

# Comparing transit spectroscopy pipelines at the catalogue level: evidence for systematic differences

Lorenzo V. Mugnai <sup>1,2,3,4</sup>★ Mark R. Swain,<sup>5</sup> Raissa Estrela<sup>5</sup> and Gael M. Roudier<sup>5</sup>

<sup>1</sup>*School of Physics and Astronomy, Cardiff University, Queens Buildings, The Parade, Cardiff CF24 3AA, UK*

<sup>2</sup>*Dipartimento di Fisica, La Sapienza Università di Roma, Piazzale Aldo Moro 2, I-00185 Roma, Italy*

<sup>3</sup>*INAF – Osservatorio Astronomico di Palermo, Piazza del Parlamento 1, I-90134 Palermo, Italy*

<sup>4</sup>*Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK*

<sup>5</sup>*Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, USA*

Accepted 2024 April 15. Received 2024 April 9; in original form 2023 October 11

## ABSTRACT

The challenge of inconsistent results from different data pipelines, even when starting from identical data, is a recognized concern in exoplanetary science. As we transition into the *JWST* era and prepare for the *ARIEL* space mission, addressing this issue becomes paramount because of its implications on our understanding of exoplanets. Although comparing pipeline results for individual exoplanets has become more common, this study is the first to compare pipeline results at the catalogue level. We present a comprehensive framework to statistically compare the outcomes of data analysis reduction on a population of exoplanets and we leverage the large number of observations conducted using the same instrument configured with *HST*-WFC3. We employ three independent pipelines: IRACLIS, EXCALIBUR, and CASCADE. Our combined findings reveal that these pipelines, despite starting from the same data and planet system parameters, yield substantially different spectra in some cases. However, the most significant manifestations of pipeline differences are observed in the compositional trends of the resulting exoplanet catalogues. We conclude that pipeline-induced differences lead to biases in the retrieved information, which are not reflected in the retrieved uncertainties. Our findings underscore the critical need to confront these pipeline differences to ensure the reproducibility, accuracy, and reliability of results in exoplanetary research. Our results demonstrate the need to understand the potential for population-level bias that pipelines may inject, which could compromise our understanding of exoplanets as a class of objects.

**Key words:** techniques: spectroscopic – software: data analysis – infrared: planetary systems.

## 1 INTRODUCTION

A successful methodology for detecting atomic and molecular species and unveiling the atmospheric chemistry of exoplanets involves the use of multiband transit photometry and spectroscopy (e.g. Charbonneau et al. 2002; Tinetti et al. 2007; Swain, Vasist & Tinetti 2008; Swain et al. 2009, 2021; Tsiaras et al. 2016a; Chachan et al. 2019; Mugnai et al. 2021a). Current space instrumentation, such as the *Spitzer Space Telescope* and *Hubble Space Telescope (HST)*, have facilitated the atmospheric characterization of approximately 60 exoplanets over a limited wavelength range (e.g. Iyer et al. 2016; Sing et al. 2016; Barstow et al. 2017; Tsiaras et al. 2018; Edwards et al. 2022; Estrela, Swain & Roudier 2022). However, these instruments were not specifically designed for exoplanetary science, necessitating specialized data reduction pipelines to remove instrument systematics that are similar in amplitude to the astrophysical signal (Deming et al. 2013; Tsiaras et al. 2016b).

To interpret the observed spectra, spectral retrieval techniques are commonly used to estimate astrophysical parameters (e.g. Irwin

et al. 2008; Madhusudhan & Seager 2009; Lee, Heng & Irwin 2013; Line et al. 2013; Waldmann et al. 2015; Gandhi & Madhusudhan 2017; Lavie et al. 2017; Al-Refaie et al. 2021). Studies have been conducted to compare and validate different retrieval models, demonstrating their robustness and consistency (Barstow et al. 2020, 2022). However, a similar in-depth large-scale validation has not been performed for data reduction pipelines, which estimate the spectra from raw data. Uncharacterized biases introduced at this stage of data analysis can potentially undermine the correct interpretation of observations using retrieval techniques. While the recent literature does chronicle multiple validation endeavours, these comparisons are undertaken on a singular-planet basis. A remarkable example of this trend is offered by the Early Release Science of *JWST*: Ahler et al. (2023), Alderson et al. (2023), Feinstein et al. (2023), Holmberg & Madhusudhan (2023), and Rustamkulov et al. (2023). Thus, there is a compelling need for holistic, population-centric validation, which is the cornerstone of our proposed study.

The data reduction process for exoplanet transit spectroscopy has a number of steps where differences in methods have the potential to produce differences in final outcomes. A non-exhaustive list of specific areas where method differences might influence the final

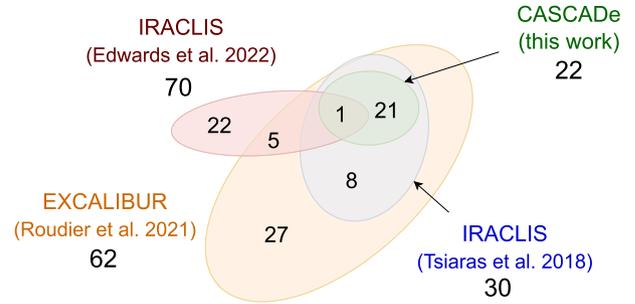
\* E-mail: [mugnail@cardiff.ac.uk](mailto:mugnail@cardiff.ac.uk)

outcome includes; spectral extraction and background subtraction, interpolation errors associated with placing spectra on a common wavelength grid, system parameter values, astrophysical models such as the detailed formulation of the limb-darkening relation, outlier rejection methods, the value or width of any priors applied, which parameters are locked and which are retrieved, the dimensionality and form of the instrument model, and the formulation of the sampler. Differences in the method can lead to differences in astrophysical interpretation (e.g. Swain et al. 2021; Mugnai et al. 2021a; Libby-Roberts et al. 2022) and raise the question of which result more accurately represents the astrophysical reality. Further reinforcing the notion that differences in methods can impact the astrophysical interpretation, the literature is replete with examples of exoplanet descriptions being revised due to different approaches to the modelling and removal of systematics in data reduction pipelines (e.g. Diamond-Lowe et al. 2014; Stevenson et al. 2014a,b; Tsiaras et al. 2018). A notable example is the hypothesis of six exomoon candidates proposed by Fox & Wiegert (2021), which was later discarded by an updated data reduction pipeline (Kipping 2020). Using different stellar or planetary parameters for the analysis is known to introduce offsets or slopes in the data set (Morello et al. 2017; Alexoudi et al. 2018). Finally, time variability can arise from stellar activity (Kirk et al. 2019; Bruno et al. 2020).

A further complication is the potential inconsistency between data sets from different instruments, which remains a problematic issue that has been tentatively discussed in the literature to assess its implications for our understanding of exoplanets (e.g. Pluriel et al. 2020; Yip et al. 2020; Saba et al. 2022). Different data analysis approaches have been suggested to take into account these discrepancies during retrieval (Yip et al. 2021), however, this issue warrants further research to ensure the reliability and reproducibility of exoplanetary science. In particular, with the anticipated contributions from the *JWST*, ensuring consistency across multi-instrument data sets will be paramount for the accurate interpretation and understanding of exoplanetary atmospheres (Constantinou, Madhusudhan & Gandhi 2023).

In fact, as we usher in the era of the *JWST* and the *ARIEL* space mission, the challenges posed by the possibility of pipeline-dependent biases operating on an entire observational catalogue must be taken seriously and understood. The advent of these next-generation telescopes promises unprecedented data quality, allowing researchers to delve into more intricate questions about exoplanets. However, the very richness of this data also amplifies the potential pitfalls of pipeline discrepancies. While this study highlights issues observed with *HST*-WFC3 data, the implications extend far beyond. In the *JWST* and *ARIEL* era, where the focus will be on planetary population studies and comparative planetology, ensuring consistency and reliability across pipelines is paramount. Addressing these discrepancies is not just about refining our current understanding of exoplanets but is crucial for harnessing the full potential of upcoming observational capabilities. Only by resolving the challenge posed by pipeline-dependent results can we truly capitalize on the advanced data, asking deeper questions and drawing more precise conclusions about the Universe's myriad exoplanets.

In this study, we aim to explore the existence of systematic biases originating from the analysis processes of different pipelines in the examination of exoplanet catalogues. It is crucial to note that our objective is not to pinpoint the precise origins of differences between pipelines, a task that would necessitate a detailed comparison of standardized intermediate data products to locate where discrepancies are injected.



**Figure 1.** This study integrates four data sets of transmission spectra, derived from *HST*-WFC3 observations, as shown on the left. The first data set (blue), based on the IRACLIS pipeline, is adopted from Tsiaras et al. (2018) and encompasses 30 transmission spectra. Also based on the IRACLIS pipeline, a second data set (red) was curated and published in Edwards et al. (2022), contributing six planetary spectra to this study. The reader should note here that the sum in the red ellipse gives 28 and not 70. This is because here we are considering only the planets consistently analysed in Edwards et al. (2022), which are 28, and not the planetary spectra that the authors used in their work, but were processed somewhere else. A third data set (orange), constructed with the EXCALIBUR pipeline, is outlined in Roudier et al. (2021) and contains 62 spectra. Of these, 30 correspond to planets analysed in Tsiaras et al. (2018) and overlap with that data set, and 6 overlap with the planets from Edwards et al. (2022). The fourth data set (green), developed through the CASCADE pipeline for this work, employs the planetary parameters from Roudier et al. (2021) and reported in Tables 2 and 3. This novel data set includes 22 spectra that coincide with the other data sets. In this case, we consider only 22 planets and not 30, because for 8 planets the automatic pipeline failed for different reasons: because the goal of this work is to compare consistently analysed populations, we decided to exclude the 8 planets instead of proceeding with dedicated data processing. The figure highlights the intersections between data sets and enumerates the spectra contained in each intersection. Note that one planet, WASP-121b, is shared between all the data sets.

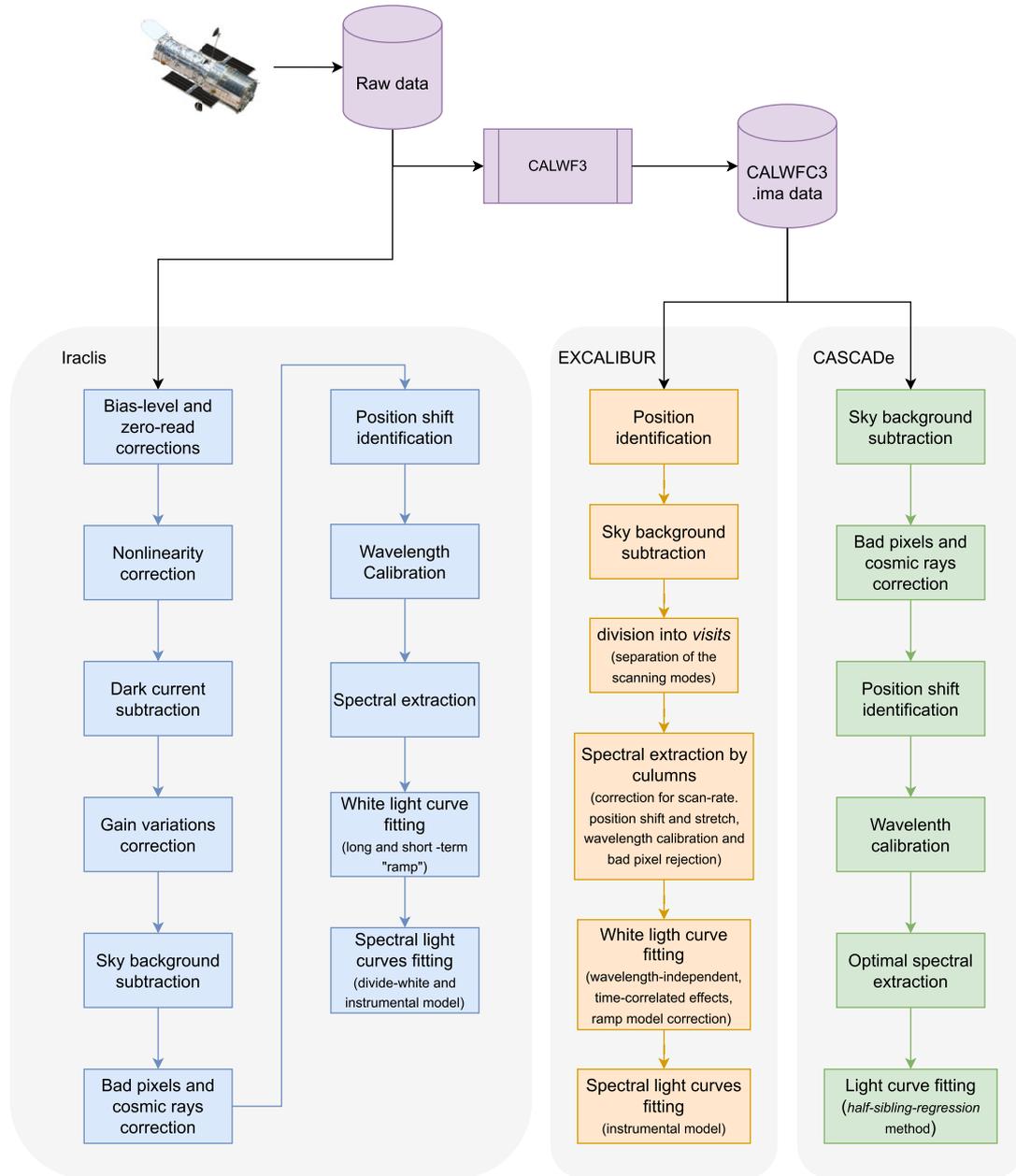
## 2 METHODS

### 2.1 The data sets

In this study, we compare the spectra produced starting from three pipelines (Iraclis, EXCALIBUR, and CASCADE) and four different data sets. A summary representation is reported in Fig. 1.

