [M/H] = $-0.02^{+0.03}_{-0.02}$ dex, and R = $1.03^{+0.01}_{-0.02}$ R_{Jup}. The extended wavelength coverage provided by MIRI (accounting for 41% of the bolometric flux) completes the SED, yielding a precise luminosity estimation of $\log(L/L_\odot) = -6.166 \pm 0.002$ dex. Combined with a previous estimate of the system age (414 ± 23 Myr), cooling models predict a mass of M = 7.3 ± 0.3 M_{Jup}.

Conclusions. The preferred models suggest a metallicity consistent with that of the primary, potentially supporting a binary-like formation scenario. Remaining discrepancies across spectral bands and between model grids suggest incomplete chemistry modeling and highlight the need for improved treatments of alkali condensation and diabatic processes for models at these low effective temperatures.

Key words. Techniques: high angular resolution, spectroscopic; Methods: data analysis, observational, statistical; Planets and satellites: atmospheres, gaseous planets

1. Introduction

Although only a few planetary-mass companions have been detected and characterized using direct imaging[1] (approximately 80), this sample includes diverse objects with effective temperatures (T_{eff}) ranging from 270 to 2700 K and orbital separations spanning from a few tens to thousands of au. The majority of these companions are young (< 300 Myr) and the observed diversity of properties (e.g., log($L/L_\odot$), T_{eff}, log(g), C/O, [M/H], *e*) is a product of each object's formation and evolution history (e.g., Spiegel & Burrows 2012; Allers & Liu 2013; Ruffio et al. 2026).

The COCONUTS-2 (L 34-26 and WISEPA J075108.79-763449.6) system stands out as an outlier in the field of direct imaging. It consists of a cold ~8 M_{Jup} (Zhang et al. 2025), T9 (Kirkpatrick et al. 2011) companion orbiting an M3 star with an estimated age of 414 ± 23 Myr (Kiman et al. 2026), thus providing a unique opportunity to study the atmosphere of a mature super-Jupiter. Originally thought to be two separate free-floating objects, Zhang et al. (2020, 2021a) demonstrated, through proper motion and parallax measurements, that COCONUTS-2 A and b are gravitationally bound. They estimated the semimajor axis of COCONUTS-2 b to be 7506^{+5205}_{-2060} au and its effective temperature to be T_{eff} = 434±9 K, making it the second-coldest exoplanet and the one with the largest directly imaged separation discovered to date. The large separation poses a challenge for planet formation models. COCONUTS-2b may have formed within the circumstellar disk of A, either through core accretion (Pollack et al. 1996; Ida & Lin 2008; Alibert et al. 2005, 2013; Benz et al. 2014) or disk instability (Kuiper 1951; Cameron 1978; Boss 1997), followed by outward migration due to dynamical interactions (Vorobyov 2013). Alternatively, COCONUTS-2b could have formed via gravitational instability (Padoan & Nordlund 2002) in a collapsing molecular cloud, similarly to a stellar binary (Zuckerman & Song 2009; Kirkpatrick et al. 2011). A less likely scenario is a capture during a flyby, as suggested by Marocco et al. (2024). The recent

[1] https://exoplanetarchive.ipac.caltech.edu/

Article published by EDP Sciences, to be cited as https://doi.org/10.1051/0004-6361/202658927

Table 1. Literature values of the physical properties of the COCONUTS-2 system

COCONUTS-2 A		Refs.
Spectral type	M3	(a)
T_{eff} (K)	3406±69	(c)
log(g)	4.83±0.03	(h)
[M/H]	0.00±0.08	(d)
Distance (pc)	10.888±0.002	(f)
Radial velocity (km.s^{-1})	1.19±0.61	(e)
Age (Myr)	414±23	(j)
Mass ($M_\odot$)	0.37±0.011	(h)
Radius ($R_\odot$)	0.388±0.011	(h)
COCONUTS-2 b		
Spectral type	T9	(b)
T_{eff} (K)	483^{+44}_{-53}	(i)
log(g)	$4.19^{+0.18}_{-0.13}$	(i)
Semimajor axis (au)	7506^{+5205}_{-2060}	(h+)
Period (Myr)	$1.1^{+1.3}_{-0.4}$	(h+)
Age (Myr)	414±23	(j)
Mass (M_{Jup})	8±2	(i)
Radius (R_{Jup})	$1.11^{+0.03}_{-0.04}$	(i)

Notes. [(a)] Torres et al. (2006) , [(b)] Kirkpatrick et al. (2011) , [(c)] Gaidos et al. (2014) , [(d)] Hojjatpanah et al. (2019) , [(e)] Schneider et al. (2019) , [(f)] Gaia Collaboration et al. (2021) , [(g)] Bailer-Jones et al. (2021) , [(h)] Zhang et al. (2021a) , [(i)] Zhang et al. (2025) and [(j)] Kiman et al. (2026) . The semimajor axis and orbital period were estimated using orbital predictions based on the eccentricity distributions from Dupuy & Liu (2011); Zhang et al. (2021a) measured a projected separation of 594".

extensive study from Zhang et al. (2025) using the combination of WISE and Spitzer photometry with Gemini/FLAMINGO-2 spectroscopy (0.98–2.51 μm, $R_\lambda \sim 200 - 1200$) alongside multiple atmospheric grids revealed that COCONUTS-2 b likely has a substellar or near-stellar C/O ratio and metallicity ([M/H]). This may suggests that it accreted oxygen-rich and carbon-depleted gas from within ice lines (Line et al. 2021; August et al. 2023), and was not influenced by core erosion processes or planetesimal bombardment (e.g., Alibert et al. 2013; Miller & Fortney 2011; Thorngren & Fortney 2019; Madhusudhan 2019; Schneider & Bitsch 2021; Zhang et al. 2023). This could indicate a planet-like formation followed by outward migration. However, the spread of values in their retrieved metallicities accross models (ranging from -0.4 to 0.1) remains fairly consistent with the metallicity of COCONUTS-2 A (0.00 ± 0.08; Hojjatpanah et al. 2019), and therefore cannot be used to rule out a scenario of a binary-like formation, as the metallicities of the host star and the companion coincide.

We present in this study new observations from JWST/MIRI-LRS that we analyzed jointly with existing spectroscopic data to consolidate the estimate of the physical properties of COCONUTS-2 b. The combination of the very high S/N of these data and the low effective temperature of COCONUTS-2 b presents a challenge for classical forward modeling approaches (Petrus et al. 2024). These temperature and age regimes remain largely unexplored, and self-consistent models may suffer from systematic deviations, either due to missing or incomplete physics (e.g., disequilibrium chemistry, inhomogeneous atmospheres) or due to limitations in parameter space exploration. Our analysis method is presented in Sect. 2, and the results of our analysis are given in Sect. 3. We discuss the implications of our results in Sect. 4, and present our conclusions in Sect. 5. Additional material is provided in the Appendices.

2. Methods

2.1. Observation and data reduction

We used both archival and newly obtained spectro-photometric data on the target. This section describes each of them. Table 2 provides a concise summary of all of them.

Photometry: COCONUTS-2 b was observed with both Spitzer and WISE as part of a survey of 184 late T and Y dwarfs presented in Kirkpatrick et al. (2019). We used the Spitzer [3.6] and [4.5] points, as well as the WISE W1 (3.32 ± 0.33 μm) and W2 (4.56 ± 0.52 μm). The W3 and W4 bands were excluded as their bandpasses are not fully covered by the Sonora Elf Owl grid; W4 lies entirely outside the modeled wavelength range. We note that this has little impact on the retrieved parameters for the other grids.

Gemini/FLAMINGOS-2: FLAMINGOS-2 is a low- to medium-resolution spectrograph located on the 8.1 m Gemini-South telescope. We used archival FLAMINGOS-2 observations of COCONUTS-2 b obtained as part of programs GS-2021B-FT-111 and GS-2021B-FT-204 and first presented in Zhang et al. 2025. The instrument provides simultaneous coverage of the Y-, J-, H-, and K-bands (1.01–2.50 μm). COCONUTS-2 b was observed on 16 and 19 December 2021 with the 2 pixel-wide slit (0.36"×263") and an ABBA nodding pattern, resulting in an average spectral resolution of $R_\lambda \sim 900$ (Zhang et al. 2025). Because COCONUTS-2 A and b are widely separated, the adaptative optics (AO) system was not needed. To extract the final spectrum, a standard data reduction was applied using the *Gemini-IRAF*[2] package.

JWST/NIRSpec: As part of the "Explaining the Diversity of Cold Worlds" survey (Program ID: 2124, PI J. Faherty, DOI), COCONUTS-2 b was observed with JWST/NIRSpec. Observations took place on 9 July 2023 for a total integration time of 2777.706 s and were reduced and presented in Kiman et al. 2026. They used the G395H grating with the F290LP filter element resulting in an average spectral resolution of $R_\lambda \sim 3000$ and a wavelength coverage of 2.88–5.14 μm. We convolved the spectra to a lower resolution of $R_\lambda \sim 1500$ in our analysis to match the sampling of the ATMO2020++ grids.

JWST/MIRI-LRS: MIRI-LRS on the JWST is a low-resolution spectrograph operating in the mid-infrared. It was used to observe COCONUTS-2 b on 29 March 2024 (Program ID: 3514, PI M. Bonnefoy, DOI) for a total integration time of 3335.598 s. The observation strategy was based on using a two-dither pattern with two integrations of 300 groups per exposure. The data were reduced using a combination of the official JWST

[2] https://gemini-iraf-vm-tutorial.readthedocs.io/en/stable/

STScI pipeline,[3] supplemented with customized functions described in Voyer et al. (2025) Sect. 2.2. As noted in other studies (Xuan et al. 2024; Voyer et al. 2025), the retrievals showed a significant wavelength-dependent offset between the line positions of the data and the theoretical models. To correct these offsets we include a third-order polynomial for the wavelength correction in the retrieval. The coefficients of this correction are retrieved once and then fixed for subsequent analysis following the methods in Xuan et al. (2024); Voyer et al. (2025). The wavelength solution from Table 1[4] of Voyer et al. (2025) was used for the pixel-to-wavelength calibration. Beyond 12.5 μm, the extracted spectrum is dominated by background emission. With only two integrations per exposure, the background cannot be reliably averaged out, and the resulting data quality is insufficient for scientific analysis. We therefore excluded these longer wavelengths from the analysis.

2.2. Framework

To infer the atmospheric properties of COCONUTS-2 b, we used a custom version of the Forward Modeling tool for Spectral Analysis, `ForMoSA`[5] (Ravet et al. 2025). It relies on the nested-sampling (Skilling 2004) approach using the Python package `PyMultinest`[6] (Buchner 2016). Each inversion utilizes 1000 live points, adaptive sampling efficiency mode, an evidence tolerance of 0.5; uniform priors were used for the atmospheric parameters. We also fit for the planet radius R constrained by the flux dilution factor $\left(\frac{\mathrm{R}}{d}\right)^2$ using a fixed distance of 10.888 ($\pm$0.002) pc (see Table 1). Nested sampling allows the direct computation of the log-Bayesian evidence, $\ln \mathcal{Z}$, thus enabling a robust quantitative comparison between models via the log-Bayes factor, $\ln \mathcal{B}_{1,2} = \ln \mathcal{Z}_1 - \ln \mathcal{Z}_2$. This metric inherently accounts for both model fit and complexity (i.e., number of free parameters). To facilitate an intuitive interpretation, the Bayes factor is often approximately translated into a sigma significance level, representing the strength of preference for one model over another (Sellke et al. 2001; Trotta 2008; Benneke & Seager 2013). However, as recently pointed out by Kipping & Benneke (2025), such a conversion may only provide an upper limit on the sigma values. The true values are generally lower, which can lead to an overestimation of the significance of model preferences. We therefore chose to rely on the log-Bayesian evidence as a more conservative proxy for model significance. Importantly, the trend we find is also consistent with the Akaike information criterion (AIC; Akaike 1974), the Bayesian information criterion (BIC; e.g., Mollière et al. 2025), and the simplified Bayesian predictive information criterion (BPICS; Ando 2011; Thorngren et al. 2026) across all tested models (see Table 3). For the rest of the study, we define these quantities relative to the favored model:

$$\begin{aligned}
&\ln \mathcal{B} = \ln \mathcal{Z}_{max} - \ln \mathcal{Z},\\
&\Delta\mathrm{AIC} = \mathrm{AIC} - \mathrm{AIC}_{min},\\
&\Delta\mathrm{BIC} = \mathrm{BIC} - \mathrm{BIC}_{min},\\
&\Delta\mathrm{BPICS} = \mathrm{BPICS} - \mathrm{BPICS}_{min}.
\end{aligned} \tag{1}$$

The subscripts *max* and *min* denote the maximum and minimum values across all models. With this convention, the preferred model has $\ln \mathcal{B} = 0$, while all other models have $\ln \mathcal{B} > 0$.