#### 2.1.1 Iraclis: Tsiaras et al. (2018)

The Iraclis data base, as presented by Tsiaras et al. (2018), contains transmission spectra of 30 gaseous planets, generated using the IRACLIS pipeline. This pipeline executes a series of steps, including zero-read subtraction, reference pixel correction, non-linearity correction, dark current subtraction, gain conversion, sky-background subtraction, calibration, flat-field correction, and bad pixels and cosmic ray correction. The pipeline then extracts the flux from the spatially scanned spectroscopic images to produce the final transit light curves per wavelength band. The light curves are fitted using literature values, with the planet-to-star radius ratio and the transit mid-time as the only free parameters, apart from the coefficients for Hubble systematics. The limb-darkening coefficients are selected from the quadratic formula by Claret (2000), using the stellar parameters. These models have been recently incorporated into the EXOTETHYS package (Morello et al. 2020) utilized by IRACLIS. The spectral light curves are then fitted using the divide-white method introduced by Kreidberg et al. (2014), with the inclusion of a normalization factor



**Figure 2.** Comparison of data reduction steps across pipelines. The data reduction steps for the IRACLIS pipeline are delineated following Tsiaras et al. (2016b), for EXCALIBUR as per Swain et al. (2021), and for CASCADE within the appendices of Carone et al. (2021). These pipelines, while employing conceptually similar strategies for data processing, differ significantly in the specifics of their implementation: examples of these different implementations are reported between parenthesis in the diagram. An in-depth analysis of these differences falls outside the scope of this paper, and readers are referred to the respective primary sources for detailed methodologies.

for the slope. A summary of the data reduction steps is reported in Fig. 2

### 2.1.2 IRACLIS: Edwards et al. (2022)

A subsequent data set, also processed using the IRACLIS pipeline, was introduced by Edwards et al. (2022). This data set contains 70 transmission spectra of gaseous planets processed with the IRACLIS pipeline, with 29 derived from Tsiaras et al. (2018) and 13 obtained from other sources such as Tsiaras et al. (2019), Anisman et al. (2020), Edwards et al. (2020), Pluriel et al. (2020), Skaf et al. (2020), Changeat et al. (2020b), Guilluy et al. (2021), Yip et al. (2021), and

Saba et al. (2022). This study focuses exclusively on the latest 28 transmission spectra, as they were specifically processed for the work presented in Edwards et al. (2022) using an updated version of the software. Of these spectra, only six were used because they coincide with spectra contained in other data sets. One of the six planets is also shared with the Tsiaras et al. (2018) paper: WASP-121b. A deeper discussion on this planet is reported in Section 4.2.

### 2.1.3 EXCALIBUR: Roudier et al. (2021)

The EXCALIBUR 2021 catalogue, presented by Roudier et al. (2021), contains 62 transmission spectra obtained using the EX-

CALIBUR pipeline.<sup>1</sup> EXCALIBUR uses a fully automated, uniform processing approach with persistent intermediate data products to maintain the chain of inference. Further information EXCALIBUR can be found in Roudier et al. (2021), Swain et al. (2021), and Huber-Feely et al. (2022) and a scheme of the pipeline steps is reported in Fig. 2. Of the 62 planets in the EXCALIBUR 2021 catalogue, 30 spectra correspond to planets also processed in Tsiaras et al. (2018). These planets are XO-1b, WASP-31b, HAT-P-38b, HAT-P-41b, HD 209458b, WASP-69b, WASP-76b, WASP-121b, WASP-43b, WASP-52b, WASP-74b, WASP-101b, HD 149026b, HAT-P-17b, HAT-P-12b, WASP-63b, HD 189733b, HAT-P-26b, GJ 436b, HAT-P-11b, HAT-P-32b, WASP-67b, WASP-39b, WASP-80b, GJ 3470b, WASP-29b, WASP-12b, HAT-P-3b, HAT-P-18b, and HAT-P-1b. Additionally, the Excalibur data set shares six planets with the data set presented in Edwards et al. (2022): WASP-121b, WASP-107b, GJ 1214b, KELT-1b, K2-24b, and WASP-18b.

#### 2.1.4 CASCADE: this work

In this work, we also constructed a data base using an automated procedure within the CASCADE<sup>2</sup> pipeline, maintaining the same planetary parameters utilized in Roudier et al. (2021) and listed later in the text in Tables 2 and 3. CASCADE represents an instrument-independent reduction pipeline that has demonstrated its versatility through application to data sets from both the *HST* and the *Spitzer Space Telescope* (Lahuis et al. 2020; Carone et al. 2021). Furthermore, its efficacy has been validated through tests on simulations from the *JWST*'s Mid-Infrared Instrument (*JWST* MIRI). Similar to EXCALIBUR, the CASCADE pipeline initiates the data reduction process with the ‘ima’ intermediate data product. This product is generated by the CALWFC3 data reduction pipeline, marking a departure from pipelines such as IRACLIS, which undertake the data calibration themselves. A novel feature of CASCADE is its implementation of a data-driven method (the *half-sibling-regression* method), a pioneering approach introduced by Schölkopf et al. (2016). This method leverages the causal connections inherent within a data set to calibrate the spectral time series data, potentially enhancing the accuracy and reliability of the data reduction process. For a full description of CASCADE, refer to Carone et al. (2021), while Fig. 2 reports a summary of the data reduction steps as reported in that work appendices. This data set comprises 22 transmission spectra, which are also found in Tsiaras et al. (2018): WASP-31b, HAT-P-41b, WASP-76b, WASP-121b, which is also found in Edwards et al. (2022), WASP-52b, WASP-74b, HD 149026b, HAT-P-17b, HAT-P-12b, WASP-63b, HAT-P-26b, GJ 436b, HAT-P-11b, HAT-P-32b, WASP-67b, WASP-39b, WASP-80b, GJ 3470b, WASP-29b, HAT-P-3b, HAT-P-18b, and HAT-P-1b. All of these spectra overlap with the EXCALIBUR data processed in Roudier et al. (2021). We present in this work only 22 spectra produced with the CASCADE pipeline, instead of fully reproducing the other data sets because we decided to focus on the planets presented in Tsiaras et al. (2018), as these spectra are publicly available and the reader can reproduce our results. Moreover, we decided to run the CASCADE pipeline fully automatically, with the only exception of the control on the stellar and planetary parameter adapted from Roudier et al.

<sup>1</sup>The data are available at <http://excalibur.ipac.caltech.edu/>, referring to EXCALIBUR Run ID 187.

<sup>2</sup>For this work, we use the code version 1.1.15, which is available at <https://jbouwman.gitlab.io/CASCADE/>.

(2021). So, we decided to exclude every planetary observation that requires custom analysis.

#### 2.1.5 Catalogues summary

We included 35 planets in our study. The three data sets share 22 planets: WASP-67b, WASP-31b, HD 149026b, HAT-P-41b, HAT-P-1b, WASP-76b, WASP-74b, HAT-P-12b, HAT-P-17b, HAT-P-26b, WASP-39b, WASP-52b, GJ 436b, HAT-P-32b, HAT-P-11b, HAT-P-3b, WASP-63b, HAT-P-18b, WASP-121b, WASP-80b, GJ 3470b, and WASP-29b. Also, all the data sets include WASP-121b: a detailed discussion around this planet can be found in Section 4.2. We aim to leverage these intersections between the data sets to infer pipeline information on a statistical basis. Such intersections are highlighted in Fig. 1. To help with the identification of the data base in the following, we use the notation

- (i) ‘Iraclis’ for the data base presented in Tsiaras et al. (2018);
- (ii) ‘Edwards2022’ for the data base presented in Edwards et al. (2022);
- (iii) ‘Excalibur’ for the data base presented in Roudier et al. (2021);
- (iv) ‘Cascade’ for the data base produced for this work with the CASCADE pipeline.

Also, a list of the planet spectra contained in each data set is reported in Table 1, while lists of the parameters used from Roudier et al. (2021)<sup>3</sup> are reported in Tables 2 and 3.

A schematic overview of the main steps involved in the three data processing pipelines is presented in Fig. 2. This illustration highlights the utilization of analogous methodologies for data detrending across the pipelines. None the less, the specifics of the implementation for each pipeline vary considerably. An in-depth discussion of these implementation details falls beyond the scope of this article.

## 2.2 Comparison strategy

To enable a comparison of the spectra produced by various pipelines, we bin the spectra to match the spectral resolution used in the Tsiaras et al. (2018) data set, which corresponds to a resolving power of 70 at 1.4  $\mu\text{m}$ . This procedure resulted in 25 data points for each spectrum. Consequently, the Excalibur data set shows one or more empty bins for the following planets due to the EXCALIBUR outlier rejection approach (Swain et al. 2021): HAT-P-12b, HAT-P-32b, WASP-107b, WASP-31b, WASP-39b, and WASP-43b.

Our analysis proceeded in two primary stages. First, we compared the statistical properties of the spectra produced by different pipelines. The results of this comparison are reported in Section 3.1. Secondly, to evaluate the implications of the observed differences on the retrieved information, we conducted consistent retrievals and compared the derived properties (Section 3.2). The described strategy, from the creation of the data sets, is summarized in Fig. 3. As shown in the figure, we decided to not perform the retrieval study for the Edwards2022 catalogue. Further details behind this decision are reported in Section 4.1.

For the retrieval process, we utilized ALFNOOR (Changeat et al. 2020a; Mugnai et al. 2021b), a TAUREX 3 (Al-Refaie et al. 2021) wrapper that streamlines the retrieval procedure. We considered all planets to have primary atmospheres with  $\text{He}/\text{H}_2 = 0.17$ , with

<sup>3</sup>The data are also available with their references at <http://excalibur.ipac.caltech.edu/>, referring to EXCALIBUR Run ID 155.

**Table 1.** List of all the planets included in this work and of the data sets containing their transmission spectra. The header includes the catalogue name (first row), the catalogue reference paper (second row), and the pipeline used (third row). TS18 refers to Tsiaras et al. (2018), ED22 to Edwards et al. (2022), and RO21 to Roudier et al. (2021).