2.3. Likelihood mapping

Since using only the injected error bars coming from the data-reduction pipelines likely underestimates the total uncertainty, we introduce in `ForMoSA` the following modified log-likelihood function:

$$\begin{aligned}
&\ln \mathcal{L}^{\mathrm{classic}} = \sum_i \ln \mathcal{L}_i^{\mathrm{classic}}\\
&\ln \mathcal{L}_i^{\mathrm{classic}} = -\frac{N_i}{2} \ln\left(\frac{\chi_i^2}{N_i}\right) - \frac{1}{2}\ln(|\boldsymbol{\sigma_i}|) - \frac{N_i}{2}\ln(2\pi) - \frac{N_i}{2}\\
&\text{with} \quad \chi_i^2 = \sum \left(\frac{\boldsymbol{d_i} - \boldsymbol{m_i}}{\boldsymbol{\sigma_i}}\right)^2.
\end{aligned} \tag{2}$$

Here the index i takes values from 1 to 4 and corresponds to the three spectroscopic observations and four photometric points; $\boldsymbol{d_i}$, $\boldsymbol{\sigma_i}$, and $\boldsymbol{m_i}$ refer to the data, error, and model vectors, respectively, each of length N_i. This likelihood formulation proposed by Ruffio et al. (2019) is obtained by marginalizing the standard log-likelihood over a noise scaling factor $\boldsymbol{\sigma_s} = s\boldsymbol{\sigma}$. The maximum likelihood estimator for this scaling is

$$\hat{s}_i = \frac{\chi_i^2}{N_i}. \tag{3}$$

This approach partially compensates for model–data mismatches by inflating the error bars, thus propagating them into the posterior distributions. Other approaches exist in the literature, including alternative multiplicative or additive prescriptions (e.g., Piette & Madhusudhan 2020; Zhang et al. 2025). Here we adopted a marginalization over a simple multiplicative factor as a computationally efficient choice that does not assume a particular functional form of the inflation. In a forward modeling framework, the quality of the fit is primarily limited by the fidelity of the atmospheric models rather than by the formal observational uncertainties (Petrus et al. 2025), making the inclusion of these noise scaling parameters a pragmatic way to account for residual model deficiencies without biasing the posteriors. However, they entirely neglect potential spectral correlations introduced by the spectrograph or by systmematic deviations between data and model spread over large bandwidths. To address this limitation, we employed Gaussian processes (GPs) to model correlated noise between neighboring pixels. Building upon previous work (Czekala et al. 2015; Wang et al. 2020; de Regt et al. 2024, 2025; Rotman et al. 2025), we define the covariance matrix as

$$C_{i;\,m,n} = \sigma_{i;\,n}^2\, \delta_{m,n} + a_i^2 \tilde{\sigma_i}^2 \exp\left[-\frac{\Delta\lambda_{m,n}^2}{2\ell_i^2}\right], \tag{4}$$

where each $C_{i;\,m,n}$ represents the total covariance between pixels m and n in the i-th spectrum. The matrix comprises two components: diagonal terms $\sigma_{i;\,n}^2\, \delta_{m,n}$ that account for uncorrelated noise, and radial basis function (RBF) kernel terms that model spectral correlations. The RBF kernel is parameterized by an amplitude a_i and a length scale ℓ_i, and is normalized by the squared median uncertainty $\tilde{\sigma_i}^2$; $\Delta\lambda_{m,n}$ denotes the wavelength separation

[3] https://zenodo.org/records/17515973
[4] https://content.cld.iop.org/journals/2041-8205/982/2/L38/revision2/apjladbd46t1_mrt.txt
[5] https://formosa.readthedocs.io/en/latest/
[6] https://johannesbuchner.github.io/PyMultiNest/

Table 2. Summary of the new and archival data.

Instrument	Type	Spectral coverage	Spectral resolution	S/N	Refs.
WISE	(W1) L-band photometry	3.32±0.33 μm	<30	~69	(a)
	(W2) M-band photometry	4.56±0.52 μm	<30	~167	(a)
Spitzer	(c1) L-band photometry	3.51±0.34 μm	<30	~69	(a)
	(c2) M-band photometry	4.44±0.43 μm	<30	~125	(a)
Gemini/FLAMINGOS-2	YJHK-band spectroscopy	1.01–2.50 μm	200–1200	2–10	(b)
JWST/NIRSpec	G395H/F290LP LM-band spectroscopy	2.88–5.14 μm	2000–3500	10–60	(c)
JWST/MIRI-LRS	P750L N-band spectroscopy	5.00–12.5 μm	30–200	50–700	(d)

Notes. From left to right, the columns give the name of the instrument, the type (spectroscopy or photometry), the wavelength coverage, the (effective) spectral resolution R_λ, and the signal-to-noise ratio (S/N). References: [(a)] Kirkpatrick et al. (2019) , [(b)] Zhang et al. (2025) , [(c)] Kiman et al. (2026) [(d)] (this work) .

between data points m and n. Injecting this covariance matrix $\boldsymbol{C_i}$ into the standard log-likelihood, we obtain

$$\ln \mathcal{L}^{\rm GP} = \sum_i \ln \mathcal{L}_i^{\rm GP}$$

$$\ln \mathcal{L}_i^{\rm GP} = -\frac{\chi_i^2}{2} - \frac{1}{2}\ln(|\boldsymbol{C_i}|) - \frac{N_i}{2}\ln(2\pi)$$

$$\text{with} \quad \chi_i^2 = (\boldsymbol{d_i} - \boldsymbol{m_i})^T \boldsymbol{C_i}^{-1} (\boldsymbol{d_i} - \boldsymbol{m_i}). \tag{5}$$

This likelihood is very similar to the one presented in de Regt et al. (2025), except that we do not fit for a noise scaling factor. This choice is motivated by the fact that including a noise scaling term introduces strong biases in the model-data residuals for MIRI-LRS (see bottom panel of Fig. 1, bottom panels of Fig. 4, and discussion in Sect. 4.2). Equations 2 and 5 assume that each dataset is independent, a reasonable assumption given that each observation was taken with a different instrument at a different time. Each $\log(a_i)$ and $\log(\ell_i)$ are retrieved as free parameters during the inversion.

3. Results

We divided our analysis into three parts. First, we performed an exploration using different atmospheric models evaluated with a simple noise scaling (Eq. 2). Then we carried out a more detailed investigation using GP-aided forward modeling (Eq. 5), focusing on the best-fitting models identified in the first stage. Finally, we used the preferred solutions to rederive the bolometric luminosity using the full spectral energy distribution (SED) and, in combination with evolutionary models, infer the corresponding bulk properties. The aim of this strategy is to mitigate some of the limitations of current forward modeling approaches (such as model systematics and correlated noise) to derive more robust constraints on the atmospheric parameters of COCONUTS-2 b.

Similarly to Zhang et al. (2025), we masked five regions of the FLAMINGOS-2 observation and performed the fit on the following ranges: 1.05–1.12 μm, 1.18–1.33 μm, and 1.52–1.66 μm (see Fig. 1). Fluxes from wavelengths shorter than 1.05 μm were masked to avoid the area with low S/N near the edge of the detector and valleys between the Y, J, H, and K peaks due to the residuals from telluric correction and the relatively large flux uncertainties in the spectrum. We also ignored any variability and did not adjust for relative flux offsets between instruments as no variability has been reported for COCONUTS-2 b.

3.1. Model comparison

Table 3. Model comparison and significance criteria for all models in Sect. 3.1.

Model	$\ln \mathcal{B}$	ΔAIC	ΔBIC	ΔBPICS
BT-Settl	3381	6776	6751	6775
Sonora Diamondback	2383	4767	4744	4768
Sonora Elf Owl	331	665	668	656
ATMO2020++ (no PH_3)	139	286	286	286
ATMO2020++	0	0	0	0

Notes. We used the formula from Spiegelhalter et al. (2002) to define the effective number of parameters in the computation of BPICS.

Figures 1 and 2 and Tables 3 and A.3 summarize the inversion results using the five different models. While most models manage to reproduce the overall shape of the combined observations, we observed significant data–model mismatches across all instruments. We observe an important dispersion between models and their retrieved posterior distributions (see Fig. 2), with almost no overlap. This huge dispersion suggests that the models are unable to robustly extract the physical parameters. Comparing the log-Bayes factors (see Table 3), the ATMO2020++ models (with and without PH_3) are statistically preferred among those tested. This result is in line with a previous atmospheric analysis of the target (Zhang et al. 2025), although the model with PH_3 seems to be preferred ($\ln \mathcal{B} = 138$). The PH_3 detection significance is discussed in Sect. 4.3.

The retrieved effective temperatures range from 450 to 540 K for models that converged within their parameter grids (excluding Sonora Diamondback, which converges at the grid edge with $T_{\rm eff} > 600$ K). This is broadly consistent with previous expectations from evolutionary tracks, which predict $T_{\rm eff} = 483^{+44}_{-53}$ K (Zhang et al. 2025). The lower limit of $T_{\rm eff} > 600$ K obtained with Sonora Diamondback probably explains the broader constraint observed in luminosity (see purple histogram in lower left panel of Fig. 2) due to the degeneracy with the radius. We note that this temperature is highly inconsistent with evolution-

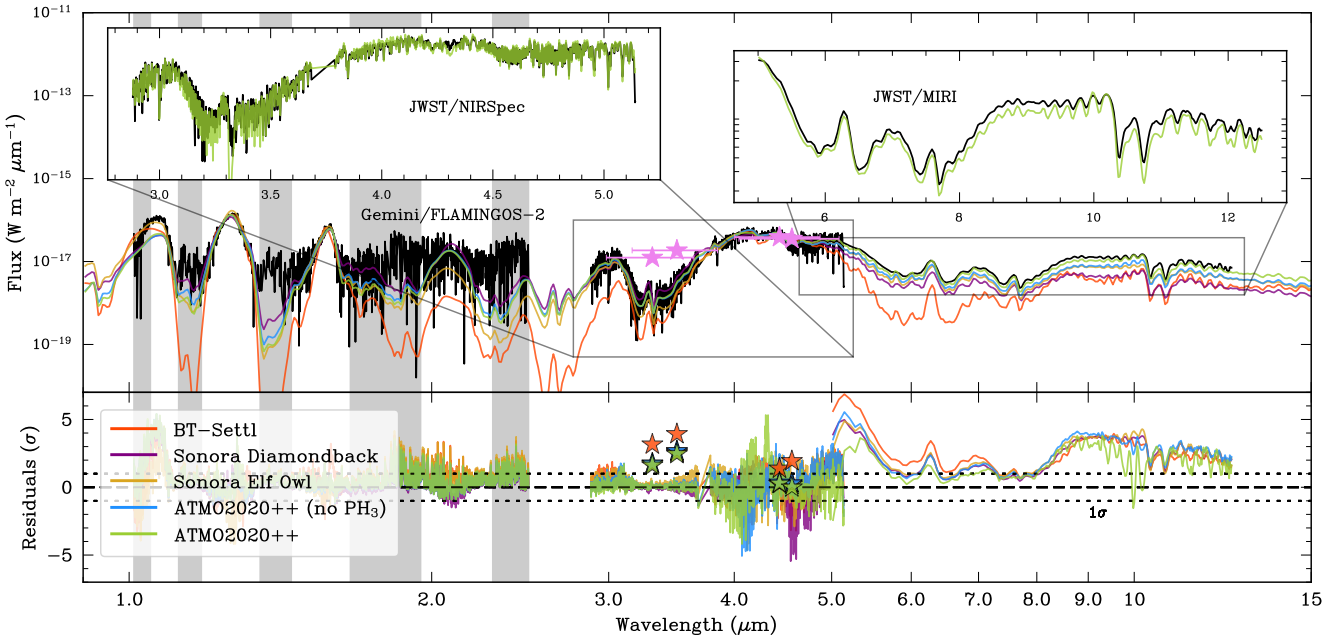

Fig. 1. *Top panel:* Forward modeling results of the combined COCONUTS-2 b observations. The black solid lines represent the spectroscopic data (from left to right: Gemini/FLAMINGOS-2, JWST/NIRSpec, and JWST/MIRI-LRS), while the colored lines represents each $R_\lambda \sim 100$ model (classic). The pink and white stars indicate the photometric observations (WISE and Spitzer). The gray shaded regions indicate the masked wavelengths during the spectral inversion. *Bottom panel:* Residuals for each fit. The dotted lines represents the $\pm\sigma$ (68%) confidence interval

ary predictions. The surface gravity is poorly constrained, with retrieved values between 4.5 and 5.0 dex, significantly higher than the evolutionary estimate of log(g) = $4.19^{+0.18}_{-0.13}$ dex. Notably, inversions using BT-Settl and Sonora Elf Owl reached the edges of their grids, with posteriors at > 4.5 dex and < 3.25 dex, respectively. However, comparisons between atmospheric and evolutionary log(g) estimates should be treated with caution as they trace different physical processes (e.g., atmospheric structure versus radius contraction; Petrus et al. 2025).

Except for Sonora Elf Owl, all grids that explored metallicity converged toward slightly supersolar values ([M/H] $\in$ 0.2–0.4 dex). Since we used the same dataset, it is reasonable to assume that the observed differences in metallicity between the various inversions arise from systematic offsets between the models. Because metallicity is correlated with log(g) (Zhang et al. 2021b), the systematically higher surface gravities retrieved for most models in this setup may also bias the inferred metallicity. Moreover, metallicity is highly sensitive to the wavelength range and datasets included in the inversion (Petrus et al. 2024). Sonora Elf Owl is the only grid that also probes the carbon-to-oxygen ratio (C/O), yielding a subsolar value of 0.317 ± 0.005.

Of the two grids that include cloud modeling, only Sonora Diamondback explores their effects explicitly through the sedimentation efficiency parameter of its silicate clouds (f_{sed}), which ranges from 1 (thick clouds) to 8 (thin clouds). Our best-fit value of $f_{sed} = 3.6^{+0.1}_{-0.2}$ suggests a moderately cloudy atmosphere. However, since both cloudy models (BT-Settl and Sonora Diamondback) are among the least favored in our analysis, this constraint should be interpreted with caution as it may not provide accurate information about the actual cloud content and properties of COCONUTS-2 b.

The vertical (eddy) diffusion coefficient K_{zz} (in cm^2.s^{-1}) parameterizes the efficiency of atmospheric vertical mixing. In the Sonora Elf Owl grid, this parameter is explored logarithmically from 10^2 cm^2.s^{-1} (inefficient mixing) to 10^9 cm^2.s^{-1} (efficient mixing). We retrieved a value of $K_{zz} = (9.8^{+0.5}_{-0.7}) \times 10^3$ cm^2.s^{-1}, indicative of moderate vertical mixing. The vertical mixing can also be retrieved analytically from the inferred log(g) using Fig. 1 of Phillips et al. (2020) for the two ATMO2020++ grids. Both ATMO2020++ models suggest a more vigorous vertical mixing with $K_{zz} = (2.5 \pm 0.1) \times 10^5$ cm^2.s^{-1} and $K_{zz} = (1.3 \pm 0.1) \times 10^5$ cm^2.s^{-1} for the model with and without PH_3, respectively. This strong vertical mixing can also be seen in Fig. A.5, where all the key chemical abundances are quenched. However, as noted and explored in Mukherjee et al. (2022, 2024), K_{zz} is expected to vary significantly with pressure, depending on whether the local atmospheric layer is radiative or convective. This effect is not accounted for in this setup. In addition, the retrieved K_{zz} may partially encode viewing-geometry effects, further contributing to the degeneracy of this parameter (Tan & Showman 2021; Suárez et al. 2023; Petrus et al. 2025). Recent studies have investigated how mixing strength may vary in more complex 3D atmospheres (Visscher et al. 2010; Mukherjee et al. 2022; Lacy & Burrows 2023).

The top left panel of Fig. 2 compares the retrieved pressure–temperature (P–T) profiles for Sonora Diamondback (purple), Sonora Elf Owl (yellow), and ATMO2020++ (green). All the profiles agree well with each other.

3.2. GP analysis

In Sect. 3.1 we identified ATMO2020++ and its variation ATMO2020++ (no PH_3) as the two preferred models; ATMO2020++ (no PH_3) is the only model that does not converge outside its grid range for any parameter. Using the GP framework presented in Sect. 2.2, we reanalyzed the combined dataset. Assuming that each dataset is independent, we fit for each covariance matrix separately, resulting in three length-scales and three amplitudes as additional extra-grid parameters.

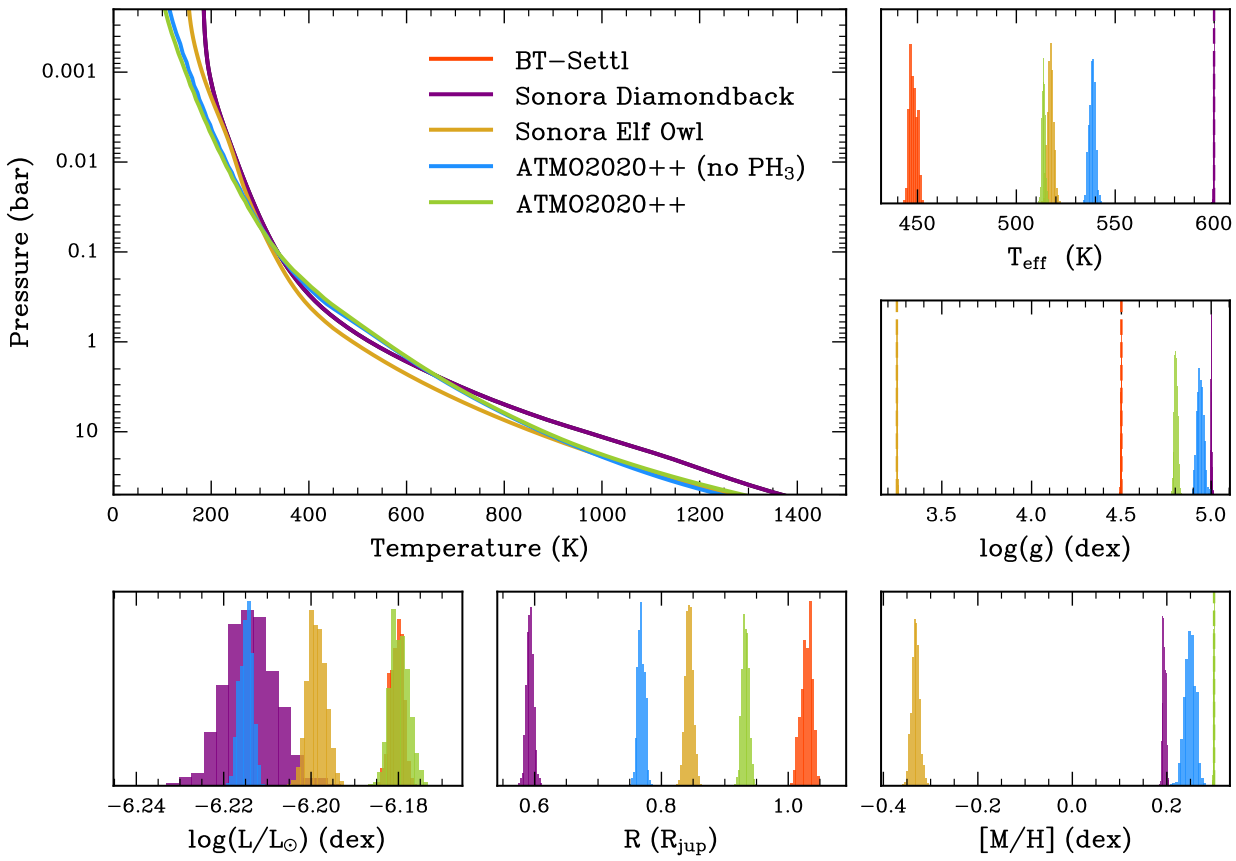

Fig. 2. *Top left panel:* Interpolated pressure–temperature (P–T) profiles for models providing a P–T grid (excluding BT-Settl) in Sect. 3.1. *Remaining panels:* Posteriors of key parameters. The vertical dashed lines indicate the boundaries of the respective model grids when encountered during the inversion.

Table 4. Model comparison and significance criteria for all models in Sect. 3.2.

Model	$\ln \mathcal{B}$	ΔAIC	ΔBIC	ΔBPICS
BT-Settl GP	1219	2445	2426	2448
Sonora Diamondback GP	732	1463	1457	1447
ATMO2020++ (no PH_3) GP	319	643	630	648
ATMO2020++ GP	94	197	123	196
Sonora Elf Owl GP	0	0	0	0

Notes. We used the formula from Spiegelhalter et al. (2002) to define the effective number of parameters in the computation of BPICS.

This section focuses on the two ATMO2020++ models; however, we report the final parameters obtained with the other grids in Tables A.3 and A.4, and additional elements of discussion are provided in Appendix D. Notably, based on the various significance criteria (Table 4), the Sonora Elf Owl models become preferred when GPs are included, highlighting the importance of the GP framework in the modeling. However, these models still provide a comparatively poor fit, both visually (large residual offsets) and in terms of the inferred physical parameters (inconsistent with cooling model predictions). We therefore retain the focus of this section on the two ATMO2020++ grids. Final posteriors are compared with the classical approach and evolutionary models predictions in Fig. 3 and summary plots can be found in Figs. A.1, A.2, and 4. In particular, in Fig. 4, the GP approach significantly improves the fit to the MIRI-LRS observation, where the flux was previously underestimated. The effect is less noticeable for the other two spectra. In the first column of Table A.3, both GP-aided results exhibit a significant decrease in the relative log-Bayes factor, suggesting that accounting for correlated noise is relevant for this target. In this configuration, the ATMO2020++ model including PH_3 is again favored over the version without it, with a relative log-Bayes factor of $\ln \mathcal{B} = 319 - 94 = 225$ (see Table 4).

Focusing on ATMO2020++, Fig. 3 shows both a significant shift and broadening of the posterior distributions between the classical and GP approaches for all atmospheric parameters. We obtain $T_{\rm eff} = 496^{+5}_{-3}$ K and $\log(g) = 4.30^{+0.04}_{-0.02}$ dex; consistent with predictions from evolutionary models at 0.9 and 3.0σ, respectively, using the values reported in Table A.2 (490^{+3}_{-4} K and $4.16^{+0.03}_{-0.04}$ dex; see Sect. 3.3 for a more detailed description). The metallicity shifts markedly, from a supersolar [M/H] > 0.30 dex in the classical approach to solar [M/H] = $-0.02^{+0.03}_{-0.02}$ dex with the GP model, probably linked to the associated decrease in log(g). From our inversion, we infer a radius of $R = 1.03^{+0.01}_{-0.02}$ $R_{\rm Jup}$ (consistent at 3.3σ), which translates into an estimated luminosity of $\log(L/L_\odot) = -6.163 \pm 0.002$ dex when combining the datasets. This luminosity estimate is also consistent with the value of -6.162 ± 0.006 dex obtained by Zhang et al. 2025 with the same model.

The retrieved hyperparameters for each observation listed in Table A.4 are consistent between the two ATMO2020++ models. These hyperparameters are also consistent across the other model grids (see Table A.4 and Appendix D). Among the

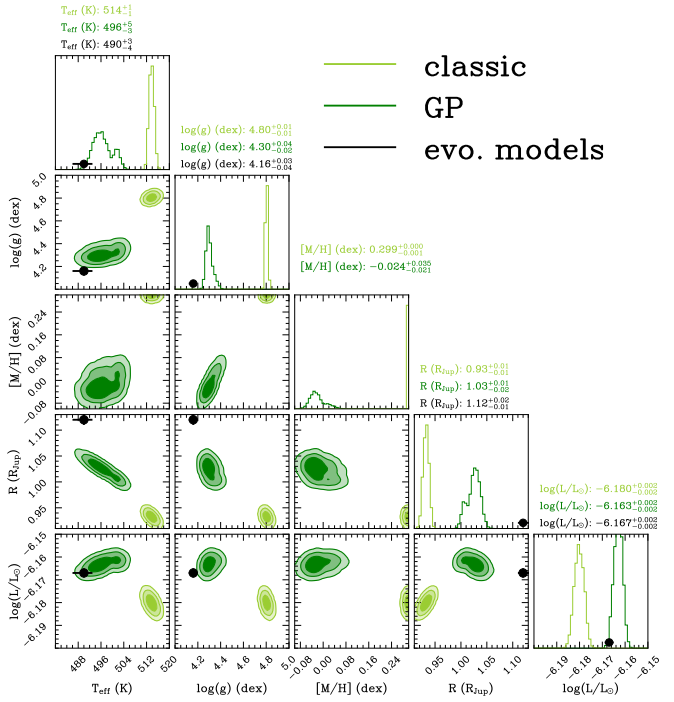

Fig. 3. Corner plot comparing forward modeling using the ATMO2020++ grid with (dark green) and without (light green) Gaussian processes (GPs). The black points indicate the mean predictions from evolutionary models with their associated error bars (see Sect. 3.3).

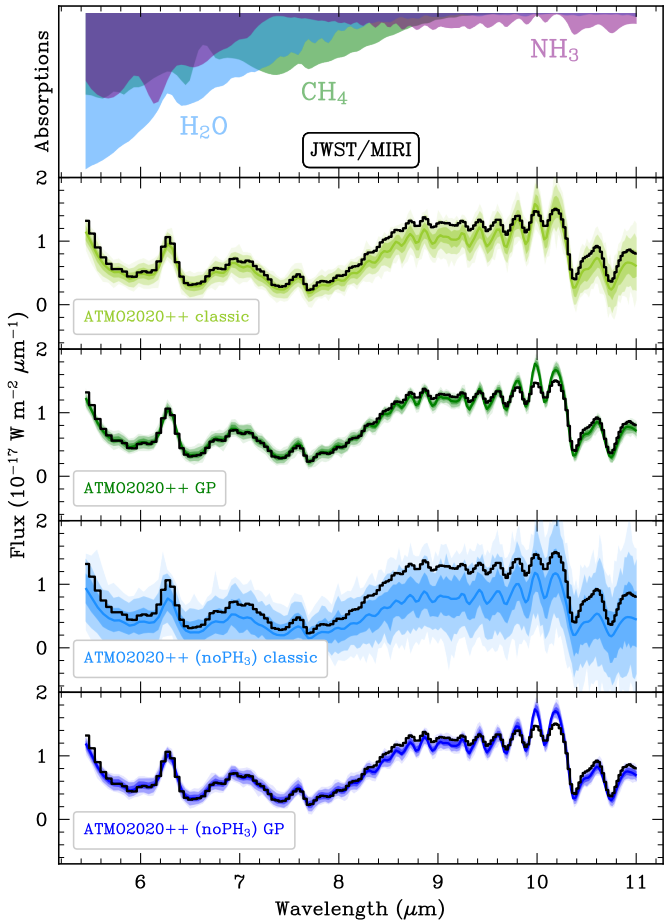

Fig. 4. Same as Fig. A.1, but for a comparison between the classical and GP-aided approaches on the JWST/MIRI-LRS spectrum using ATMO2020++ (in green) and ATMO2020++ (no PH_3, in blue). *Top panel:* Main molecular absorption features calculated using `petitRADTRANS`. In the other panels, the black lines show the observed spectrum; the colored lines show the corresponding model spectrum. The shaded regions around the model indicate the 1σ, 2σ, and 3σ dispersions from 200 random draws from the retrieved covariance matrix $\boldsymbol{C}$. In the classical approach, the covariance matrix is assumed to be diagonal, $\boldsymbol{C} = \hat{s}^2\,\mathrm{diag}(\boldsymbol{\sigma}^2)$, where $\hat{s}^2$ is the noise scaling factor and $\boldsymbol{\sigma}$ the observational uncertainties. Zoom-in on the 5–11 μm region.

datasets, MIRI-LRS exhibits the largest fitted correlation amplitude, with $\log(a)_{\mathrm{MIRI}} = 1.24 \pm 0.02$.