Catalogue	Iraclis	Edwards2022	Excalibur	Cascade
Reference	TS18	ED22	RO21	This work
Pipeline	Iraclis	Iraclis	EXCALIBUR	CASCADE
GJ 1214b	–	✓	✓	–
GJ 3470b	✓	–	✓	✓
GJ 436b	✓	–	✓	✓
HAT-P-11b	✓	–	✓	✓
HAT-P-12b	✓	–	✓	✓
HAT-P-17b	✓	–	✓	✓
HAT-P-18b	✓	–	✓	✓
HAT-P-1b	✓	–	✓	✓
HAT-P-26b	✓	–	✓	✓
HAT-P-32b	✓	–	✓	✓
HAT-P-38b	✓	–	✓	–
HAT-P-3b	✓	–	✓	✓
HAT-P-41b	✓	–	✓	✓
HD 149026b	✓	–	✓	✓
HD 189733b	✓	–	✓	–
HD 209458b	✓	–	✓	–
K2-24b	–	✓	✓	–
KELT-1b	–	✓	✓	–
WASP-101b	✓	–	✓	–
WASP-107b	–	✓	✓	✓
WASP-121b	✓	✓	✓	✓
WASP-12b	✓	–	✓	–
WASP-18b	–	✓	✓	–
WASP-29b	✓	–	✓	✓
WASP-31b	✓	–	✓	✓
WASP-39b	✓	–	✓	✓
WASP-43b	✓	–	✓	–
WASP-52b	✓	–	✓	✓
WASP-63b	✓	–	✓	✓
WASP-67b	✓	–	✓	✓
WASP-69b	✓	–	✓	–
WASP-74b	✓	–	✓	✓
WASP-76b	✓	–	✓	✓
WASP-80b	✓	–	✓	✓
XO-1b	✓	–	✓	–

H<sub>2</sub>O and CH<sub>4</sub> as trace gasses. The parallel plane approximation was adopted for an isothermal atmosphere, composed of 100 layers ranging from 10<sup>6</sup> to 10<sup>−5</sup> Pa. We also incorporated Collision-Induced Absorption (Abel et al. 2011, 2012; Fletcher, Gustafsson & Orton 2018) and Rayleigh effects into the model. The star and planet parameters were sourced from Roudier et al. (2021), and they are listed in Tables 2 and 3. The planetary equilibrium temperature was computed for the temperature parameter as

$$T_{\text{eq}} = T_{\star} \sqrt{\frac{R_{\star}}{2a}} (1 - A)^{\frac{1}{4}}, \quad (1)$$

where  $T_{\star}$  is the stellar temperature,  $R_{\star}$  is the stellar radius,  $a$  is the planet’s semimajor axis, and  $A$  is the Bond albedo that we arbitrarily set to 0.1 for all the planets.

In the fitting procedure, we investigate the planet radius with uniform priors between 0.1 and 10 times the input value, the temperature with uniform priors between 0.5 and 1.5 times the equilibrium temperature, a cloud deck (represented as grey clouds, using logarithmic uniform priors from 10<sup>6</sup> to 10<sup>−5</sup> Pa), and the presence of H<sub>2</sub>O (Polyansky et al. 2018) and CH<sub>4</sub> (Yurchenko &

**Table 2.** List of stellar parameters used in this study. All the parameters are from Roudier et al. (2021) and listed in the EXCALIBUR archive. The reader can refer to that work and the reference therein for further details. For HAT-P-38, KELT-1, and WASP-43, the  $\log g_{\star}$  values were estimated by the EXCALIBUR pipeline due to the absence of direct measurements in the literature.

Name	$\log g_{\star}$	$T_{\star}$ (K)	$R_{\star}$ (R <sub>⊙</sub> )	[Fe/H]
GJ 1214	4.94	3026	0.22	0.39
GJ 3470	4.7	3600	0.55	0.20
GJ 436	4.79	3416	0.46	0.02
HAT-P-11	4.66	4780	0.68	0.31
HAT-P-12	4.61	4650	0.70	−0.29
HAT-P-17	4.53	5246	0.84	0.00
HAT-P-18	4.57	4803	0.75	0.10
HAT-P-1	4.36	5980	1.17	0.13
HAT-P-26	4.56	5079	0.87	−0.04
HAT-P-32	4.22	6001	1.37	−0.16
HAT-P-38	4.45	5330	0.92	0.06
HAT-P-3	4.56	5185	0.87	0.27
HAT-P-41	4.14	6390	1.68	0.21
HD 149026	4.37	6179	1.41	0.32
HD 189733	4.49	5052	0.75	−0.02
HD 209458	4.45	6091	1.19	0.01
K2-24	4.29	5625	1.16	0.34
KELT-1	4.228	6518	1.46	0.01
WASP-101	4.31	6380	1.31	0.20
WASP-107	4.5	4430	0.66	0.02
WASP-121	4.24	6459	1.46	0.13
WASP-12	4.38	6300	1.59	0.30
WASP-18	4.47	6431	1.29	0.13
WASP-29	4.5	4800	0.79	0.11
WASP-31	4.76	6302	1.25	−0.08
WASP-39	4.48	5400	0.90	−0.10
WASP-43	4.646	4400	0.60	−0.05
WASP-52	4.58	5000	0.79	0.03
WASP-63	4.01	5550	1.86	0.08
WASP-67	4.35	5200	0.88	−0.07
WASP-69	4.5	4700	0.86	0.15
WASP-74	4.39	5990	1.42	0.39
WASP-76	4.13	6250	1.73	0.23
WASP-80	4.66	4143	0.59	−0.13
XO-1	4.51	5750	0.88	0.02

Tennyson 2014), using logarithmic uniform priors between 10<sup>−9</sup> and 10<sup>−2</sup>. We employed the Multinest (Feroz, Hobson & Bridges 2009; Buchner et al. 2014) algorithm with 1500 live points and an evidence tolerance of 0.5 for the fitting procedure. The results of these retrievals are discussed in Section 3.2.

For each retrieval, we also estimate the atmospheric detection indices (ADIs), as defined in Tsiaras et al. (2018), by comparing the log evidence for an atmospheric model retrieval with a flat retrieval, where we only fit for radius, temperature, and cloud pressure, and considering no trace gasses or molecular features.

### 3 RESULTS

All spectra considered in this study are presented in Fig. 4. Spectra obtained with different pipelines for the same planet are reported on the same panel to ease the comparison. The colour code used is consistent with Fig. 1 and with the following figures reported in the manuscript to help the reader in identifying the pipeline used: blue is for Iraclis, red for Edwards2022, orange for Excalibur, and green for

**Table 3.** List of planetary parameters used in this study. All the parameters are from Roudier et al. (2021) and listed in the EXCALIBUR archive. The reader can refer to that work and to the reference therein for further details.

Name	Period (d)	$T_{\text{eq}}$ (K)	$a$ (au)	$R$ ( $R_{\text{Jup}}$ )	$M$ ( $M_{\text{Jup}}$ )	Inc. (deg)	Ecc.	$t_0$ (Julian days)
GJ 1214b	1.58040456	561.2	0.0141	0.254	0.02	88.17	0.0	2455320.535733
GJ 3470b	3.3366496	665.5	0.0355	0.408	0.044	89.13	0.017	2455983.70421
GJ 436b	2.64389782	619.8	0.0308	0.372	0.07	86.774	0.0	2456295.431924
HAT-P-11b	4.887802443	809.4	0.0525	0.389	0.074	90.0	0.0	2454957.8132067
HAT-P-12b	3.2130598	932.5	0.0384	0.959	0.211	89.0	0.0	2454419.19556
HAT-P-17b	10.338523	920.8	0.06	1.05	0.58	89.2	0.35	2454801.16945
HAT-P-18b	5.508023	826.3	0.0559	0.995	0.197	88.8	0.084	2454715.02174
HAT-P-1b	4.46529976	1288.3	0.0556	1.319	0.525	85.634	0.0	2453979.92802
HAT-P-26b	4.23452	1016.6	0.0479	0.63	0.07	88.6	0.12	2455304.6522
HAT-P-32b	2.1500082	1789.9	0.034	1.98	0.68	88.98	0.159	2455867.402743
HAT-P-38b	4.640382	1049.9	0.0523	0.825	0.267	88.3	0.067	2455863.11957
HAT-P-3b	2.8997	1151.1	0.0389	0.94	0.65	87.24	0.0	2454218.81
HAT-P-41b	2.694047	1886.4	0.0426	2.05	1.19	87.7	0.0	2454983.86167
HD 149026b	2.87589	1649.6	0.0436	0.74	0.38	84.55	0.0	2454597.7071
HD 189733b	2.21857567	1161.5	0.0313	1.13	1.13	85.71	0.0	2453955.5256
HD 209458b	3.52474859	1438.4	0.0471	1.39	0.73	86.71	0.0	2452826.6293
K2-24b	20.88977	725.1	0.154	0.482	0.06	88.874	0.06	2456905.8855
KELT-1b	1.217513	2353.6	0.0247	1.11	27.2	87.6	0.0	2455933.61
WASP-101b	3.58572	1525.1	0.0506	1.43	0.51	85.0	0.0	2456164.6941
WASP-107b	5.72149	720.8	0.055	0.94	0.12	89.7	0.0	2456514.4106
WASP-121b	1.2749255	2298.1	0.0254	1.865	1.18	87.6	0.0	2456635.70832
WASP-12b	1.09142245	2439.1	0.0234	1.937	1.46	82.5	0.0447	2456176.6683
WASP-18b	0.94145	2413.7	0.0202	1.2	11.4	80.6	0.01	2455265.5525
WASP-29b	3.92273	937.3	0.0457	0.77	0.23	88.8	0.03	2455830.1889
WASP-31b	3.4059096	1533.1	0.0466	1.549	0.478	84.41	0.0	2455192.6887
WASP-39b	4.055259	1091.4	0.0486	1.27	0.28	87.83	0.0	2455342.9688
WASP-43b	0.813475	1343.3	0.0142	0.93	1.78	82.6	0.0	2455528.86774
WASP-52b	1.7497798	1265.6	0.0272	1.27	0.46	85.35	0.0	2455793.68143
WASP-63b	4.37808	1483.8	0.0574	1.41	0.37	87.8	0.0	2455921.6536
WASP-67b	4.61442	1006.7	0.0518	1.15	0.43	85.8	0.0	2455824.375
WASP-69b	3.86814	962.3	0.0452	1.11	0.29	86.71	0.0	2455748.8342
WASP-74b	2.13775	1741.7	0.037	1.36	0.72	79.81	0.0	2456506.8926
WASP-76b	1.809886	2125.4	0.033	1.83	0.92	88.0	0.0	2456107.85507
WASP-80b	3.06785234	805.9	0.0344	0.999	0.538	89.02	0.002	2456487.425006
XO-1b	3.94153	1146.8	0.0488	1.14	0.83	88.81	0.0	2453887.7477

Cascade. The y-axis is the panels of Fig. 4 is automatically scaled to fit the spectra with their offsets.

### 3.1 Spectra statistics

The spectral variations depicted in Fig. 4 can be broadly categorized for each planet into three distinct types:

- (i) Variations in mean values, observable as offsets between the spectra.
- (ii) Discrepancies in claimed precision, manifested as differing error bars for the spectra.
- (iii) Alterations in spectral shape, discernible as varying values in the spectral bins between different spectra, assuming no offset between them.

In the subsequent sections, we delve into a detailed analysis of these categories.

#### 3.1.1 Variations in mean values

Initially, we examine the mean values of the processed spectra, as elaborated in Fig. 4. Fig. 5 demonstrates the ratio of the spectra mean values between different pipelines for the same planet, denoted as

$MK1_{A,B}$ :

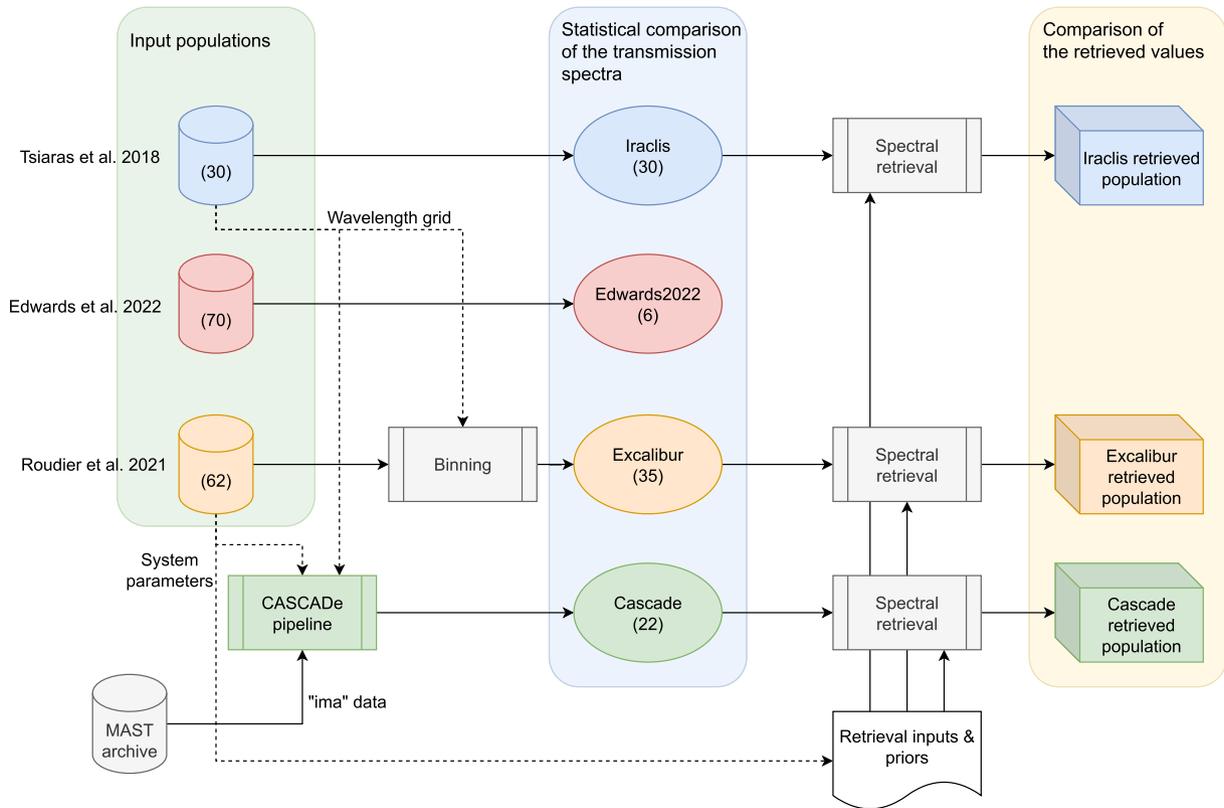
$$MK1_{A,B} = \frac{\widehat{Sp}_A(\lambda)}{\widehat{Sp}_B(\lambda)}. \quad (2)$$

Here,  $Sp(\lambda)$  stands for the measurements in the spectral bin  $\lambda$ ,  $A$  and  $B$  symbolize different data sets, and  $\widehat{Sp}_A(\lambda)$  represents the mean value across the spectral bins for that planet in the  $A$  data set.

The left panel of Fig. 5 presents histograms of the ratio of mean values from each pipeline compared to Excalibur. As reported in the legend, median values align closely with one, the expected value for ideal pipelines. Despite the histograms' non-Normal distribution, computing the standard deviation aids in the interpretation of these ratios.

While the standard deviation of the Iraclis to Excalibur ratio suggests a ratio of the radii between 0.91 and 1.10, the case for Cascade is more complex. For Cascade, 68 per cent of the ratios lie between 0.62 and 1.36, with a spread that is 3.7 times larger than what was observed in the comparison between Iraclis and Excalibur. This dispersion suggests some pipelines can yield significantly different planet radius estimates.

The subsequent panels in Fig. 5 adopt Cascade (central panel) and Iraclis (right panel) as references. The right panel echoes the findings of the left, while the central panel reveals similar performance for Excalibur and Iraclis when compared to Cascade.



**Figure 3.** The figure summarizes the strategy used to compare the planetary transmission spectra. Starting from the input populations from published works, we binned down all the spectra to the same spectral resolution, and we used the same planetary parameters used in Roudier et al. (2021) to produce the Cascade data set. We compare then the transmission spectra, obtaining the results reported in Section 3.1. Using the input parameters reported in Tables 2 and 3, we perform spectral retrieval on each spectrum of the most populated data sets, using the same boundaries and priors. Finally, we compare the derived measurements in Section 3.2.

We additionally report the number of outliers in the panel, which we define as planets with a ratio of  $<0.75$  or  $>1.25$ . This definition implies a tolerance of 25 percent on the ratio of the means. Remarkably, when comparing Cascade and Excalibur, we find that 14.29 percent of the ratios fall outside of our defined tolerance. The reason for this percentage discrepancy warrants further investigation because such different measurements imply different interpretations of the planetary classes. The discrepancy is highlighted in Fig. 4, with certain planets (e.g. HAT-P-17b, WASP-29b, WASP-39b, and WASP-74b) showing significant differences in estimated radii, potentially due to normalization differences (Carone et al. 2021). These four spectra, representing 11.43 per cent of our sample, are marked as outliers in Fig. 5. Recalculating the central panel’s statistics without these spectra yields an average ratio of  $1.00 \pm 0.06$  against Iraclis and  $1.00 \pm 0.10$  against Excalibur.

In general, the ratio of the estimated radii is well centred around one, meaning that no pipeline has a preferred bias towards bigger or smaller radii. However, the number of significantly discrepant estimations points to certain planets where the automated procedures of the pipelines yield highly varied radius measurements.

### 3.1.2 Discrepancies in uncertainty estimates

In the following paragraph, we compare the uncertainties and the values computed by the pipeline for each spectral bin.

Fig. 6 presents a comparative analysis of the error bars across each spectral channel. For each exoplanet, we examine the spectra

generated by two distinct pipelines and compare the error bars for each spectral channel, represented as  $MK2_{A,B}$ :

$$MK2_{A,B} = \frac{\sigma_A(\lambda)}{\sigma_B(\lambda)}, \quad (3)$$

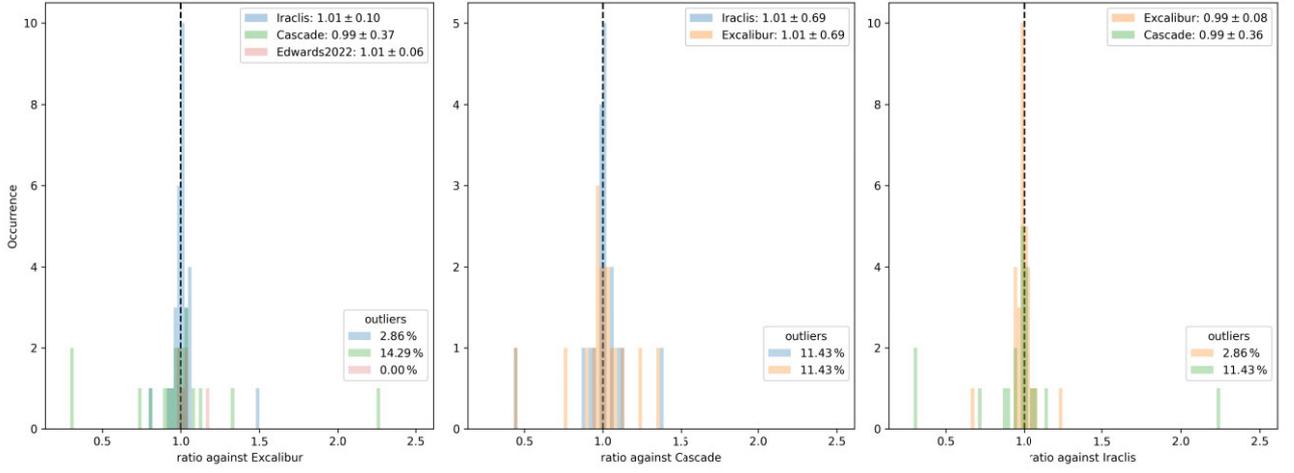
where  $A$  and  $B$  denote the error bars derived from the two pipelines. In the three panels, we use a different pipeline as a reference for each comparison and we report the comparison as histograms with their cumulative curve reported on the right y-axis.

The expectation for pipelines that are consistent with each other is to yield consistent uncertainties across the spectral channels. However, as observed from the first panel, the distribution produced by the six planets from Edwards2022 (Edwards et al. 2022) does not align with the data generated by the other pipelines. This discrepancy suggests a potential inconsistency in the pipeline used in Edwards2022 or a unique characteristic of the planets used from that work. However, this discrepancy could also be caused by small number statistics. It is also worth mentioning the case of KELT-1b, where the mean ratio between the uncertainties estimated by Edwards2022 and Excalibur is 34, which suggests pipeline-to-pipeline differences in the interpretation of the planetary system and the data.

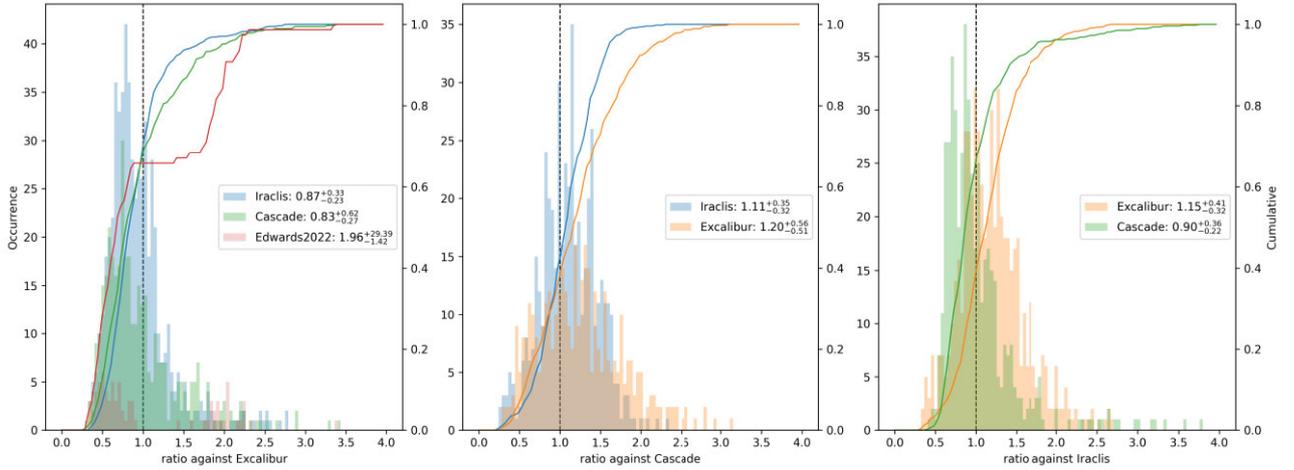
Conversely, the Iraclis and Cascade distributions have means at  $0.87^{+0.33}_{-0.23}$  and  $0.83^{+0.62}_{-0.27}$ , respectively, indicating that the Excalibur data set generally exhibits approximately  $\sim 15$  per cent bigger uncertainties across the spectral channels. The cumulative curves in the panel help highlight this behaviour by reporting the ratio that



**Figure 4.** This figure displays all spectra for the 35 planets analysed in this study. Spectra from the Iraclis data set are represented in blue, sourced from Tsiaras et al. (2018). The Edwards2022 data set spectra, taken from Edwards et al. (2022), are illustrated in red. In orange, we present the Excalibur data set spectra, obtained from Roudier et al. (2021). The unbinned spectra data points are superimposed on the binned one using the same colour. Lastly, spectra produced using the automated CASCADE pipeline for this work are shown in green.



**Figure 5.** This figure displays the ratio of mean spectral values across various exoplanets, defined in the text as  $MK1_{A,B}$  (equation 2). Each panel portrays histograms of the mean spectral value ratio. Progressing from left to right, histograms are plotted with respect to the Excalibur data set, the Cascade data set, and the Iraclis data set. The data are divided into 100 uniformly spaced bins ranging from 0 to 2.5. To emphasize the discrepancies in the statistical populations between data sets, occurrences are not normalized. The upper legend denotes the mean values along with standard deviations. The lower legend lists the outliers, which are defined in this context as spectra with means diverging by more than  $\pm 25$  per cent.



**Figure 6.** This figure illustrates the ratio between the error bars in identical spectral channels across different planets, defined as  $MK2_{A,B}$  (equation 3). Each panel presents histograms of the ratio of the error bars for the same spectral channel. Moving from left to right, histograms are shown with respect to the Excalibur data set, the Cascade data set, and the Iraclis data set. The data are categorized into 100 evenly spaced bins within a range of 0–4. Occurrences are not normalized to underline the differences in the statistical populations between data sets. Cumulative curves are superimposed to the histogram and refer to the right y-axis.

reaches the saturation limit (1.0) slower. In a similar vein, we observe that Iraclis tends to estimate uncertainties that are approximately 10 per cent bigger than those estimated by Cascade.