The top row of Fig. A.3 presents a zoomed-in image of the three retrieved covariance matrices, normalized to highlight their noise structures. The correlated noise pattern is particularly visible in the FLAMINGOS-2 matrix with correlated noise extending across multiple spectral channels. Its block-like structure arises from the spectral windows used in the analysis.

The middle row displays the corresponding mean radial profiles, overplotted with the autocorrelation function (ACF) of the residuals. Comparing these profiles allows us to make an assessment of whether the retrieved correlation lengths adequately capture the actual noise structure in the data, assuming the model provides a good fit and the residuals are therefore dominated by noise. We find that the retrieved correlation lengths generally match this noise structure, although the correlation length may be slightly underestimated for the MIRI-LRS observation. Spectrographs with the finest wavelength sampling (FLAMINGOS-2 and NIRSpec) exhibit the largest correlation lengths, often spanning multiple pixels. In contrast, the retrieved correlation length for MIRI-LRS is nearly equal to the pixel size ($\sim 0.02\,\mu$m), suggesting more localized noise correlations.

The bottom row of Fig. A.3 compares the power spectral density (PSD) of the residuals, the injected observational uncertainties, and GP realizations. At low spectral frequencies, the PSD of the residuals (in green) is well captured by the GP (in red), and white noise (in blue) is not sufficient to explain the residual noise. At higher frequencies, for both the FLAMINGOS-2 and NIRSpec spectrographs, the PSD of the residuals is below that of the injected uncertainties and the GP noise, which suggests that the adopted noise model is too conservative at very small spectral scales.

Using the GP-aided spectral analysis on the combined dataset, we adopt the following parameters for COCONUTS-2 b: $T_{\mathrm{eff}} = 496^{+5}_{-3}$ K, $\log(g) = 4.30^{+0.04}_{-0.02}$ dex, $[M/H] = -0.02^{+0.03}_{-0.02}$ dex, $R = 1.03^{+0.01}_{-0.02}$ R_{Jup}, and $\log(L/L_\odot) = -6.166 \pm 0.002$ dex (from Sect. 3.3).

3.3. SED analysis

In Sects. 3.1 and 3.2, we constrain the bolometric luminosity of COCONUTS-2 b by propagating the posterior distributions of its effective temperature and radius using the Stefan–Boltzmann law (see the last column of Table A.3). This approach implicitly assumes that the bolometric flux can be accurately described by a single effective temperature. A more direct and less assumption-driven estimate can be obtained by directly integrating the full SED.

Because the wavelength coverage is not continuous, this SED must be completed. Previous studies (e.g., Filippazzo et al. 2015; Petrus et al. 2025; Kiman et al. 2026) have extrapolated the missing regions using Wien's approximation at short wavelengths and the Rayleigh–Jeans tail at long wavelengths. This approach avoids introducing explicit atmospheric model assumptions (systematics) and is well suited for relatively warm L/T objects with smooth spectral energy distributions.

However, for very cold objects such as COCONUTS-2 b, whose emergent spectra are strongly shaped by molecular absorption, the assumption of smooth blackbody-like behavior outside the observed range may be less accurate. For this work we chose to use the posterior chain from our best-fit model (ATMO2020++ with GP; see Sect. 3.2) to generate a sample of model spectra over the widest possible wavelength range (0.2–30 μm). We then propagated this sample alongside the posterior distribution of the radius to derive a probability distribution for the bolometric luminosity. As previously, to account for potential model–data discrepancies, we also applied the corrective term given by Eq. (3) of Zhang et al. (2025), yielding

$$\log(L/L_\odot) = -6.166 \pm 0.002 \text{ dex}. \quad (6)$$

By repeating the procedure describe above on MIRI-LRS's wavelength range, we can estimate the contribution of the MIRI-LRS observation to the total bolometric flux to be 41%. This value highlights the crucial role of the new dataset in constraining the bolometric luminosity of COCONUTS-2 b.

Finally, we used the ATMO2020++ (Phillips et al. 2020), Sonora Bobcat (Marley et al. 2021) and Sonora Diamondback hybrids (Morley et al. 2024) evolutionary models to rederive the physical properties of COCONUTS-2 b. We assumed a Gaussian distribution for the luminosity and the value we estimate (Eq. 6) and the Gaussian distribution for the age, based on the predictions from Kiman et al. (2026, 414 ± 23 Myr). The inferred parameters are summarized in Table A.2. While these parameters are fairly consistent across the models, there is an observable difference in radius and effective temperature between the ATMO2020 models and the Sonora Bobcat models (see Fig. A.6). After concatenating the chains for each parameter following the Monte Carlo sampling, evolution models predict that COCONUTS-2 b has $T_{\rm eff} = 490^{+3}_{-4}$ K, $\log(g) = 4.16^{+0.03}_{-0.04}$ dex, $R = 1.12^{+0.02}_{-0.01}$ $R_{\rm Jup}$, and $M = 7.3 \pm 0.3$ $M_{\rm Jup}$. These predictions are compared to those from the atmospheric models in Fig. 3.

4. Discussion

4.1. Atmosphere and formation of COCONUTS-2 b

In Sect. 3.1 we discuss the significant discrepancies in the atmospheric predictions from the different tested models. The cloud-free models (Sonora Elf Owl, ATMO2020++, and ATMO2020++ without PH_3) are significantly preferred based on Bayesian evidence, although the differences between them are also noticeable in the overall fit (see Fig. 1).

In particular, in the Y-band (FLAMINGOS-2) and N-band (MIRI), all models underestimate the observed flux. As suggested in Zhang et al. (2025), the discrepancies observed in the Y-band may be due to uncertainties in the modeling of the pressure-dependent Na and K lines. Although the preferred and most recent grids we use (ATMO2020++ and Sonora Elf Owl) include a self-consistent decrease in the abundance of these chemical species in the upper atmospheric layers through condensation and precipitation processes, we argue that increased precipitation or additional processes are needed to reduce the abundance of Na and/or K to match the observations. The flux underestimation in MIRI's wavelength range is not seen in the GP inversions (Sect.3.2, Fig.4), suggesting that ATMO2020++ can reproduce this part of the spectrum reasonably well without invoking additional physical or chemical processes. This discrepancy between the classical and GP approaches arises from differences in the treatment of the noise as, in the classical approach, the noise marginalization effectively reduces the importance of the MIRI-LRS data during the inversion (see Sect. 4.2).

We pushed the atmospheric analysis using a GP-aided framework in Sect. 3.2 on the ATMO2020++ and ATMO2020++ (no PH_3) models, which resulted in a significant improvement in both the fit quality and consistency with evolutionary predictions. For both best-fit models, we retrieved metallicities consistent with the metallicity of the primary (0.00 ± 0.08 dex; Hojjatpanah et al. 2019) with $[M/H] = -0.02^{+0.03}_{-0.02}$ dex and $[M/H] = -0.09 \pm 0.02$ dex for the models including and excluding PH_3, respectively. This general agreement suggests that COCONUTS-2 b has not experienced significant enrichment through core erosion or planetesimal accretion, and is instead compatible with a binary-like formation scenario. However, [M/H] is among the most degenerate parameters in our analysis analysis (Fig. 2 and Table A.3), with retrieved values spanning from supersolar to subsolar depending on the model employed. For instance, both Sonora models favor a more pronounced subsolar metallicity with $[M/H] = -0.43 \pm 0.01$ dex and $[M/H] = -0.30 \pm 0.02$ dex for Diamondback and Elf Owl, respectively, when using GP (see Table A.3). This could in turn suggest a different formation pathway (e.g., planetary, late accretion). Overall, this parameter is strongly coupled to the entire atmospheric composition, which means it is highly sensitive to the treatment of chemical processes, particularly non-equilibrium chemistry, which is nevertheless properly included in all the favored models.

4.2. Noise modeling

Accounting for all noise sources within a single inversion framework is a challenging task (e.g., Czekala et al. 2015; Gully-Santiago & Morley 2022). One of the simplest and least computationally expensive approaches is to include a noise scaling parameter in the Bayesian estimator, which can either be fitted directly (e.g., Zhang et al. 2025) or marginalized over (e.g., Ruffio et al. 2019). This scaling factor can be additive (e.g., Zhang et al. 2025) or, more commonly, multiplicative (e.g., Ruffio et al. 2019), and aims to compensate for missing and uncorrelated noise sources.

Marginalizing over this factor has been shown to significantly improve the robustness of retrieved posteriors in low (Landman et al. 2024; Denis et al. 2025) and in high S/N regimes (de Regt et al. 2024, 2025). However, our study demonstrates that such an approach is less suited to dealing with multimodal datasets. Marginalizing over this scaling factor effectively equalizes the S/N contribution of each dataset. While MIRI-LRS intrinsically has the highest S/N and would therefore be expected to dominate the fit, the optimization process increases its noise-scaling factor (see Table A.1), effectively reducing its weight and enhancing the contribution of the FLAMINGOS-2 and NIRSpec spectra. This effect is clearly visible in Fig. 4: in the classical approach, the noise-scaling parameter allows the model to underestimate the MIRI-LRS flux, whereas this bias is not present in the GP approach, which does not marginalize over the noise-scaling factor.

For this study we used GPs to account for unmodeled spectral correlations, fitting for correlation lengths and amplitudes for each spectroscopic observations. We assessed the reliability of the retrieved correlation lengths from the ACF and PSD of the residuals (see Sect. 3.2 and Fig. A.3). In contrast, the correlation amplitudes depend on the overall scale of the residuals; their reliability is therefore best evaluated through injection tests as illustrated in Appendix C.

Accounting for correlated noise in the two ATMO2020++ models statistically improved both the fit quality and the consistency of most of the retrieved parameters (Table A.3). These correlations may arise from a combination of instrumental effects, and model systematics. In practice, the two contributions are difficult to disentangle from the residuals alone since the empirical ACF and PSD reflects both (residuals = observation - model). For the FLAMINGOS-2 and NIRSpec instruments, the residual autocorrelation and the GP reveal correlation lengths are both larger than (or on the order of) the instrumental line spread function (~2–3 pixels, see middle row of Fig. A.3), indicating the presence of additional, more complex noise structures in the data that must be accounted for in the analysis.

More broadly, GP applications in exoplanet atmospheric characterization can extend beyond noise modeling. They have, for instance, been extensively used in transit spectroscopy to model systematic noise sources and correct for stellar contamination (e.g., Gibson et al. 2012). In the context of forward modeling, they can also be employed to remove local deviant spectra using the overall trends of the grids, therefore mitigating the effect of these spectra on the posterior distributions (Czekala et al. 2015). Moreover, GPs can propagate interpolation uncertainties into the final posteriors, as demonstrated in Czekala et al. (2015) with their `Starfish`[7] framework. These last two aspects, while promising, were not explored for the present work. Future forward modeling efforts will benefit from incorporating these GP-based developments, to better characterize the noise structure and to more robustly disentangle observational systematics from model-driven discrepancies.

4.3. PH_3 analysis

Phosphine (PH_3) was expected to be the primary phosphorous molecule in the low-temperature atmospheres of brown dwarfs and giant exoplanets (Fegley & Lodders 1996; Visscher et al. 2006). However, most observational searches for PH_3 in these planetary-mass objects (e.g., Morley et al. 2018; Wallack et al. 2024) have provided only upper limits that are ~100 times lower than the abundances predicted by atmosphere models (Beiler et al. 2024). In their paper Beiler et al. 2024 discussed the atmospheric mechanisms that could explain the observed underabundance of PH_3, including a vertical eddy diffusion coefficient that varies with altitude, incorrect chemical pathways, elements condensing out in forms such as $NH_4H_2PO_4$, or incorrect quenching approximations.

Throughout our analysis, the ATMO2020++ model including PH_3 remained statistically preferred compared to the same model without this molecule. The strongest features of PH_3 are located between 4 and 4.5 μm with a deep broad line at 4.3 μm. This deep 4.3 μm feature is clearly visible on the final fit (second and third panels from the top of Fig. A.2), but does not appear in the data at this location, clearly suggesting that PH_3 is indeed absent from the observation. On the other hand, the model without PH_3 manages to fit this region (fourth and fifth panels from the top of Fig. A.2), but consistently overestimate the flux between 4 μm and 4.2 μm (with or without GP) where CH_4 lines are present.

To evaluate whether this broader 4–4.2 μm region could be driving the preference for the model with PH_3, we masked it and reran the inversions. In this case, ATMO2020++ without PH_3 was favored with $\ln \mathcal{B} = 724$ using the classical approach and $\ln \mathcal{B} = 54$ with the GP-aided approach. These results clearly indicate that the putative detection of PH_3 absorptions is highly sensitive to wavelength coverage and model systematics. Neither model successfully reproduces the NIRSpec data in this region implying that robust constraints on PH_3 detection are difficult to obtain with this approach.

To complement this analysis, we performed a cross-correlation exploration of the NIRSpec spectrum using molecular templates. Templates for H_2O, CO, CH_4, CO_2, NH_3, and PH_3 were generated with `petitRADTRANS`[8] (Mollière et al. 2019; Blain et al. 2024) using the median temperature and chemical profiles from the ATMO2020++ (GP) inversion (see Fig. A.5). We used the line-by-line opacity computation mode to generate spectra with only one molecule at a time. Both the observed spectrum and the templates were continuum-removed and normalized, and the template spectral resolution was matched to that of the observations. To avoid the 3.6–3.8 μm gap, we performed the cross-correlation on the two NIRSpec detectors separately. The resulting cross-correlation functions (CCF) are shown in Fig. A.7. H_2O, CO, CH_4, CO_2, and NH_3 are detected at significance levels above 4σ, while no significant signal is recovered for PH_3, further supporting its absence (or reduced abundance) in the atmosphere of COCONUTS-2 b.

5. Conclusion

In this work we investigated the T9 planetary-mass companion COCONUTS-2 b. Using new MIRI-LRS observations in combination with existing spectro-photometric data within a self-consistent GP-aided framework, we refined the atmospheric properties of this object. From the GP-aided model analysis, we adopted the following parameters: $T_{\rm eff} = 496^{+5}_{-3}$ K, $\log(g) = 4.30^{+0.04}_{-0.02}$ dex, $[M/H] = -0.02^{+0.03}_{-0.02}$ dex, and $R = 1.03^{+0.01}_{-0.02}$ $R_{\rm Jup}$. These values are consistent with previous results (Zhang et al. 2021a, 2025; Kiman et al. 2026) and with the updated predictions from evolutionary models. The adopted metallicity, $[M/H] = -0.02^{+0.03}_{-0.02}$ dex, is consistent with that of the primary, supporting a binary-like formation scenario for COCONUTS-2 b. The wavelength extension provided by the MIRI-LRS observations contribute 41% of the bolometric flux. It enables a robust new constraint on the luminosity of COCONUTS-2 b: $\log(L/L_\odot) = -6.166 \pm 0.002$ dex.

However, systematic discrepancies between the data and atmospheric models in specific bands (Y and N), as well as inconsistencies in the retrieved metallicities and surface gravities across different model grids, remain the main limitations to providing a fully robust characterization of the object. These issues point to missing or incomplete chemistry in the current forward models, and highlight the need to incorporate improved alkali condensation and rainout, potentially more complex cloud structures, and more generally, diabatic processes. Such processes are increasingly being recognized as essential for interpreting the cold brown dwarf population revealed by JWST (e.g., Faherty et al. 2024; Alejandro Merchan et al. 2025; Kiman et al. 2026; Biller et al. in prep.). Accounting for them

[7] https://starfish.readthedocs.io/en/latest/

[8] https://petitradtrans.readthedocs.io/en/latest/

self-consistently will thus be equally important in preparation for the population of temperate near-water-ice-line planets expected from *Gaia* DR4 and the ELT, and, in the longer term, for extending these modeling efforts toward exo-Earth analogs.

We also find that classical noise-scaling schemes are ill-suited for heterogeneous multi-modal datasets with varying resolutions and S/N values. In contrast, parametric covariance models provide a robust alternative for small to moderate-size observation samples.

The aim of future retrieval analyses will be to constrain individual molecular abundances and explore more complex cloud prescriptions (Copeland et al., in prep.; Kühnle et al., subm.) using the data presented here and with the upcoming MIRI–MRS observations, providing further insight into the formation history of this object. Additionally, the high S/N MIRI-LRS spectrum will enable constraints on the isotope ratios (e.g., D/H, ^{12}C/^{13}C).

Acknowledgements. This work is heavily supported by a large collaboration of international people: Atmospheric modeling groups represented by P. Tremblin and M. W. Phillips (ATMO), F. Allard (deceased) (BT-Settl), J. J. Mang and C. V. Morley (Sonora). Collaborators who provided us with the data are represented by the M. Bonnefoy (MIRI-LRS), J. K. Faherty (NIRSpec), Z. Zhang (FLAMINGOS-2). This project has received funding from *Agence Nationale de la Recherche* (ANR) under grant ANR-23-CE31-0006-01 (MIRAGES). This project was provided with computing HPC and storage resources by GENCI at TGCC thanks to the grant 2024-15722 and 2025-15722 on the supercomputer Joliot Curie's SKL and ROME partition. This work is based [in part] on observations made with the NASA/ESA/CSA James Webb Space Telescope. The data were obtained from the Mikulski Archive for Space Telescopes at the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS 5-03127 for JWST. These observations are associated with GTO program 2124 and 3514. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. G.-D.M. acknowledges the partial support of the Deutsche Forschungsgemeinschaft (DFG) through grant "MA 9185/2-1". This publication is based upon work from COST Action CA22133 "PLANETS" (https://costactionplanets.github.io), supported by COST (European Cooperation in Science and Technology).

References

Ackerman, A. S. & Marley, M. S. 2001, ApJ, 556, 872
Akaike, H. 1974, IEEE Transactions on Automatic Control, 19, 716
Alejandro Merchan, S., Faherty, J. K., Suárez, G., et al. 2025, ApJ, 989, 80
Alibert, Y., Carron, F., Fortier, A., et al. 2013, A&A, 558, A109
Alibert, Y., Mordasini, C., Benz, W., & Winisdoerffer, C. 2005, A&A, 434, 343
Allard, F., Homeier, D., & Freytag, B. 2012, Philosophical Transactions of the Royal Society of London Series A, 370, 2765
Allers, K. N. & Liu, M. C. 2013, ApJ, 772, 79
Ando, T. 2011, American Journal of Mathematical and Management Sciences, 31, 13
August, P. C., Bean, J. L., Zhang, M., et al. 2023, ApJ, 953, L24
Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, AJ, 161, 147
Batalha, N. E., Marley, M. S., Lewis, N. K., & Fortney, J. J. 2019, ApJ, 878, 70
Beiler, S. A., Mukherjee, S., Cushing, M. C., et al. 2024, ApJ, 973, 60
Benneke, B. & Seager, S. 2013, ApJ, 778, 153
Benz, W., Ida, S., Alibert, Y., Lin, D., & Mordasini, C. 2014, in Protostars and Planets VI, ed. H. Beuther, R. S. Klessen, C. P. Dullemond, & T. Henning, 691–713
Blain, D., Mollière, P., & Nasedkin, E. 2024, The Journal of Open Source Software, 9, 7028
Boss, A. P. 1997, Science, 276, 1836
Buchner, J. 2016, Statistics and Computing, 26, 383
Cameron, A. G. W. 1978, Moon and Planets, 18, 5
Czekala, I., Andrews, S. M., Mandel, K. S., Hogg, D. W., & Green, G. M. 2015, ApJ, 812, 128
de Regt, S., Gandhi, S., Snellen, I. A. G., et al. 2024, A&A, 688, A116
de Regt, S., Snellen, I. A. G., Allard, N. F., et al. 2025, A&A, 696, A225
Denis, A., Vigan, A., Costes, J., et al. 2025, A&A, 696, A6
Dupuy, T. J. & Liu, M. C. 2011, ApJ, 733, 122
Faherty, J. K., Burningham, B., Gagné, J., et al. 2024, Nature, 628, 511
Fegley, Jr., B. & Lodders, K. 1996, ApJ, 472, L37
Filippazzo, J. C., Rice, E. L., Faherty, J., et al. 2015, ApJ, 810, 158
Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2021, A&A, 649, A1
Gaidos, E., Mann, A. W., Lépine, S., et al. 2014, MNRAS, 443, 2561
Gibson, N. P., Aigrain, S., Roberts, S., et al. 2012, MNRAS, 419, 2683
Gully-Santiago, M. & Morley, C. V. 2022, ApJ, 941, 200
Hojjatpanah, S., Figueira, P., Santos, N. C., et al. 2019, A&A, 629, A80
Houllé, M., Vigan, A., Carlotti, A., et al. 2021, A&A, 652, A67
Ida, S. & Lin, D. N. C. 2008, ApJ, 673, 487
JWST TECERST, Ahrer, E.-M., Alderson, L., et al. 2023, Nature, 614, 649
Kiman, R., Beichman, C. A., Ruiz Diaz, A., et al. 2026, AJ, 171, 60
Kipping, D. & Benneke, B. 2025, arXiv e-prints, arXiv:2506.05392
Kirkpatrick, J. D., Cushing, M. C., Gelino, C. R., et al. 2011, ApJS, 197, 19
Kirkpatrick, J. D., Martin, E. C., Smart, R. L., et al. 2019, ApJS, 240, 19
Kuiper, G. P. 1951, Proceedings of the National Academy of Science, 37, 1
Lacy, B. & Burrows, A. 2023, ApJ, 950, 8
Landman, R., Stolker, T., Snellen, I. A. G., et al. 2024, A&A, 682, A48
Leggett, S. K., Tremblin, P., Phillips, M. W., et al. 2021, ApJ, 918, 11
Line, M. R., Brogi, M., Bean, J. L., et al. 2021, Nature, 598, 580
Madhusudhan, N. 2019, ARA&A, 57, 617
Marley, M. S., Saumon, D., Visscher, C., et al. 2021, ApJ, 920, 85
Marocco, F., Kirkpatrick, J. D., Schneider, A. C., et al. 2024, ApJ, 967, 147
Meisner, A. M., Leggett, S. K., Logsdon, S. E., et al. 2023, AJ, 166, 57
Miller, N. & Fortney, J. J. 2011, ApJ, 736, L29
Mollière, P., Kühnle, H., Matthews, E. C., et al. 2025, A&A, 703, A79
Mollière, P., Wardenier, J. P., van Boekel, R., et al. 2019, A&A, 627, A67
Morley, C. V., Mukherjee, S., Marley, M. S., et al. 2024, ApJ, 975, 59
Morley, C. V., Skemer, A. J., Allers, K. N., et al. 2018, ApJ, 858, 97
Mukherjee, S., Batalha, N. E., Fortney, J. J., & Marley, M. S. 2023, ApJ, 942, 71
Mukherjee, S., Fortney, J. J., Batalha, N. E., et al. 2022, ApJ, 938, 107
Mukherjee, S., Fortney, J. J., Morley, C. V., et al. 2024, ApJ, 963, 73
Padoan, P. & Nordlund, Å. 2002, ApJ, 576, 870
Petrus, S., Chauvin, G., Bonnefoy, M., et al. 2025, A&A, 701, A208
Petrus, S., Whiteford, N., Patapis, P., et al. 2024, ApJ, 966, L11
Phillips, M. W., Tremblin, P., Baraffe, I., et al. 2020, A&A, 637, A38
Piette, A. A. A. & Madhusudhan, N. 2020, MNRAS, 497, 5136
Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, Icarus, 124, 62
Ravet, M., Bonnefoy, M., Chauvin, G., et al. 2025, A&A, 704, A325
Rotman, Y., Welbanks, L., Line, M. R., et al. 2025, ApJ, 989, 201
Ruffio, J.-B., Macintosh, B., Konopacky, Q. M., et al. 2019, AJ, 158, 200
Ruffio, J.-B., Xuan, J. W., Chachan, Y., et al. 2026, Nature Astronomy [arXiv:2601.08227]
Schneider, A. C., Shkolnik, E. L., Allers, K. N., et al. 2019, AJ, 157, 234
Schneider, A. D. & Bitsch, B. 2021, A&A, 654, A71
Sellke, T., Bayarri, M. J., & Berger, J. O. 2001, The American Statistician, 55, 62, publisher: [American Statistical Association, Taylor & Francis, Ltd.]
Skilling, J. 2004, in American Institute of Physics Conference Series, Vol. 735, Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, ed. R. Fischer, R. Preuss, & U. V. Toussaint (AIP), 395–405
Spiegel, D. S. & Burrows, A. 2012, ApJ, 745, 174
Spiegelhalter, D., Best, N., Carlin, B., & {Van Der Linde}, A. 2002, Journal of the Royal Statistical Society. Series B: Statistical Methodology, 64, 583
Suárez, G., Vos, J. M., Metchev, S., Faherty, J. K., & Cruz, K. 2023, ApJ, 954, L6
Tan, X. & Showman, A. P. 2021, MNRAS, 502, 2198
Thorngren, D. & Fortney, J. J. 2019, ApJ, 874, L31
Thorngren, D. P., Sing, D. K., & Mukherjee, S. 2026, ApJS, 283, 10
Torres, C. A. O., Quast, G. R., da Silva, L., et al. 2006, A&A, 460, 695
Trotta, R. 2008, Contemporary Physics, 49, 71
Visscher, C., Lodders, K., & Fegley, Jr., B. 2006, ApJ, 648, 1181
Visscher, C., Moses, J. I., & Saslow, S. A. 2010, Icarus, 209, 602
Vorobyov, E. I. 2013, A&A, 552, A129
Voyer, M., Changeat, Q., Lagage, P.-O., et al. 2025, ApJ, 982, L38
Wallack, N. L., Batalha, N. E., Alderson, L., et al. 2024, AJ, 168, 77
Wang, J. J., Ginzburg, S., Ren, B., et al. 2020, AJ, 159, 263
Wogan, N. F., Mang, J., Batalha, N. E., et al. 2025, Research Notes of the American Astronomical Society, 9, 108
Xuan, J. W., Perrin, M. D., Mawet, D., et al. 2024, ApJ, 977, L32
Zhang, Z., Liu, M. C., Claytor, Z. R., et al. 2021a, ApJ, 916, L11
Zhang, Z., Liu, M. C., Hermes, J. J., et al. 2020, ApJ, 891, 171
Zhang, Z., Liu, M. C., Marley, M. S., Line, M. R., & Best, W. M. J. 2021b, ApJ, 921, 95
Zhang, Z., Mollière, P., Hawkins, K., et al. 2023, AJ, 166, 198
Zhang, Z., Mukherjee, S., Liu, M. C., et al. 2025, AJ, 169, 9
Zuckerman, B. & Song, I. 2009, A&A, 493, 1149