### 3.1.3 Differences in spectral shape

Fig. 7 illustrates the difference in the estimated values for each spectral channel between two pipelines, normalized by the maximum estimated uncertainties in the channel. To compare the spectra, we also subtract the median values to remove any offset. This is represented as  $MK3_{A,B}$ :

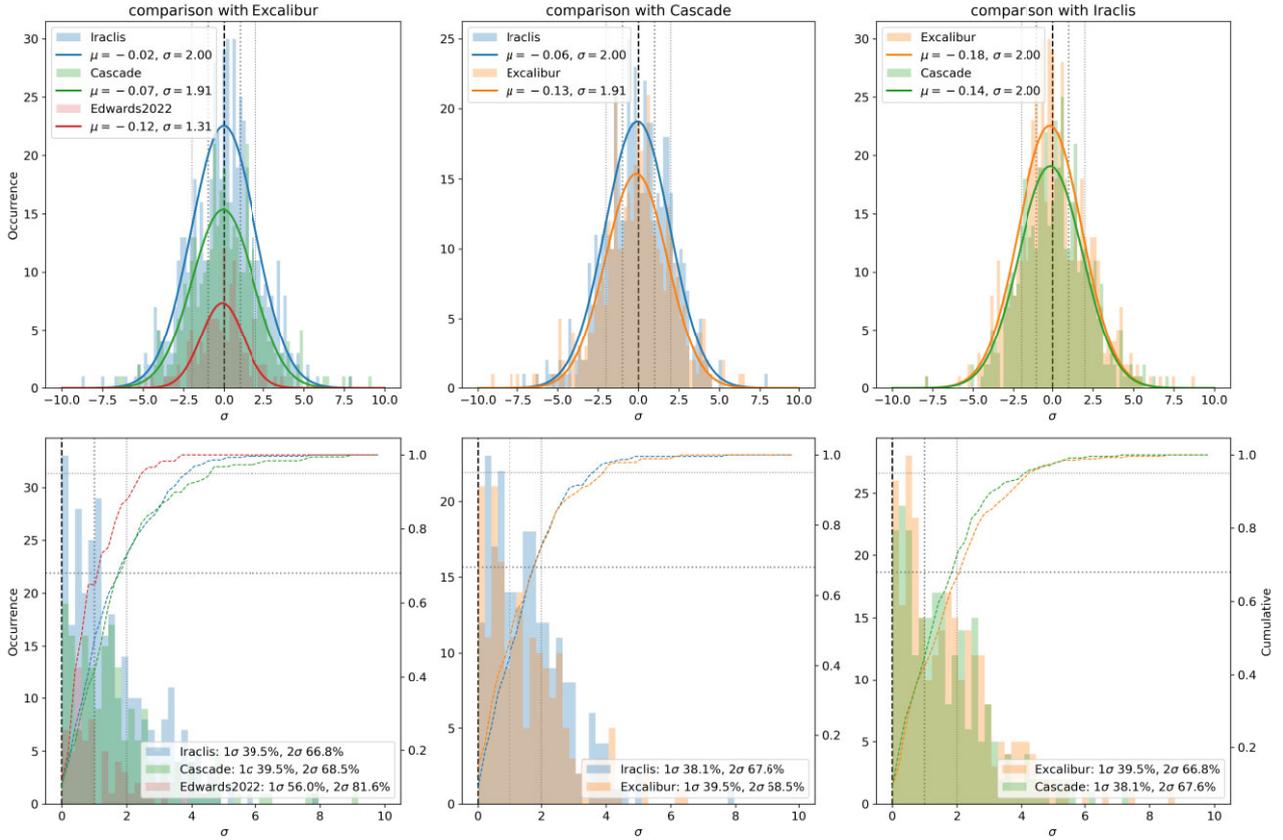
$$MK3_{A,B} = \frac{(Sp_A(\lambda) - \widehat{Sp}_A) - (Sp_B(\lambda) - \widehat{Sp}_B)}{\max[\sigma_A(\lambda), \sigma_B(\lambda)]} \quad (4)$$

in the top panels and as  $MK4_{A,B}$ :

$$MK4_{A,B} = \frac{|(Sp_A(\lambda) - \widehat{Sp}_A) - (Sp_B(\lambda) - \widehat{Sp}_B)|}{\max[\sigma_A(\lambda), \sigma_B(\lambda)]} \quad (5)$$

in the bottom panels, where the absolute value of the difference is considered.  $Sp(\lambda)$  signifies the values of the spectral bin, while  $\sigma(\lambda)$  denotes the standard deviation within that bin.  $A$  and  $B$  are placeholders for the two pipelines being compared.

Upon observation, it is evident that there is no significant offset in the distributions. However, the histograms exhibit a broader spread than anticipated. In fact, for consistent spectra, we expect 68 per cent of the data to be consistent within  $1\sigma$ . Specifically, we find that only approximately 40 per cent of the spectral bins yield compatible estimates within the  $1\sigma$  uncertainties. This suggests that the pipelines may not be as consistent with each other as desired, or



**Figure 7.** This figure presents histograms of normalized spectral bin differences, where normalization is achieved using the largest standard deviation reported by any pipeline for that specific spectral bin. The top panels show the difference defined as defined in the text as  $MK3_{A,B}$  (equation 4). Here,  $Sp_\lambda$  denotes the values of the spectral bin and  $\sigma_\lambda$  represents the standard deviation within that bin.  $A$  and  $B$  correspond to the two pipelines under comparison. The data are distributed into 100 evenly spaced bins within a range of  $-10$  and  $10$ . To emphasize the distinctions in statistical populations between data sets, occurrences are not normalized. Normal distributions are fitted to the histograms and the fit parameters are reported in the legends. The bottom panels show the absolute difference computed as  $MK4_{A,B}$  (equation 5), using 50 bins evenly spaced between 0 and 10. The cumulative distribution is overplotted in each panel. The legend provides percentages of data points falling within  $1\sigma$  and  $2\sigma$ , also highlighted by the vertical lines in all the panels. The horizontal lines in the bottom panel show the 68 and the 95 per cent levels desired for  $1\sigma$  and  $2\sigma$ , respectively.

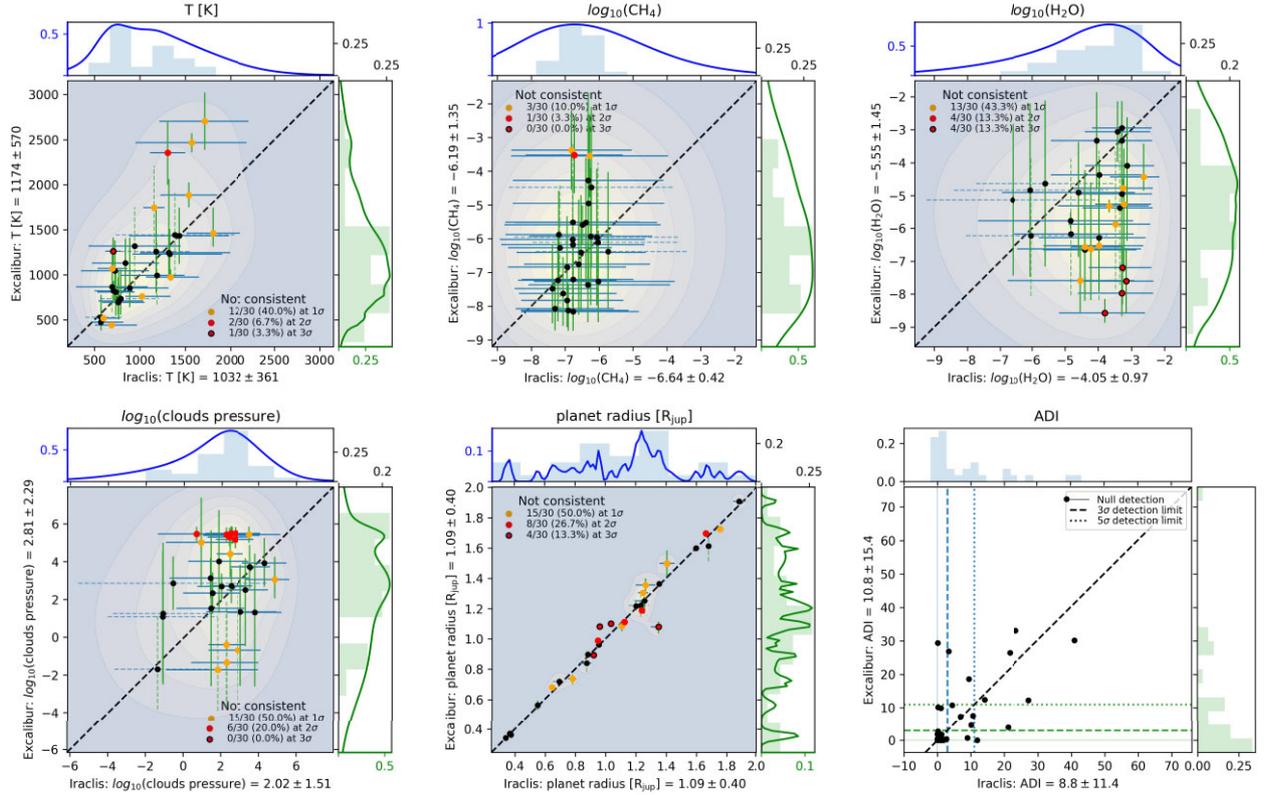
that the uncertainties are underestimated. The large distribution, also highlighted by the Normal functions fitted to the data in Fig. 7 shows no bias, as they are centred around 1, but they also highlight that the spectral features estimated by the pipelines are consistent within  $2\sigma$  for around 68 per cent of the spectral bins. The same estimate is evident from the bottom panels, where are reported the cumulative of the absolute normalized difference. From the horizontal dotted lines set at 0.68 and 0.95, we notice that the cumulative reaches the 0.68 line close to  $2\sigma$ , indicated by a dotted vertical line.

The sole exception to this trend is the comparison between the Excalibur and Edwards2022 pipelines, where compatibility within  $1\sigma$  is achieved for up to 62 per cent of the spectral bins. This higher degree of compatibility could indicate similarities in the methodologies or algorithms used by these two pipelines. However, this comparison might also suffer from the limitations of small number statistics and may not be representative of overall pipeline behaviour.

It is pertinent to once again highlight the case of KELT-1b. This exoplanet, with a radius of  $1.11 R_{\text{Jup}}$  and a mean density  $24 \text{ g cm}^{-3}$  (Siverd et al. 2012), exhibits a Transit Spectroscopy Metric of 3 as per Kempton et al. (2018), a value that would not predict

detectable spectral modulation in its transit spectrum through *HST* observations. Results from Excalibur (Roudier et al. 2021) reveal a flat spectrum, with a mean value of  $0.55917 \pm 0.00004$ , and the residual standard deviation for individual spectral light curves generally around 2.5 times the photon noise. In contrast, Edwards et al. (2022) identified a more modulated spectrum, with a mean value of  $0.577 \pm 0.009$ , still compatible with a flat line according to Edwards et al. (2022) appendix B10, but with clear channel-to-channel correlations in transit depth observed. While analysing differences in individual targets can be constructive, caution is advised in overinterpreting results based on single-target comparisons. Given the myriad factors that can influence comparisons on individual subjects, we emphasize that pipeline-to-pipeline comparison should ideally be conducted through the analysis of entire catalogues. This approach provides a more robust and replicable framework for evaluating the consistency and reliability of data across different analytical methodologies. Indeed, through systematic comparison across complete catalogues, we can identify trends, consistencies, and discrepancies that are crucial for further refining our observation and analysis techniques, thereby promoting a deeper and more accurate understanding of exoplanetary environments.

## Comparison of Iraclis - Excalibur



**Figure 8.** Comparison of retrieval results for Iraclis and Excalibur retrieval results for various planets, marked with blue and green error bars for  $x$ - and  $y$ -axis uncertainties, respectively. Compatibility of data sets indicated by alignment with dotted bisector line and colour-coded discrepancies: orange for between 1 and 2 standard deviations ( $\sigma$ ), light red for  $2\sigma$ – $3\sigma$ , and dark red for more than  $3\sigma$ . Parameter space sampling is represented by normalized 2D Normal distributions forming a coloured background. Side panels display projected marginal distributions referencing the colored axis at the bottom and histogram references along the black axis at the top. The  $x$ - and  $y$ -axis labels also report the mean retrieved values with standard deviations. The final panel contrasts ADIs with thresholds for  $3\sigma$  (dashed) and  $5\sigma$  (dotted) atmospheric detection.

### 3.2 Retrieval comparison

To evaluate the discrepancies in the retrieved quantities, we juxtapose the retrieval results for two distinct data sets within the same panel. The results are reported in Figs 8–10, where each panel displays the retrieved quantities along with their uncertainties for each planet. Blue error bars correspond to the  $x$ -axis, while green error bars are associated with the  $y$ -axis. The same colour scheme is applied to the side panels (described later in this section). Dotted error bars are used to highlight those retrieved quantities with a null ADI, indicating that a flat line fits the data better than an atmospheric spectrum.

A data point is deemed compatible with the dotted black bisector line if the estimates for the two data sets align. If the bisector falls over  $1\sigma$  from the measurements, the data point is highlighted in orange. If the bisector is over  $2\sigma$ , the point is marked in light red, and if it is further than  $3\sigma$ , the data point is depicted in red with a black edge. The colour-coding scheme used in this analysis provides a visual representation of the level of agreement between the two data sets. It is important to note that the colour of a data point does not necessarily reflect the accuracy or reliability of the estimates, but rather the degree of discrepancy between the two data sets. A list of all the planets that are not consistent within  $3\sigma$  for each panel and each figure is reported in Table 4. Upon examining Section 3.1.1, we find that HAT-P-17b, WASP-29b, WASP-39b, and WASP-74b were identified as outliers in terms of their spectral mean values. This

observation aligns with the expectation that these planets exhibit inconsistencies in their retrieved radius estimations, as highlighted by their discordance within a  $3\sigma$  range in Table 4

To estimate the extent of the parameter space sampled, each data point is replaced with a two-dimensional Normal distribution, with the uncertainties serving as standard deviations in each direction. These Normal distributions are then normalized to 1 over the number of data points on the plots and summed. This process generates the coloured background of the panels, which is normalized to 1.

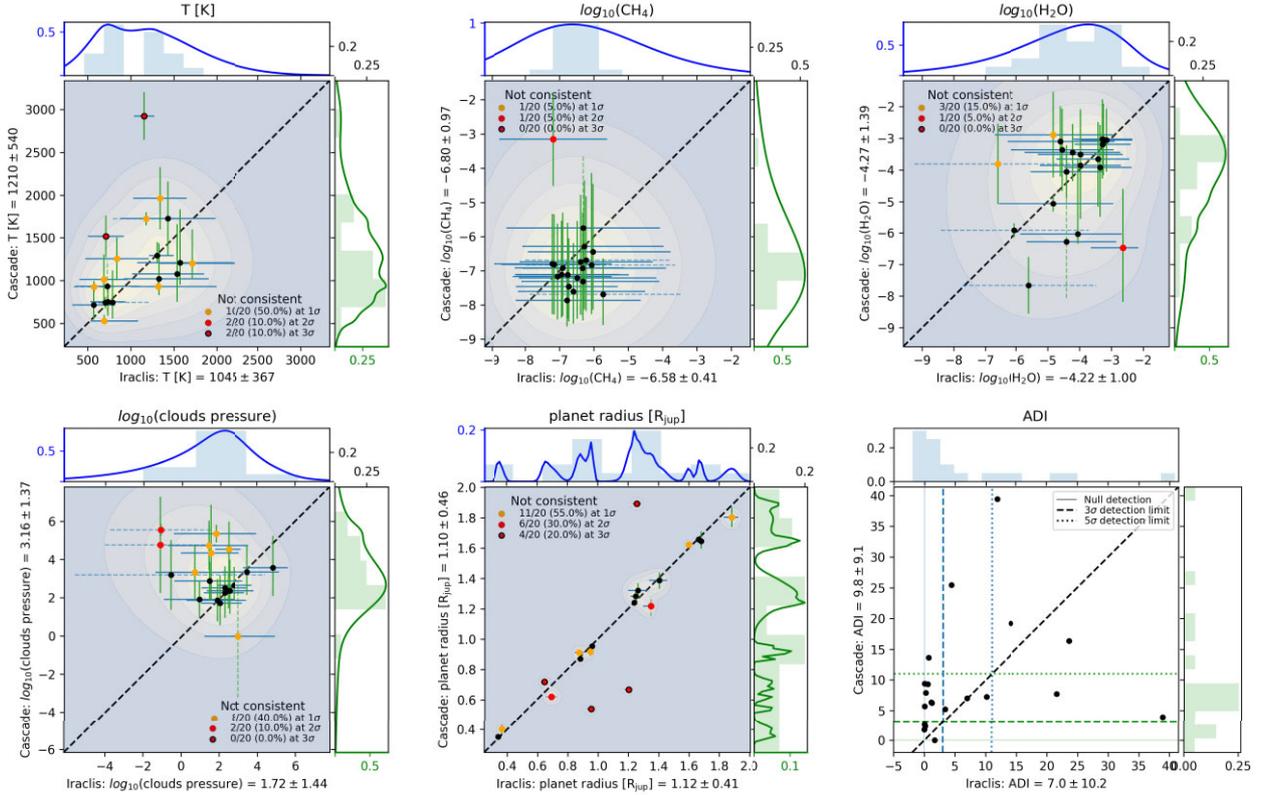
The background is subsequently projected in the two directions and displayed in the side panels to represent the marginal distribution. These panels also showcase the histograms for the data point distribution as references.

Lastly, the final panel in Fig. 8–10 presents the comparison between the estimated ADIs. The dashed lines represent the threshold for a  $3\sigma$  detection of an atmosphere, while the dotted line signifies the limit for a  $5\sigma$  detection. The solid line represents a null detection ( $\text{ADI} = 0$ ). The lines are colour-coded in blue for the horizontal axis and green for the vertical, as the other panels.

#### 3.2.1 Iraclis and Excalibur

Fig. 8 presents a comparison between the retrieval results of Iraclis and Excalibur. The first panel showcases the temperature fit. Only

## Comparison of Iraclis - Cascade



**Figure 9.** Comparison of retrieval results for Iraclis and Cascade retrieval results for various planets, marked with blue and green error bars for  $x$ - and  $y$ -axis uncertainties, respectively. Compatibility of data sets indicated by alignment with dotted bisector line and colour-coded discrepancies: orange for between 1 and 2 standard deviations ( $\sigma$ ), light red for  $2\sigma$ – $3\sigma$ , and dark red for more than  $3\sigma$ . Parameter space sampling is represented by normalized 2D Normal distributions forming a coloured background. Side panels display projected marginal distributions referencing the colored axis at the bottom and histogram references along the black axis at the top. The  $x$ - and  $y$ -axis labels also report the mean retrieved values with standard deviations. The final panel contrasts ADIs with thresholds for  $3\sigma$  (dashed) and  $5\sigma$  (dotted) atmospheric detection.

WASP-52b does not have compatible estimates within  $3\sigma$ , however, 40 per cent of the planets are not compatible within  $1\sigma$ . Overall, the retrieved temperatures between the data sets are generally compatible, but it is noticeable that Excalibur tends to explore the temperature range between 2000 and 3000 K, while Iraclis spectra do not exceed 2000 K. Additionally, the uncertainties on Excalibur retrieved values are typically smaller than the uncertainties on Iraclis retrieved values.

In terms of the atmospheric features fit, for  $\text{CH}_4$ , Iraclis consistently finds abundances around or below  $10^{-6}$ , while Excalibur appears to investigate the range  $10^{-8}$  to  $10^{-3}$ . Similarly, for water, Iraclis identifies water at about  $10^{-5}$  to  $10^{-3}$  for most of the planets, with some exceptions down to  $10^{-6}$ . Excalibur uniformly samples the range between  $10^{-8}$  (which is a prior dominated region of the parameter space) and  $10^{-3}$ . There are four planets not compatible by more than  $3\sigma$ , but they are worth noticing because they are in the prior dominated region for Excalibur ( $<10^{-7}$ ) and in a clear detection area for Iraclis ( $>10^{-4}$ ). These are HAT-P-1b, HAT-P-41b, HD 209458b, and WASP-121b, as listed in Table 4. However, it is worth mentioning here that the *HST*-WFC3 spectra analysed in this study are all collected in the wavelength range primarily sensitive to water absorption, and they count only 25 data points each in the wavelength range. Therefore, the debate between water and methane can be based only on one or two data points.

For the cloud pressure panel, it is observed that Iraclis estimates a peak around 316 Pa, while Excalibur’s distribution of estimates is biased towards higher values, such as 10 000 Pa. Although all planets are compatible within  $3\sigma$ , there are six planets not compatible by more than  $2\sigma$  and they all are in the  $>1000$  Pa range for Excalibur estimates.

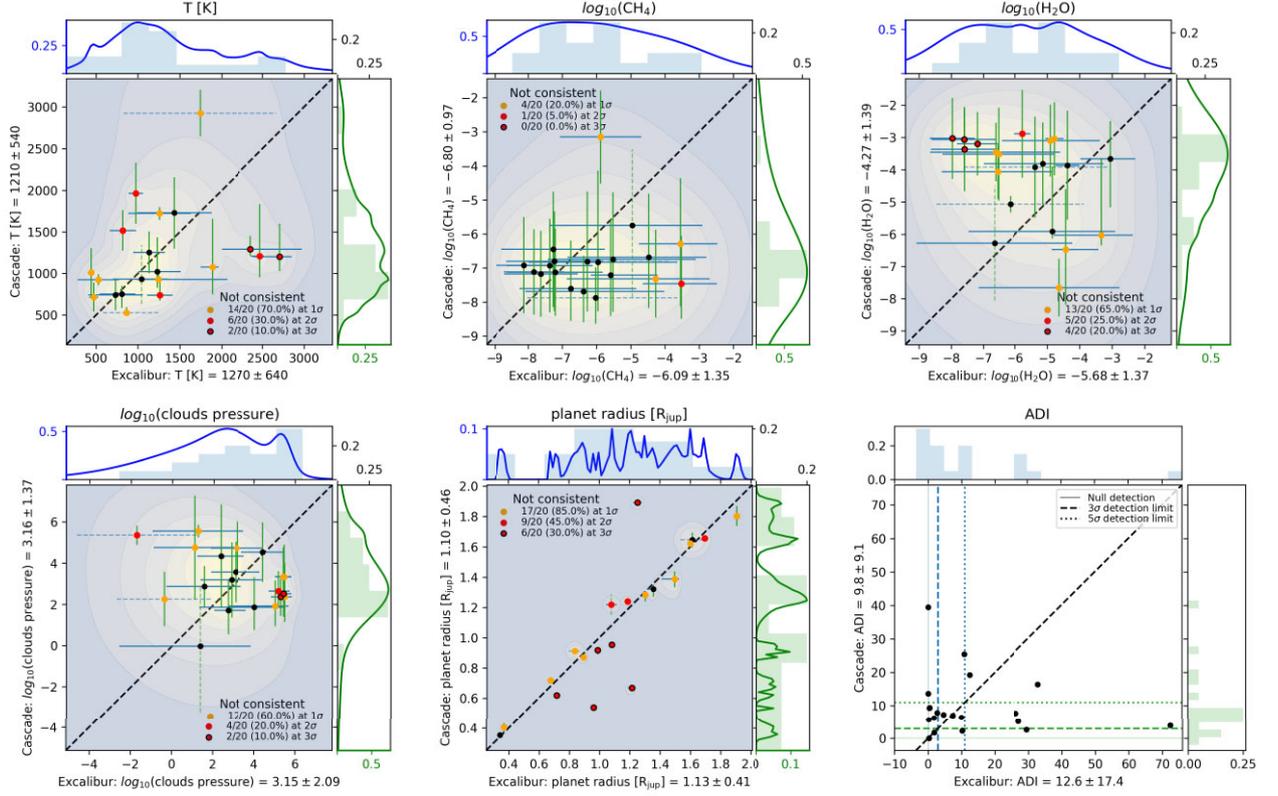
The subsequent panel reports the planetary radius fit. All these estimates have small error bars, and even though all the values are close to the bisector, there are 8 planets out of 30 that are not compatible by more than  $2\sigma$ . However, no bias is observed in this panel, as the data points are spread along all the parameter space uniformly.

Finally, the ADI panel indicates that there are some planets for which Excalibur claims a strong detection, while Iraclis does not. This is a reflection of the previous panels which highlights the impact of these differences on our understanding of these exoplanets. The most extreme disagreement, however, is the single planet for which Iraclis claims a  $3\sigma$  detection, while Excalibur has a null adi: WASP-76b.

### 3.2.2 Iraclis and Cascade

Fig. 9, which compares the retrieval results of Iraclis and Cascade, reveals that only 50 per cent of the planets have retrieved temperature

## Comparison of Excalibur - Cascade



**Figure 10.** Comparison of retrieval results for Excalibur and Cascade retrieval results for various planets, marked with blue and green error bars for  $x$ - and  $y$ -axis uncertainties, respectively. Compatibility of data sets indicated by alignment with dotted bisector line and colour-coded discrepancies: orange for between 1 and 2 standard deviations ( $\sigma$ ), light red for  $2\sigma$ – $3\sigma$ , and dark red for more than  $3\sigma$ . Parameter space sampling is represented by normalized 2D Normal distributions forming a coloured background. Side panels display projected marginal distributions referencing the colored axis at the bottom and histogram references along the black axis at the top. The  $x$ - and  $y$ -axis labels also report the mean retrieved values with standard deviations. The final panel contrasts ADIs with thresholds for  $3\sigma$  (dashed) and  $5\sigma$  (dotted) atmospheric detection.