[1] Laboratoire J.-L. Lagrange, Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, 06304 Nice, France
[2] IPAG, Université Grenoble-Alpes, CNRS, F-38000 Grenoble, France
[3] Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany
[4] Department of Physics & Astronomy, University of Rochester, Rochester, NY 14627, USA
[5] Department of Astrophysics, American Museum of Natural History, New York, NY 10024, USA
[6] Université Paris Cité, Université Paris-Saclay, CEA, CNRS, AIM, F-91191 Gif-sur-Yvette, France
[7] Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh, EH9 3HJ, UK
[8] Department of Astronomy, California Institute of Technology, Pasadena, CA 91125, USA
[9] Department of Physics, Astronomy and Mathematics, University of Hertfordshire, Hatfield, UK
[10] Department of Astronomy, University of Texas at Austin, Austin, TX 78712, USA
[11] Institute of Particle Physics and Astrophysics, ETH Zürich, Wolfgang-Pauli-Str 27, 8049 Zürich Switzerland
[12] LIRA, Observatoire de Paris, Univ. PSL, CNRS, Sorbonne Universit´e, Univ. Paris Diderot, Sorbonne Paris Cit´e, 5 place Jules Janssen, 92195 Meudon, France
[13] Leiden Observatory, Leiden University, P.O. Box 9513, 2300 RA, Leiden, The Netherlands
[14] NASA-Goddard Space Flight Center, Greenbelt, MD 20771, USA
[15] Instituto de Estudios Astrofísicos, Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército Libertador 441, Santiago, Chile
[16] Millennium Nucleus on Young Exoplanets and their Moons (YEMS), Santiago, Chile
[17] Aix Marseille Univ, CNRS, CNES, LAM, Marseille, France
[18] Department of Physics & Astronomy, Johns Hopkins University, Baltimore, MD, 21218, USA
[19] Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD 21218, USA
[20] Fakultät für Physik, Universität Duisburg-Essen, Lotharstraße 1, 47057 Duisburg, Germany
[21] Physikalisches Institut, Universität Bern, Gesellschaftsstr. 6, CH-3012 Bern, Switzerland
[22] AURA for the European Space Agency (ESA), ESA Ofcice, Space Telescope Science Institute, 3700 San Martin Drive, Baltimore, MD, 21218 USA
[23] European Southern Observatory, Alonso de Córdova 3107, Casilla 19, Santiago 19001, Chile

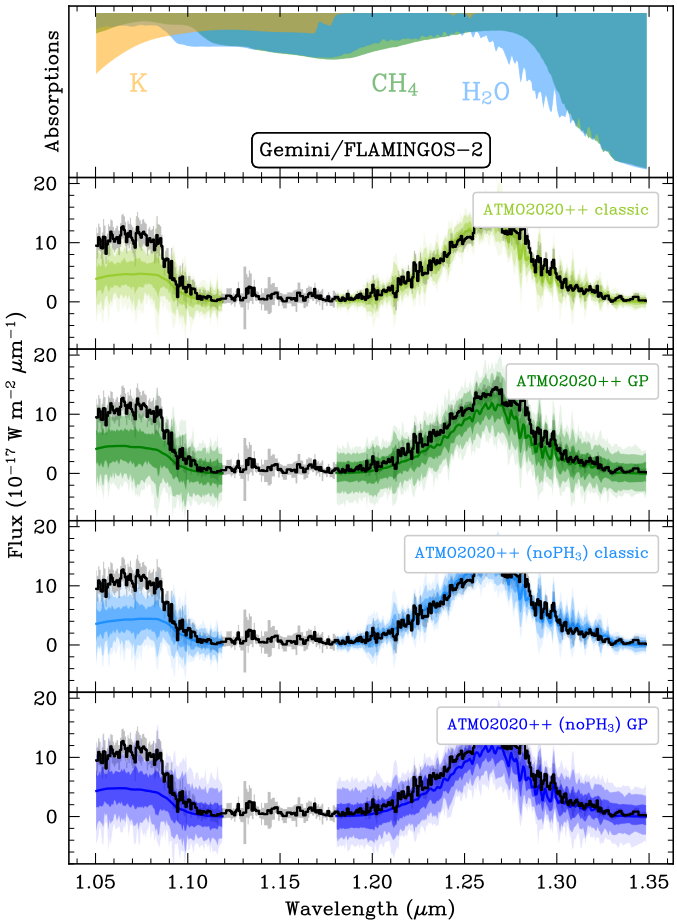

Fig. A.1. Same as Fig. 4, but with Gemini/FLAMINGOS-2. Zoom-in on the Y- and J-bands.

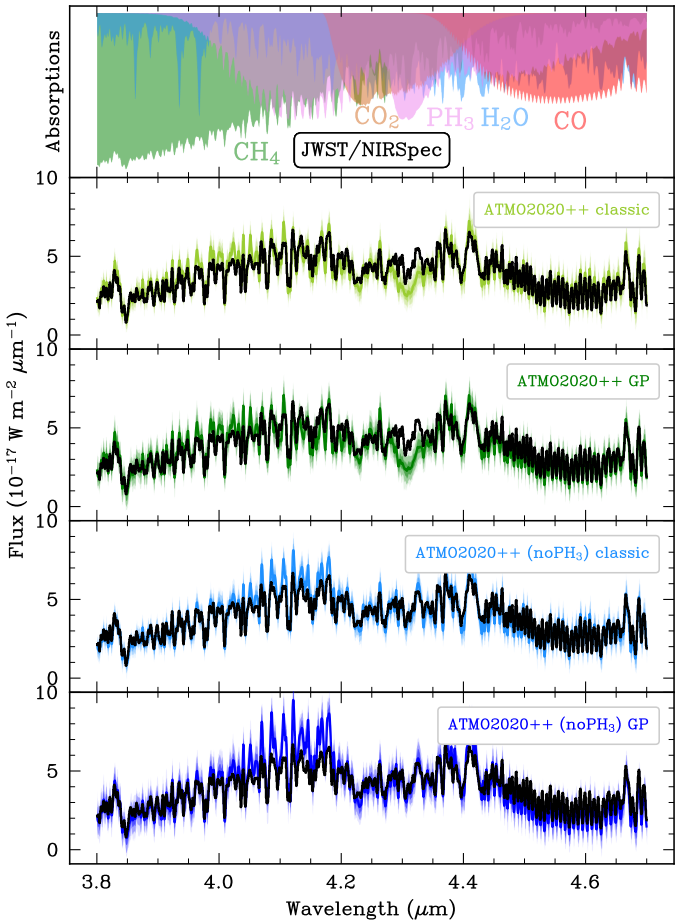

Fig. A.2. Same as Fig. 4, but with JWST/NIRSpec. Zoom-in on the 3.8–4.7 μm region.

Appendix A: Additional figures and tables

Table A.1. Noise scaling factors ($\hat{s}_i = \frac{\chi^2_i}{N_i}$) from the classical inversions of Sect. 3.1.

Model	$\hat{s}_{\text{FLAMINGOS-2}}$	$\hat{s}_{\text{NIRSpec}}$	$\hat{s}_{\text{MIRI}}$
BT-Settl	1.63	118.44	48715.21
Sonora Diamondback	1.24	70.91	33987.07
Sonora Elf Owl	1.00	21.97	18858.41
ATMO2020++ (no PH_3)	1.03	20.84	2422.83
ATMO2020++	1.04	19.83	11812.63

Appendix B: Models

We used five atmospheric grids with different physical and chemical makeup :

- BT-Settl (Allard et al. 2012) is a grid that explores complex cloud microphysics and includes non-equilibrium chemistry as well as vertical mixing. Clouds are simulated by dividing the atmosphere into multiple layers and computing the distribution and size of grains by comparing their characteristic times of condensation, coalescence, dispersion and sedimentation.
- Sonora Diamondback (Morley et al. 2024) model grid includes vertical mixing and parameterized clouds. Clouds are parameterized using Ackerman & Marley (2001) approach with the sedimentation efficiency parameter f_{sed}. This model assumes radiative–convective and chemical equilibrium. We used a custom version of this grid extended at lower temperature by rerunning the full forward model.
- Sonora Elf Owl (Mukherjee et al. 2024) is a cloudless grid that includes vertical mixing induced disequilibrium chemistry with subsolar to supersolar [M/H] and C/O. The atmospheric models have been computed using the open-source radiative-convective equilibrium model *PICASO*[9] (Batalha et al. 2019; Marley et al. 2021; Mukherjee et al. 2023; JWST TECERST et al. 2023). We used version v.2[10] of the grid with a correction to the disequilibrium abundance of CO_2 and no PH_3 (Wogan et al. 2025).
- ATMO2020++ and ATMO2020++ (no PH_3) (Phillips et al. 2020; Leggett et al. 2021; Meisner et al. 2023) are extensions of the ATMO2020 model which additionally incorporate a non-adiabatic thermal structure of the atmospheres. These models are cloud-free but uses non-equilibrium chemical reactions of CO-CH_4 and N_2-NH_3, driving fingering convection in the atmosphere and changing the temperature gradient, to emulate their reddening effect.

[9] https://natashabatalha.github.io/picaso/
[10] https://zenodo.org/records/15150865