**Table 4.** List of planets not consistent within  $3\sigma$  according to Figs 8–10.

Data sets	$T$	$\text{CH}_4$	$\text{H}_2\text{O}$	Clouds pressure	Planet radius
Iraclis Excalibur (Fig. 8)	WASP-52b	–	HD209458b HAT-P-41b HAT-P-1b WASP-121b	–	WASP-80b WASP-69b WASP-43b WASP-67b
Iraclis Cascade (Fig. 9)	WASP-76b WASP-39b	–	–	–	HD149026b WASP-74b WASP-39b HAT-P-17b
Excalibur Cascade (Fig. 10)	HAT-P-41b WASP-121b	–	HAT-P-41b WASP-63b HAT-P-1b WASP-121b	HAT-P-18b HAT-P-1b	WASP-74b WASP-80b HAT-P-18b WASP-29b WASP-39b HAT-P-17b

estimates that are consistent within  $1\sigma$ . It is also noticeable that two planets are not consistent within more than  $3\sigma$ . One of these is WASP-39b, for which Cascade finds a temperature of 1450 K, while Iraclis finds 763 K. For completeness, Excalibur claims 840 K. The case of WASP-76b is even more striking: Cascade finds 2875 K, while Iraclis finds 1186 K. For comparison, Excalibur estimates 1893 K.

The atmospheric features fit for water and methane are largely consistent between the two data sets, with the notable exception of WASP-39b, which consistently falls between  $2\sigma$  and  $3\sigma$ . The discrepancies in retrieved atmospheric content for WASP-39b are correlated with the difference in retrieved temperature. Interestingly, no differences are found in the molecular content for WASP-

76b, for which both data sets find no methane but detect water around  $10^{-4}$ .

The cloud pressure panel in Fig. 9 reveals discrepancies for eight planets, two of which are not consistent within more than  $2\sigma$ . These differences primarily occur in the same area of the parameter space, where Cascade detects clouds from  $10^3$  to  $10^6$  Pa.

The radius panel exhibits more discrepancies than the others: here, 55 per cent of the planets have measurements that are not consistent within more than  $1\sigma$ , of which four planets are not consistent within  $3\sigma$ . Contrary to what was observed in Fig. 8, here we notice that some measurements are far from the bisector, which may result in different estimates for the planetary class. Two notable examples are WASP-74b and WASP-39b. For WASP-74b, Cascade estimates a radius of  $1.89 R_{\text{jup}}$ , while Iraclis estimates  $1.24 R_{\text{jup}}$ . For WASP-39b, Cascade estimates  $0.67 R_{\text{jup}}$  and Iraclis estimates  $1.20 R_{\text{jup}}$ .

These differences in radius estimates result in discrepancies in the ADI estimates. In fact, we observe in the last panel that for some planets, Iraclis claims a very strong detection for an atmosphere, while Cascade cannot claim any detection, and vice versa. In particular, it is worth noticing that there are two data points for which Iraclis claims a null ADI while Cascade does not: HAT-P-3b, for which Cascade measures 9.4, and GJ 436b, for which it claims 5.7.

### 3.2.3 Excalibur and Cascade

Finally, Fig. 10 provides a comparison between the Excalibur and Cascade data sets. It should be noted that both pipelines have been run with the same system parameters to extract the spectra. The first panel already reveals that only 30 per cent of the temperature estimates are consistent within  $1\sigma$ . There does not appear to be a preferential bias in one of the two data sets, as differences in the estimates up to a factor of 2 are observed along both axes.

In the molecular feature panels, we observe, similarly to Fig. 8, that while Excalibur seems more sensitive to methane, Cascade only has estimates between  $10^{-6}$  and  $10^{-8}$ . The only exception is WASP-39b, for which Cascade estimates a methane abundance of  $10^{-3}$ , while Excalibur places it at  $10^{-6}$ . For comparison, Iraclis claims  $10^{-7}$  for this planet.

In the water panel, there is a group of inconsistent measurements in the area of no detection for Excalibur ( $10^{-8}$  to  $10^{-6}$ ) and strong detection for Cascade ( $10^{-4}$  to  $10^{-3}$ ). However, there are also three  $1\sigma$  discrepancies in the opposite direction, where Cascade claims no water in the atmosphere ( $10^{-8}$  to  $10^{-6}$ ) and Excalibur claims a large abundance ( $10^{-5}$  to  $10^{-3}$ ).

The panels for cloud pressure and radius tell a similar story to what was observed in Fig. 9. This is expected because, as shown in Fig. 8, Excalibur and Iraclis have all the radii measurements close to the bisector.

In the ADI panel, we notice two particularly anomalous data points. These are HAT-P-12b and WASP-76b. For these, Excalibur records a null ADI, while Cascade claims a more than  $5\sigma$  detection, with an ADI of 13.6 for HAT-P-12b, and of 39.4 for WASP-76b. For the same planets, Iraclis reports 0.6 and 11.8, respectively.

## 4 DISCUSSION

### 4.1 About Edwards2022 data set

In Section 2.2, we mention our decision to exclude the Edwards2022 data set from our retrieval analysis. This choice stems from a

confluence of reasons, notably the data set's smaller sample size of six planets and the distribution depicted in Fig. 6, which deviates from the patterns observed in other data sets. Such a limited data set does not offer a sufficiently robust basis to deduce population attributes or discern any statistically meaningful deviations from prior Iraclis analyses.

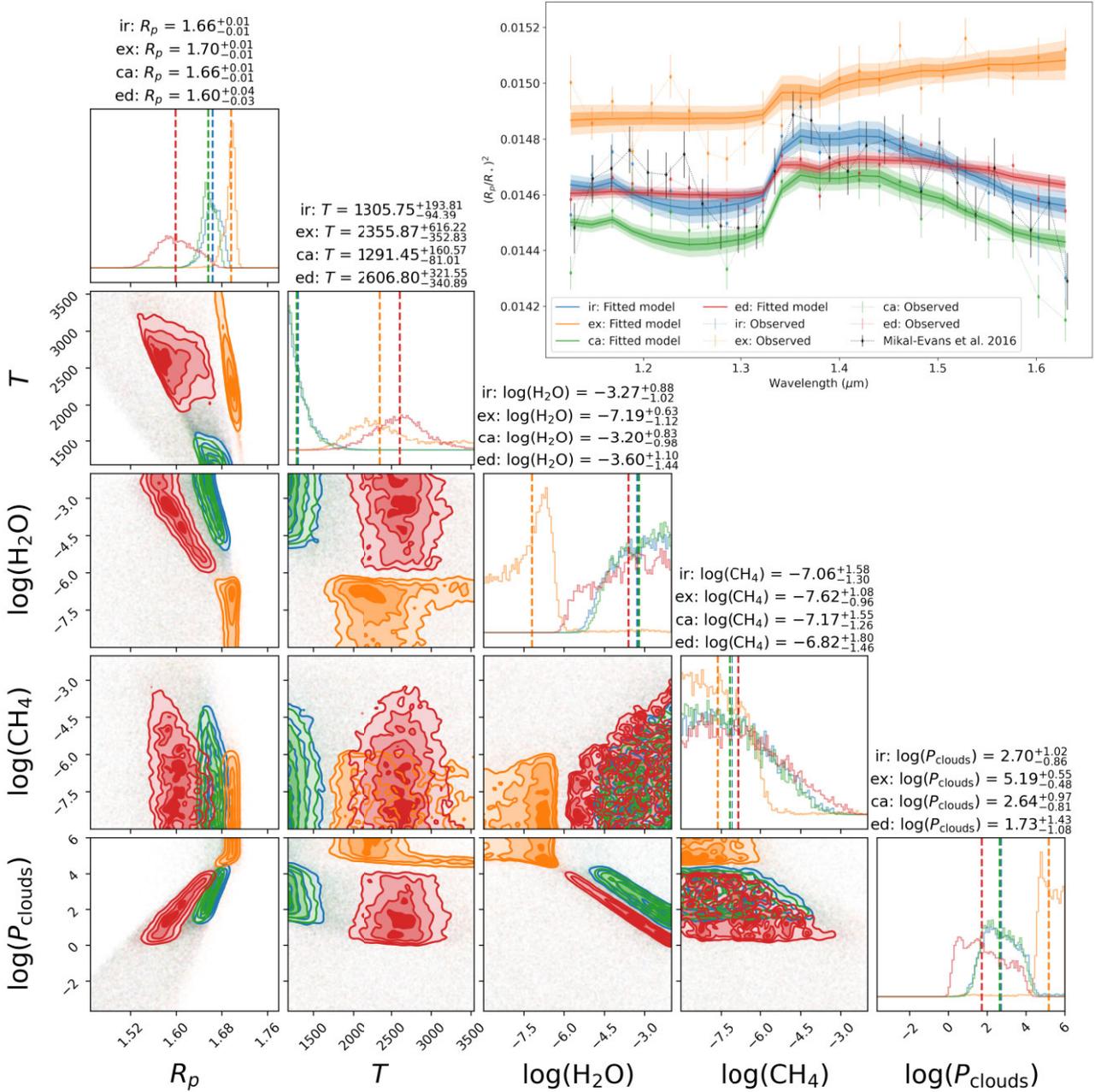
### 4.2 The case of WASP-121b

In our analysis, WASP-121b emerges as a unique entity, being the sole planet shared across all four data sets. Consequently, we present a comparative study of the spectral retrieval results for all its spectra. It is pertinent to reiterate that both the Iraclis and Edwards2022 data sets employ the IRACLIS pipeline for data reduction. Notably, WASP-121b is among the planets previously analysed in Tsiraras et al. (2018) and subsequently reprocessed in Edwards et al. (2022). The repeated data detrending is attributed to the data's availability: while Tsiraras et al. (2018) utilized a single transit observation from proposal ID: 14 468 by Thomas Mikal-Evans, Edwards et al. (2022) incorporated two transit observations from proposal ID: 15134, again by Thomas Mikal-Evans. As detailed in Edwards et al. (2022), the analysis extended beyond incorporating two new data sets into the WASP-121b spectrum; it also involved a comprehensive reevaluation of the data set previously processed in Tsiraras et al. (2018). This reevaluation was undertaken to ensure uniformity in methodology, parameters, and limb-darkening coefficients across all three transit fits. A comparative review of tables 8 and 9 from Edwards et al. (2022) with table 2 from Tsiraras et al. (2018) confirms the consistency in stellar and planetary parameters employed for detrending. Given this uniformity, we posit that the observed differences likely stem from variances in the pipeline [potentially due to an updated version used by Edwards et al. (2022)] or discrepancies in the estimation of limb-darkening coefficients.

Another salient point is that the spectrum in the Cascade data set was generated using the CASCADE pipeline, incorporating system parameters from Roudier et al. (2021). This ensures that the initial parameters used to produce the Excalibur data set spectra are consistent.

Upon executing the spectral retrieval as detailed in Section 2.2, we present the resultant corner plots and the derived fitted spectra in Fig. 11. The corner plot distinctly demarcates the regions explored by Excalibur (orange) and Edwards (red). In contrast, Iraclis (blue) and Cascade (green) exhibit remarkable congruence across all panels.

The retrieved parameters are not universally consistent. They yield spectra that pave the way for varied interpretations, especially concerning the planetary temperature, water content, and to a lesser extent, the radius. It is noteworthy that for this planet, Table 3 enumerates an effective temperature (calculated via equation 1) of 2298 K and a radius of  $1.865 R_{\text{jup}}$ . All pipelines deduce a diminished radius for the planet. Only Edwards2022 and Excalibur yield a temperature in alignment with the projected effective one. In stark contrast, Iraclis and Cascade infer a temperature nearly half the anticipated value, thereby reaching the boundary of our retrieval priors. However, it is known that equilibrium temperatures are often biased to cooler than expected temperatures (MacDonald, Goyal & Lewis 2020). On the molecular abundance front, Excalibur uniquely seems to negate the presence of water. Similarly, Excalibur posits that a cloud deck should reside at a pressure higher than that estimated by other pipelines. Broadly, Iraclis and Cascade appear harmonious across every panel, as evidenced by the overlapping blue and green hues. However, a closer inspection of the top-right panel in Fig. 11 reveals that the blue and green fitted atmospheric modulations are not congruent



**Figure 11.** The corner plot delineates the retrieval results for all WASP-121b spectra. The colour coding remains consistent with the rest of the paper: blue represents the Iraclis data set, red signifies Edwards2022, orange denotes Excalibur, and green symbolizes Cascade. Atop each panel column, the fitted values for each data set are displayed: ‘ir’ corresponds to the Iraclis data set, ‘ex’ to Excalibur, ‘ca’ to Cascade, and ‘ed’ to Edwards. The top right panel illustrates the fitted transmission spectra derived from the parameters retrieved for each data set, represented as solid lines. The corresponding filled areas indicate the  $1\sigma$  and  $2\sigma$  uncertainties. The observed data points, along with their uncertainties, are depicted using the same colour scheme. The observed data points are connected with coloured dotted lines, to help the reader. The black data points in the top right panel are from Evans et al. (2016) for comparison.

within the  $2\sigma$  range. Despite the apparent alignment in retrieved values, the two spectra depict markedly distinct atmospheres.

## 5 CONCLUSIONS

In an ideal scenario, all pipelines being compared would be run using the same system parameters and the same exoplanet observations; our analysis is able to make this rigorous, direct comparison for the EXCALIBUR and CASCADE pipelines but, due to the legacy nature

of the published IRACLIS results, performing a similar analysis was beyond the scope of this initial study. None the less, we believe that including the Iraclis results in the comparison exercise helps illustrate the multifaceted nature of comparing pipelines on a catalogue basis.

Our study shows the significant impact of different data reduction pipelines on the transmission spectra of exoplanets. The analysis of differences in mean values indicates that no pipeline has a bias towards larger or smaller radii compared to the others. However, significant inconsistencies in estimated radius values are common.

Regarding the different error bars, we observe that the CASCADE pipeline seems to produce marginally smaller error bars, compared to other pipelines, while EXCALIBUR produces the largest. However, in general, the error bars can be considered consistent within the data sets. The spectral values comparison shows that the shape of the spectra produced with different pipelines seems to be consistent within  $2\sigma$  – however, this approximate consistency is misleading.

From our retrieval analysis, we find there can be significant differences in basic planet parameters such as radius and temperature. In terms of atmospheric composition, we find that IRACLIS and Cascade typically yield analogous chemistry results and appear to favour water detection over methane. Conversely, spectra derived from the EXCALIBUR pipeline seem receptive to a broader mixing ratio spectrum for both molecules. It is pivotal to clarify that this study is not an evaluation of which pipeline offers superior or more trustworthy results. Instead, our objective is to gauge the ramifications of pipeline-induced differences at a population level, paving the way for the comparative planetology era anticipated with telescopes like *JWST* and *ARIEL*. Further analysis of WASP-121b reveals that, while different pipelines like Excalibur and Cascade can produce varying spectra, consistent atmospheric parameters can still be retrieved from similar spectra, as demonstrated by the CASCADE and IRACLIS pipelines. This indicates that despite the large parameter space, comparable spectra can lead to consistent parameter estimation for WASP-121b.

The most significant finding from our study is a clear demonstration that pipelines have the potential to inject systematic bias in the inference of population-level composition trends. The inconsistencies between data sets from different instruments, as well as the potential skewing of analysis due to residual systematics, present challenges that warrant further research.

Reflecting on our initial goals, we confirm that our investigation was aimed at identifying potential systematic biases in the results of exoplanet catalogues, stemming from specific pipelines. This study was not intended to trace the genealogy of pipeline-to-pipeline differences, an analysis that would require a more in-depth comparison of standardized intermediate data products.

In light of our findings, we recommend caution when interpreting results derived from different pipelines; in particular, agreement does not imply an absence of bias. We strongly advocate for further research into the systematics that influence the injection of bias and the origin of differences in pipeline results: a comprehensive and methodical examination into the causes of these discrepancies is essential through comparing standardized intermediate data products, in addition to the final transit spectra. Currently, the EXCALIBUR pipeline stands out as the only one to include such standardized intermediate data products in its transmission spectra catalogue, an approach we deem vital for further refining our observation and analysis techniques. Such efforts will not only clarify the sources of biases but also enhance the reliability of pipeline outputs, making them indispensable tools for advancing our understanding of exoplanetary atmospheres. It is through this meticulous scrutiny and resolution of pipeline-induced biases that we can achieve a robust and scientifically sound foundation for exoplanetary characterization.

In the future, we hope that this work will contribute to the development of more consistent and reliable practices in the field of exoplanetary science, ultimately leading to a more accurate understanding of exoplanets and their atmospheres. In particular, we encourage the development of pipelines to perform similar studies to compare the pipeline results not only for single planets but for entire populations, to make use of the statistics to extrapolate some pipeline properties and limitations.

## ACKNOWLEDGEMENTS

The authors thank Dr Angelos Tsiaras and Dr Jeroen Bouwman for the useful discussions and collaboration for the development of this paper. The authors also thank Dr Billy Edwards for sharing the processed spectra produced in Edwards et al. (2022). LVM was supported by ASI grant no. 2021.5.HH.O and UKSA grant no. ST/W002507/1.

## DATA AVAILABILITY

This work is based upon observations with the NASA/ESA *Hubble Space Telescope*, obtained at the Space Telescope Science Institute (STScI) operated by AURA, Inc. The publicly available *HST* observations were obtained from the Hubble Archive which is part of the *Mikulski Archive for Space Telescopes (MAST)*.

## REFERENCES

- Abel M., Frommhold L., Li X., Hunt K. L., 2011, *J. Phys. Chem. A*, 115, 6805
- Abel M., Frommhold L., Li X., Hunt K. L., 2012, *J. Chem. Phys.*, 136, 044319
- Ahrer E.-M. et al., 2023, *Nature*, 614, 653
- Alderson L. et al., 2023, *Nature*, 614, 664
- Alexoudi X. et al., 2018, *A&A*, 620, A142
- Al-Refaie A. F., Changeat Q., Waldmann I. P., Tinetti G., 2021, *ApJ*, 917, 37
- Anisman L. O., Edwards B., Changeat Q., Venot O., Al-Refaie A. F., Tsiaras A., Tinetti G., 2020, *AJ*, 160, 233
- Barstow J. K., Aigrain S., Irwin P. G. J., Sing D. K., 2017, *ApJ*, 834, 50
- Barstow J. K., Changeat Q., Garland R., Line M. R., Rocchetto M., Waldmann I. P., 2020, *MNRAS*, 493, 4884
- Barstow J. K., Changeat Q., Chubb K. L., Cubillos P. E., Edwards B., MacDonald R. J., Min M., Waldmann I. P., 2022, *Exp. Astron.*, 53, 447
- Bruno G. et al., 2020, *MNRAS*, 491, 5361
- Buchner J. et al., 2014, *A&A*, 564, A125
- Carone L. et al., 2021, *A&A*, 646, A168
- Chachan Y. et al., 2019, *AJ*, 158, 244
- Changeat Q., Al-Refaie A., Mugnai L. V., Edwards B., Waldmann I. P., Pascale E., Tinetti G., 2020a, *AJ*, 160, 80
- Changeat Q., Edwards B., Al-Refaie A. F., Morvan M., Tsiaras A., Waldmann I. P., Tinetti G., 2020b, *AJ*, 160, 260
- Charbonneau D., Brown T. M., Noyes R. W., Gilliland R. L., 2002, *ApJ*, 568, 377
- Claret A., 2000, *A&A*, 363, 1081
- Constantinou S., Madhusudhan N., Gandhi S., 2023, *ApJ*, 943, L10
- Deming D. et al., 2013, *ApJ*, 774, 95
- Diamond-Lowe H., Stevenson K. B., Bean J. L., Line M. R., Fortney J. J., 2014, *ApJ*, 796, 66
- Edwards B. et al., 2020, *AJ*, 160, 8
- Edwards B. et al., 2022, *AJ* 269 31
- Estrela R., Swain M. R., Roudier G. M., 2022, *ApJ*, 941, L5
- Evans T. M. et al., 2016, *ApJ*, 822, L4
- Feinstein A. D. et al., 2023, *Nature*, 614, 670
- Feroz F., Hobson M. P., Bridges M., 2009, *MNRAS*, 398, 1601
- Fletcher L. N., Gustafsson M., Orton G. S., 2018, *ApJS*, 235, 24
- Fox C., Wiegert P., 2021, *MNRAS*, 501, 2378
- Gandhi S., Madhusudhan N., 2017, *MNRAS*, 472, 2334
- Guilluy G. et al., 2021, *AJ*, 161, 19
- Holmberg M., Madhusudhan N., 2023, *MNRAS*, 524, 377
- Huber-Feely N., Swain M. R., Roudier G., Estrela R., 2022, *AJ*, 163, 22
- Irwin P. G. J. et al., 2008, *J. Quant. Spectrosc. Radiat. Transfer*, 109, 1136
- Iyer A. R., Swain M. R., Zellem R. T., Line M. R., Roudier G., Rocha G., Livingston J. H., 2016, *ApJ*, 823, 109
- Kempton E. M. R. et al., 2018, *PASP*, 130, 114401
- Kipping D., 2020, *ApJ*, 900, L44

- Kirk J., López-Morales M., Wheatley P. J., Weaver I. C., Skillen I., Louden T., McCormac J., Espinoza N., 2019, *AJ*, 158, 144
- Kreidberg L. et al., 2014, *Nature*, 505, 69
- Lahuis F., Bouwman J., Lagage P. O., Martin-Lagarde M., Min M., Waldman I., 2020, in Pizzo R., Deul E. R., Mol J. D., de Plaa J., Verkouter H. eds, ASP Conf. Ser. Vol. 527, *Astronomical Data Analysis Software and Systems XXIX*. Astron. Soc. Pac., San Francisco, p. 179
- Lavie B. et al., 2017, *AJ*, 154, 91
- Lee J.-M., Heng K., Irwin P. G. J., 2013, *ApJ*, 778, 97
- Libby-Roberts J. E. et al., 2022, *AJ*, 164, 59
- Line M. R. et al., 2013, *ApJ*, 775, 137
- MacDonald R. J., Goyal J. M., Lewis N. K., 2020, *ApJ*, 893, L43
- Madhusudhan N., Seager S., 2009, *ApJ*, 707, 24
- Morello G., Tsiaras A., Howarth I. D., Homeier D., 2017, *AJ*, 154, 111
- Morello G., Claret A., Martin-Lagarde M., Cossou C., Tsiaras A., Lagage P. O., 2020, *AJ*, 159, 75
- Mugnai L. V. et al., 2021a, *AJ*, 161, 284
- Mugnai L. V., Al-Refaie A., Bocchieri A., Changeat Q., Pascale E., Tinetti G., 2021b, *AJ*, 162, 288
- Pluriel W. et al., 2020, *AJ*, 160, 112
- Polyansky O. L., Kyuberis A. A., Zobov N. F., Tennyson J., Yurchenko S. N., Lodi L., 2018, *MNRAS*, 480, 2597
- Roudier G. M., Swain M. R., Gudipati M. S., West R. A., Estrela R., Zellem R. T., 2021, *AJ*, 162, 37
- Rustamkulov Z. et al., 2023, *Nature*, 614, 659
- Saba A. et al., 2022, *AJ*, 164, 2
- Schölkopf B., Hogg D. W., Wang D., Foreman-Mackey D., Janzing D., Simon-Gabriel C.-J., Peters J., 2016, *Proc. Natl. Acad. Sci.*, 113, 7391
- Sing D. K. et al., 2016, *Nature*, 529, 59
- Siverd R. J. et al., 2012, *ApJ*, 761, 123
- Skaf N. et al., 2020, *AJ*, 160, 109
- Stevenson K. B., Bean J. L., Seifahrt A., Désert J.-M., Madhusudhan N., Bergmann M., Kreidberg L., Homeier D., 2014a, *AJ*, 147, 161
- Stevenson K. B., Bean J. L., Fabrycky D., Kreidberg L., 2014b, *ApJ*, 796, 32
- Swain M. R., Vasisht G., Tinetti G., 2008, *Nature*, 452, 329
- Swain M. R. et al., 2009, *ApJ*, 704, 1616
- Swain M. R. et al., 2021, *AJ*, 161, 213
- Tinetti G. et al., 2007, *Nature*, 448, 169
- Tsiaras A. et al., 2016a, *ApJ*, 820, 99
- Tsiaras A., Waldmann I. P., Rocchetto M., Varley R., Morello G., Damiano M., Tinetti G., 2016b, *ApJ*, 832, 202
- Tsiaras A. et al., 2018, *AJ*, 155, 156
- Tsiaras A., Waldmann I. P., Tinetti G., Tennyson J., Yurchenko S. N., 2019, *Nat. Astron.*, 3, 1156
- Waldmann I. P., Tinetti G., Rocchetto M., Barton E. J., Yurchenko S. N., Tennyson J., 2015, *ApJ*, 802, 107
- Yip K. H., Tsiaras A., Waldmann I. P., Tinetti G., 2020, *AJ*, 160, 171
- Yip K. H., Changeat Q., Edwards B., Morvan M., Chubb K. L., Tsiaras A., Waldmann I. P., Tinetti G., 2021, *AJ*, 161, 4
- Yurchenko S. N., Tennyson J., 2014, *MNRAS*, 440, 1649

This paper has been typeset from a  $\text{\TeX}/\text{\LaTeX}$  file prepared by the author.