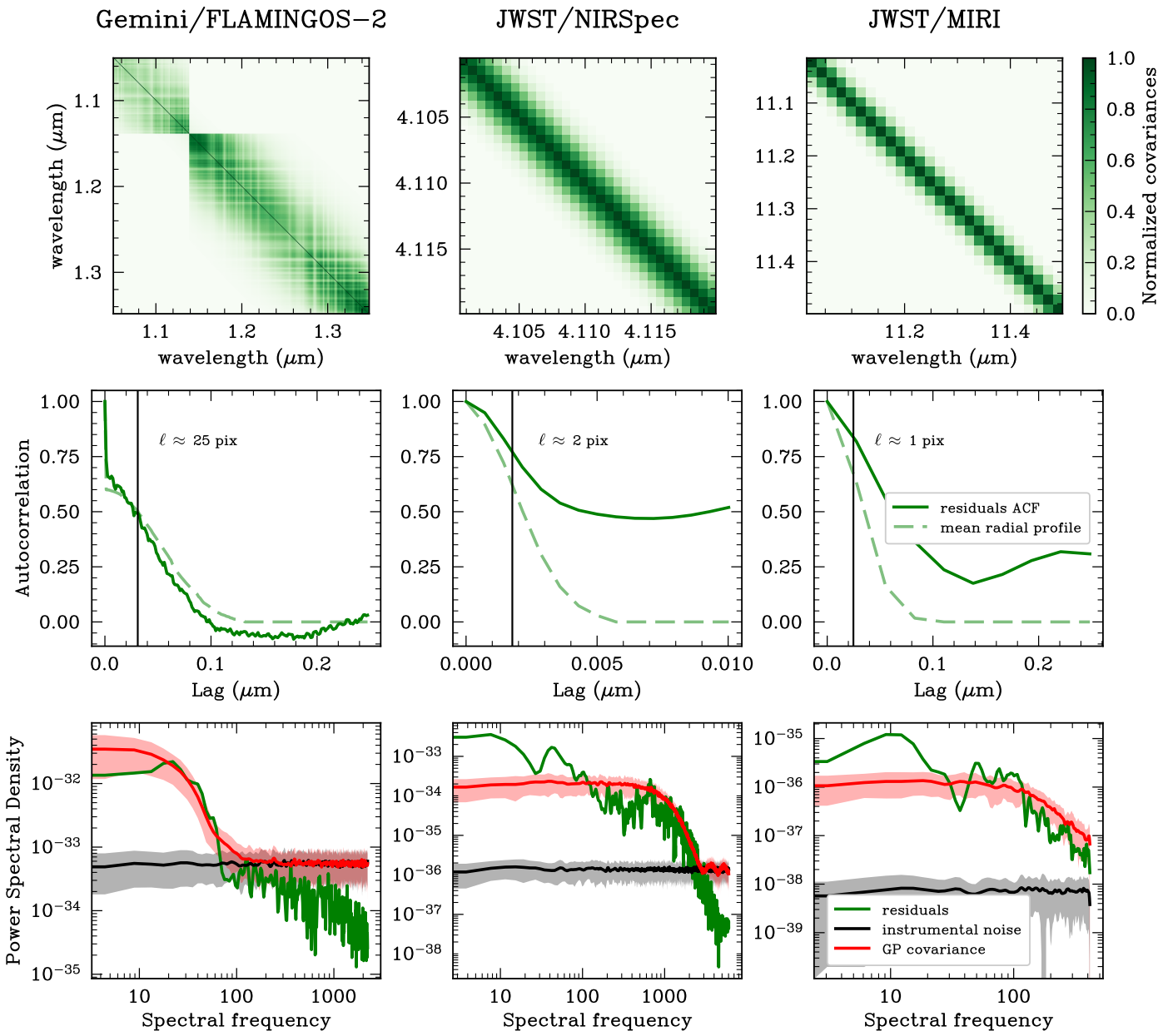

Fig. A.3. GP diagnostics. *Top row:* Zoom-in on the retrieved (normalized) covariance matrices from the GP-aided inversion using the ATMO2020++ model. The block-diagonal structure reflects the discontinuous spectral windows of the Gemini/FLAMINGOS-2 observations (see Sect. 3). *Middle row:* Autocorrelation of the residuals (data – model) in green, with the corresponding GP mean radial profile overplotted as a dotted green line. The vertical black lines indicate the GP correlation length-scales. *Bottom row:* Power spectral densities (PSD) of the residuals compared with random realizations from the injected instrumental noise (blue, $\pm 1\sigma$ shaded) and from the fitted GP covariance (red, $\pm 1\sigma$ shaded).

Appendix C: GP injection test

This section describes the injection test we performed to probe the performance of the newly implemented GP.

To ensure that the correlation amplitude and length scale retrieved during our inversion are consistent with the correlated noise present in real spectra, we constructed a synthetic Gemini/FLAMINGOS-2 spectrum. This mock spectrum was interpolated from the ATMO2020++ grid using atmospheric parameters of $T_{\rm eff}$ = 483 K, log(g) = 4.19, and [M/H] = 0.0 (see Table 1).

The model spectrum was convolved and sampled at a spectral resolution of $R_\lambda \sim 900$ to match that of FLAMINGOS-2, resulting in a noise-free spectrum $\boldsymbol{d}$. We then simulated observational noise using two components, as described by

$$\boldsymbol{d}_{\rm noisy} = \boldsymbol{d} + \boldsymbol{n}_{\rm sky} + \boldsymbol{n}_{\rm inst}, \tag{C.1}$$

where $\boldsymbol{n}_{\rm sky}$ and $\boldsymbol{n}_{\rm inst}$ represent the sky and instrumental noise contributions, respectively, and $\boldsymbol{d}_{\rm noisy}$ is the final synthetic noisy spectrum.

We model the sky noise as uncorrelated Gaussian noise, assuming an average S/N of 10

$$\boldsymbol{n}_{\rm sky} \sim \mathcal{N}\left(0,\ \sigma^2\right), \quad \text{with} \quad \sigma^2 = \left(\frac{\bar{d}}{\rm S/N}\right)^2, \tag{C.2}$$

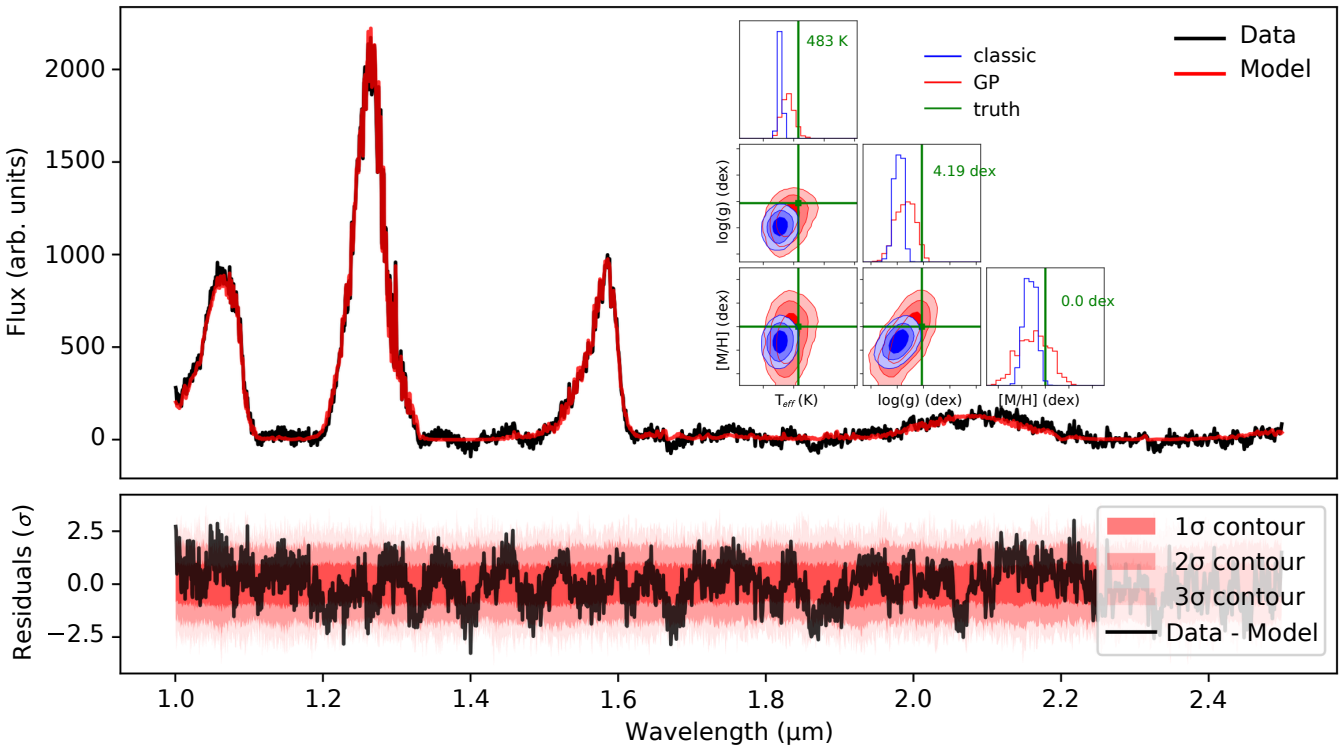

Fig. A.4. Injection test summary figure. *Top panel:* Comparison between the mock spectrum (black) and the fitted GP-aided model (red). *Top subpanel:* Corner plot of the grid parameters from the inversions with (red) and without (blue) GP. Green lines indicate the injected values. *Bottom panel:* Normalized residuals between the mock data and the fitted GP-aided model (black). Shaded regions represent the 1σ, 2σ, and 3σ dispersions from 200 random draws from the covariance matrix C.

Table A.2. COCONUTS-2b properties derived from the different evolutionary models explored.

Model	T_{eff} (K)	log(g) (dex)	R (R_{Jup})	M (M_{Jup})
ATMO2020 CEQ	493 ± 1	4.18 ± 0.02	1.103 ± 0.004	7.4 ± 0.2
ATMO2020 NEQ weak	493 ± 1	4.18 ± 0.02	1.103 ± 0.004	7.4 ± 0.2
ATMO2020 NEQ strong	493 ± 1	4.18 ± 0.02	1.103 ± 0.004	7.4 ± 0.2
Sonora Bobcat [M/H] = -0.5 dex	491 ± 1	4.18 ± 0.02	1.109 ± 0.004	7.6 ± 0.2
Sonora Bobcat [M/H] = 0.0 dex	488 ± 1	$4.16^{+0.01}_{-0.02}$	1.123 ± 0.004	7.3 ± 0.2
Sonora Bobcat [M/H] = +0.5 dex	487 ± 1	4.13 ± 0.02	1.128 ± 0.004	7.0 ± 0.2
Sonora Diamondback "hybrid" [M/H] = -0.5 dex	493 ± 1	4.18 ± 0.02	1.105 ± 0.004	7.5 ± 0.2
Sonora Diamondback "hybrid" [M/H] = 0.0 dex	489 ± 1	4.15 ± 0.02	$1.121^{+0.004}_{-0.003}$	7.2 ± 0.2
Sonora Diamondback "hybrid" [M/H] = +0.5 dex	488 ± 1	4.13 ± 0.02	1.130 ± 0.004	6.9 ± 0.2
Sonora Diamondback "hybrid-grav" [M/H] = -0.5 dex	493 ± 1	4.18 ± 0.02	1.105 ± 0.004	7.4 ± 0.2
Sonora Diamondback "hybrid-grav" [M/H] = 0.0 dex	487 ± 1	4.14 ± 0.02	$1.131^{+0.004}_{-0.003}$	7.2 ± 0.2
Sonora Diamondback "hybrid-grav" [M/H] = +0.5 dex	480 ± 1	4.11 ± 0.02	1.164 ± 0.004	7.0 ± 0.2
Concatenated	490^{+3}_{-4}	$4.16^{+0.03}_{-0.04}$	$1.12^{+0.02}_{-0.01}$	7.3 ± 0.3

Notes. "Concatenated" refers to the values and constraints obtained by merging the parameter chains after the Monte Carlo exploration.

where $\bar{d}$ = mean($\boldsymbol{d}$). This component is the only one considered in the injected error bars.

To simulate the effect of correlated instrumental noise, we introduce a second component modeled with a GP characterized by a correlation length $\ell = 10^{-2}$ μm and amplitude $a = \sigma$ (i.e., strong correlated noise). This is represented by

$$\boldsymbol{n}_{\rm inst} \sim \mathcal{N}(0,\ \boldsymbol{C}), \quad \text{with} \quad C_{ij} = a^2 \exp\left[-\frac{\Delta\lambda_{i,j}^2}{2\ell^2}\right], \tag{C.3}$$

where $\Delta\lambda_{i,j}$ denotes the wavelength separation between data points i and j. Finally, we inverted this mock data with the

Table A.3. Inversion results of Sect. 3.

Parameter	ln $\mathcal{B}$	T_{eff}	log(g)	[M/H]	C/O	f_{sed}	$\log(K_{zz})$	R	$\log(L/L_\odot)$
Units		(K)	(dex)	(dex)			$\log(cm^2.s^{-1})$	(R_{Jup})	(dex)
BT-Settl priors		$U(200, 1000)$	$U(3.5, 4.5)$					$U(0, 2)$	
BT-Settl posteriors									
classic	7643	447^{+3}_{-2}	> 4.49	(0.00)	(0.55)	(microphys.)	(profile)	1.03 ± 0.01	$-6.180^{+0.001}_{-0.002}$
GP	1219	463 ± 3	$4.48^{+0.01}_{-0.02}$	(0.00)	(0.55)	(microphys.)	(profile)	0.96 ± 0.01	-6.190 ± 0.003
Sonora Diamondback priors		$U(400, 600)$	$U(3.5, 5.5)$	$U(-0.5, 0.5)$		$U(1, 8)$		$U(0, 2)$	
Sonora Diamondback posteriors									
classic	6645	> 600	$5.000^{+0.003}_{-0.002}$	0.194 ± 0.004	(0.458)	$3.6^{+0.1}_{-0.2}$	(profile)	0.59 ± 0.01	-6.214 ± 0.006
GP	732	515 ± 1	$4.500^{+0.003}_{-0.004}$	-0.43 ± 0.01	(0.458)	> 7.1	(profile)	$1.65^{+0.02}_{-0.03}$	-5.71 ± 0.01
Sonora Elf Owl priors		$U(400, 600)$	$U(3.25, 5.5)$	$U(-1.0, 1.0)$	$U(0.229, 1.145)$		$U(2, 9)$	$U(0, 2)$	
Sonora Elf Owl posteriors									
classic	4593	517^{+1}_{-2}	< 3.25	-0.33 ± 0.01	0.317 ± 0.005		$3.99^{+0.02}_{-0.03}$	0.84 ± 0.01	-6.199 ± 0.002
GP	0	549^{+2}_{-3}	< 3.25	-0.30 ± 0.02	0.42 ± 0.01		3.50 ± 0.07	0.81 ± 0.01	-6.190 ± 0.005
ATMO2020++ (no PH_3) priors		$U(250, 1200)$	$U(2.5, 5.5)$	$U(-1.0, 0.3)$				$U(0, 2)$	
ATMO2020++ (no PH_3) posteriors									
classic	4401	538 ± 2	4.94 ± 0.02	0.25 ± 0.01	(0.55)		(5.12 ± 0.04)	0.77 ± 0.01	-6.202 ± 0.002
GP	319	493 ± 3	4.14 ± 0.02	-0.09 ± 0.02	(0.55)		(6.72 ± 0.04)	1.01 ± 0.01	-6.186 ± 0.002
ATMO2020++ priors		$U(250, 1200)$	$U(2.5, 5.5)$	$U(-1.0, 0.3)$				$U(0, 2)$	
ATMO2020++ posteriors									
classic	4262	514 ± 1	4.80 ± 0.01	> 0.30	(0.55)		(5.40 ± 0.02)	0.93 ± 0.01	-6.179 ± 0.002
GP	94	496^{+5}_{-3}	$4.30^{+0.04}_{-0.02}$	$-0.02^{+0.03}_{-0.02}$	(0.55)		$(6.40^{+0.04}_{-0.08})$	$1.03^{+0.01}_{-0.02}$	-6.163 ± 0.002

Notes. *First column:* log-Bayes factors relative to the Sonora Elf Owl (GP). *Remaining columns:* Grid priors and posteriors. $U(a, b)$ refers to a uniform distribution between a and b. The error bars correspond to the lower and upper bounds in the parameter space encompassing 68% of the retrieved solutions around the best fit. Luminosities were computed using Eq. (3) from Zhang et al. (2025) to account for model–data discrepancies. Values inside parentheses are fixed by the model grids. In particular, the $\log(K_{zz})$ values of the ATMO2020++ and ATMO2020++ (no PH_3) modeling results are linearly interpolated from the inferred log(g) using Fig. 1 of Phillips et al. (2020).

Table A.4. GP hyperparameters priors and posteriors.

Parameter	$\log(a)_{FLAMINGOS-2}$	$\log(a)_{NIRSpec}$	$\log(a)_{MIRI}$	$\log(l)_{FLAMINGOS-2}$	$\log(l)_{NIRSpec}$	$\log(l)_{MIRI}$
Units				$\log(\mu m)$	$\log(\mu m)$	$\log(\mu m)$
priors	$U(-0.5, 3)$	$U(-0.5, 3)$	$U(-0.5, 3)$	$U(-4, 0)$	$U(-4, 0)$	$U(-4, 0)$
BT-Settl posteriors	0.01 ± 0.05	0.901 ± 0.001	2.23 ± 0.02	-1.86 ± 0.03	-2.691 ± 0.001	$-1.701^{+0.001}_{-0.002}$
Sonora Diamondback posteriors	$0.33^{+0.02}_{-0.03}$	0.85 ± 0.01	1.62 ± 0.02	-1.65 ± 0.01	-2.725 ± 0.001	$-1.676^{+0.003}_{-0.004}$
Sonora Elf Owl posteriors	$0.25^{+0.05}_{-0.06}$	0.60 ± 0.01	1.89 ± 0.02	$-1.56^{+0.01}_{-0.06}$	$-2.657^{+0.001}_{-0.002}$	$-1.695^{+0.002}_{-0.003}$
ATMO2020++ (no PH_3) posteriors	$0.10^{+0.07}_{-0.06}$	0.75 ± 0.01	1.28 ± 0.02	$-1.48^{+0.08}_{-0.06}$	-2.759 ± 0.001	-1.612 ± 0.003
ATMO2020++ posteriors	0.09 ± 0.07	0.70 ± 0.01	1.24 ± 0.02	$-1.51^{+0.08}_{-0.04}$	-2.753 ± 0.001	-1.606 ± 0.002

ATMO2020++ using 1000 living points and uniform priors, with (in red) and without GP (in blue). Results are summarized in Fig. A.4 and Table A.5.

The top panel of Fig. A.4 displays the fitted spectrum alongside the corner plots for both inversions. When spectral covariances are omitted, *ForMoSA* fails to recover the injected values, and the posterior uncertainties are noticeably underestimated across all parameters. In contrast, incorporating a GP model improves performance: although some bias remains, the true values lie within the broader posterior distributions. Additionally, the GP-aided inversion successfully retrieves the covariance hyperparameters, yielding final estimates of $a = 1.0 \pm 0.1$ (true value: 1) and $\ell = 10.7^{+0.7}_{-0.1}$ nm (true value: 10 nm). The bottom panel of Fig. A.4 presents the normalized residuals between the data and the model for the GP-aided inversion, clearly revealing correlated structures (in black). The red shaded contours represent the 1σ, 2σ, and 3σ dispersions of an ensemble of 200 random draws from the final (fitted) covariance matrix $\boldsymbol{C}$. The inclusion of the GP kernel effectively captures both the structure and amplitude of the residuals.

Appendix D: GP with additional atmospheric models

In Sect. 3.2, we focused our analysis on the two ATMO2020++ grids. Here, we provide an overview of the results obtained using the GP framework with the other grids (namely BT-Settl, Sonora Diamondback, and Sonora Elf Owl).

Table A.5. Inversion results of Appendix C.

Parameter	$\ln \mathcal{B}$	T_{eff}	log(g)	[M/H]	log(a)	log(l)
Units		(K)	(dex)	(dex)		log(μm)
priors		$U(250, 1200)$	$U(2.5, 5.5)$	$U(-1.0, 0.3)$	$U(-0.5, 2)$	$U(-4, 0)$
posteriors classic	229	471 ± 1	4.06 ± 0.03	-0.03 ± 0.01		
posteriors GP	0	476 ± 4	$4.10^{+0.05}_{-0.06}$	-0.02 ± 0.03	0.01 ± 0.05	$-1.97^{+0.03}_{-0.05}$
injected		483	4.19	0.0	0.0	-2.0

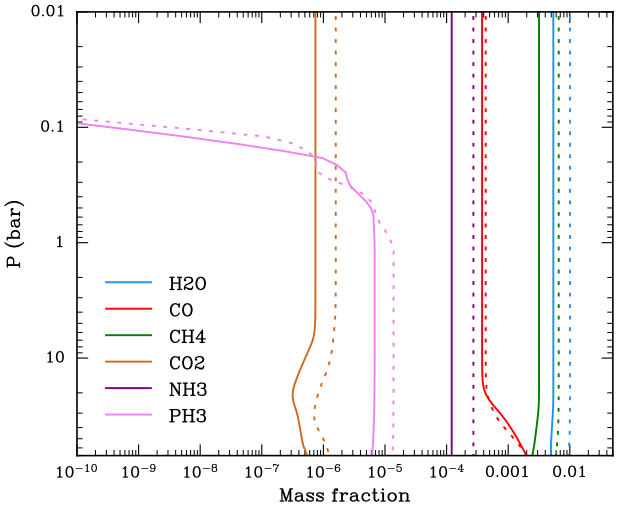

Fig. A.5. Key species mass fractions interpolated from ATMO2020++. The dashed lines correspond to the classical inversion, while the solid lines correspond to the GP-aided inversion.

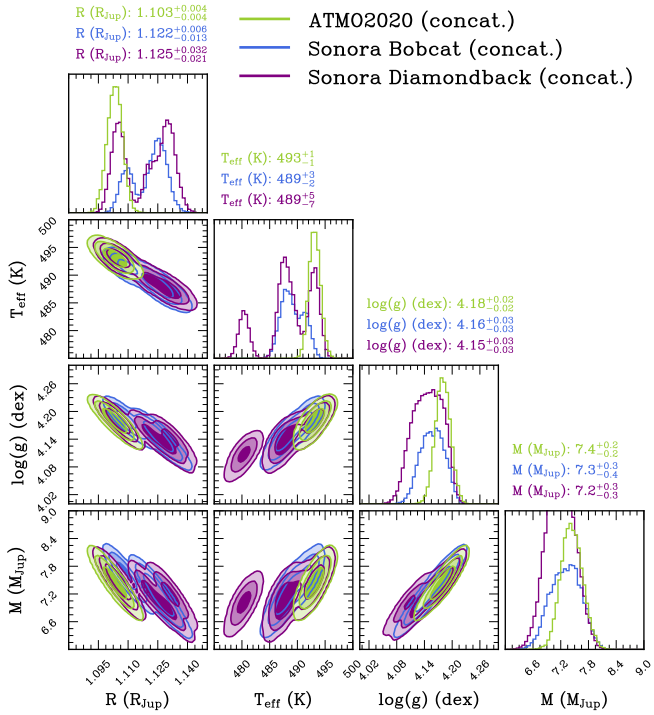

Fig. A.6. Corner plot comparing the three families of evolutionary models (ATMO2020 in green, Sonora Bobcat in blue, and Sonora Diamondback in purple). The secondary bumps in the Sonora distributions usually corresponds to the different [M/H] nodes.

According to all statistical criteria adopted in this work, Sonora Elf Owl provides the preferred fit among the tested models (see Table 4). Nevertheless, the overall quality of the fit remains poor, particularly in its ability to reproduce the MIRI observations. In addition, the retrieved atmospheric parameters, most notably T_{eff} (549^{+2}_{-3} K) and log(g) (< 3.25 dex), remain inconsistent with the predictions of all evolutionary models explored in this study (see Table A.2).

On the other hand, the GP significantly improves the consistency of the retrieved T_{eff} and log(g) for BT-Settl and Sonora Diamondback, with $T_{eff} = 463 \pm 3$ K and $\log(g) = 4.48^{+0.01}_{-0.02}$ dex for BT-Settl, and $T_{eff} = 515 \pm 1$ K and $\log(g) = 4.500^{+0.003}_{-0.004}$ dex for Sonora Diamondback.

Similar to ATMO2020++, the metallicity is found to be subsolar for all grids that explore this parameter, with [M/H] = -0.43 ± 0.01 dex and [M/H] = -0.30 ± 0.02 dex for Sonora Diamondback and Sonora Elf Owl, respectively. Contrary to the values retrieved with ATMO2020++ in this setup (see Sect. 4.1), these metallicities are not consistent with the stellar metallicity of 0.00 ± 0.08 dex (Hojjatpanah et al. 2019). Similarly, the C/O ratio remains subsolar, with C/O = 0.42 ± 0.01 retrieved using Sonora Elf Owl.

The vertical mixing inferred with Sonora Elf Owl ($K_{zz} = (3.2^{+0.6}_{-0.5}) \times 10^3$ cm^2 s^{-1}) remains significantly lower than that predicted by the two ATMO2020++ grids ($K_{zz} = (2.5^{+0.2}_{-0.4}) \times 10^6$ cm^2 s^{-1} and $K_{zz} = (5.3^{+0.5}_{-0.5}) \times 10^6$ cm^2 s^{-1}, with and without PH_3, respectively).

With the exception of the BT-Settl grid, the inferred GP correlation lengths are remarkably consistent across the different atmospheric models (see Table A.4). In contrast, we observe larger variations in the GP correlation amplitudes. This trend suggests that the correlation lengths may be primarily driven by each instrument's correlated noise pattern (see Sect. 4.2), while the correlation amplitudes may instead more closely reflect the GP compensating for each specific model mismatch.

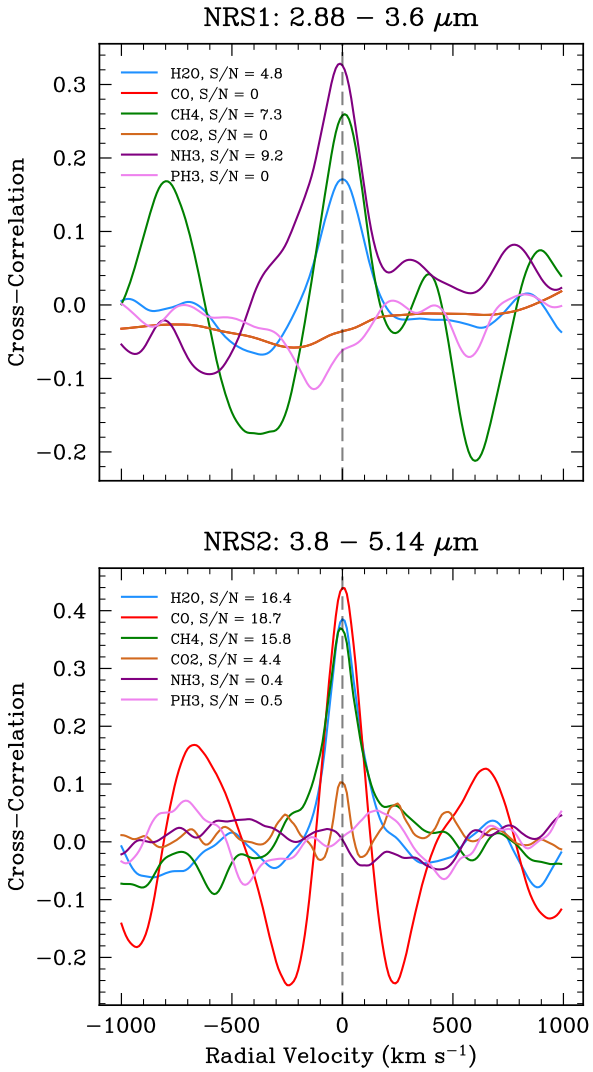

Fig. A.7. Cross-correlation functions of COCONUTS-2 b NIRSpec spectra (at full resolution) with H_2O, CO, CH_4, CO_2, NH_3, and PH_3 molecular templates generated using *petitRADTRANS*. The S/N was computed using formula (1) of Houllé et al. (2021). *Upper panel:* CCFs obtained using the NRS1 detector where H_2O, CH_4, and NH_3 are detected. *Lower panel:* CCFs obtained using the NRS2 detector where H_2O, CO, CH_4, and CO_2 are detected.