




# A lack of constraints on the cold opaque H I mass: H I spectra in M31 and M33 prefer multicomponent models over a single cold opaque component

Eric W. Koch <sup>1,2</sup>★ Erik W. Rosolowsky <sup>1</sup> Adam K. Leroy,<sup>3</sup> Jérémy Chastenet,<sup>4</sup>  
I-Da Chiang (江宜達),<sup>4</sup> Julianne Dalcanton,<sup>5</sup> Amanda A. Kepley,<sup>6</sup> Karin M. Sandstrom,<sup>4</sup>  
Andreas Schruba,<sup>7</sup> Snežana Stanimirović,<sup>8</sup> Dyas Utomo<sup>3,6</sup> and Thomas G. Williams <sup>9</sup>

<sup>1</sup>Department of Physics, University of Alberta, 4-183 CCIS, Edmonton, AB T6G 2E1, Canada

<sup>2</sup>Center for Astrophysics | Harvard & Smithsonian, 60 Garden St, Cambridge, MA 02138, USA

<sup>3</sup>Department of Astronomy, The Ohio State University, 140 West 18th Avenue, Columbus, OH 43210, USA

<sup>4</sup>Center for Astrophysics and Space Sciences, Department of Physics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

<sup>5</sup>Department of Astronomy, University of Washington, Box 351580 Seattle, WA 98195, USA

<sup>6</sup>National Radio Astronomy Observatory, 520 Edgemont Road, Charlottesville, VA 22903-2475, USA

<sup>7</sup>Max-Planck-Institut für extraterrestrische Physik, Giessenbachstraße 1, D-85748 Garching, Germany

<sup>8</sup>Department of Astronomy, University of Wisconsin, 475 N Charter St, Madison, WI 53706, USA

<sup>9</sup>Max Planck Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

Accepted 2021 April 1. Received 2021 March 29; in original form 2021 January 20

## ABSTRACT

Previous work has argued that atomic gas mass estimates of galaxies from 21-cm H I emission are systematically low due to a cold opaque atomic gas component. If true, this opaque component necessitates a  $\sim 35$  per cent correction factor relative to the mass from assuming optically thin H I emission. These mass corrections are based on fitting H I spectra with a single opaque component model that produces a distinct ‘top-hat’ shaped line profile. Here, we investigate this issue using deep, high spectral resolution H I VLA observations of M31 and M33 to test if these top-hat profiles are instead superpositions of multiple H I components along the line of sight. We fit both models and find that  $>80$  per cent of the spectra strongly prefer a multicomponent Gaussian model while  $<2$  per cent prefer the single opacity-corrected component model. This strong preference for multiple components argues against previous findings of lines of sight dominated by only cold H I. Our findings are enabled by the improved spectral resolution ( $0.42 \text{ km s}^{-1}$ ), whereas coarser spectral resolution blends multiple components together. We also show that the inferred opaque atomic ISM mass strongly depends on the goodness-of-fit definition and is highly uncertain when the inferred spin temperature has a large uncertainty. Finally, we find that the relation of the H I surface density with the dust surface density and extinction has significantly more scatter when the inferred H I opacity correction is applied. These variations are difficult to explain without additionally requiring large variations in the dust properties. Based on these findings, we suggest that the opaque H I mass is best constrained by H I absorption studies.

**Key words:** ISM: kinematics and dynamics – galaxies: individual: M33 – galaxies: individual: M31 – galaxies: ISM – radio lines: ISM.

## 1 INTRODUCTION

The fundamental difficulty in constraining the atomic interstellar medium (ISM) gas mass is the unknown temperature and density structure of the atomic gas along the line of sight. The temperature and density of atomic hydrogen (H I) are expected to vary by a factor of  $\sim 100$  between the warm neutral medium (WNM) and cold neutral medium (CNM) (Field, Goldsmith & Habing 1969; Wolfire et al. 2003), with an unstable intermediate state between these two phases (Heiles & Troland 2003) that accounts for  $<20$  per cent of the gas (Hill et al. 2018; Murray et al. 2018a). There are consequences on the resulting 21-cm H I spectra due to variation of the line opacity in these different states. The WNM is primarily optically thin, while

the CNM can become optically thick at higher densities. However, the relative importance of these two components can be difficult to assess.

The CNM can be unambiguously detected in absorption against a bright background source (e.g. Murray et al. 2018a; Jameson et al. 2019) which allows the spin temperature and opacity to be measured. By comparing to the nearby H I emission, the fraction of CNM, WNM, and intermediate phases can then be estimated. However, there is limited spatial coverage for lines of sight towards sufficiently bright background sources to measure the H I absorption,<sup>1</sup> making it difficult to constrain the total mass in each atomic ISM phase. These

<sup>1</sup>Additionally, H I self-absorption, where CNM lies in the foreground of bright WNM emission, can provide better spatial coverage where this geometry along a line of sight is available (Gibson et al. 2005).

\* E-mail: [koch.eric.w@gmail.com](mailto:koch.eric.w@gmail.com)

absorption techniques are most useful for Milky Way observations where studies can choose to focus on the brightest quasars as background sources since H I is detected along most lines of sight from our location in the Galactic disc (Heiles & Troland 2003; Murray et al. 2018a; Wang et al. 2020b). Furthermore, the comparison between H I absorption and emission to constrain the atomic ISM phases requires similar angular resolution in both observations, which is easiest to achieve in the Milky Way but becomes difficult in external galaxies. In contrast, spatially contiguous H I emission can easily be traced but it is difficult to separate the atomic ISM phases from the emission alone due to overlapping spectral components and turbulent line broadening.

Extragalactic H I observations typically resolve  $>100$  pc scales, resulting in observed spectra that trace a mixture of unresolved WNM and CNM components. The WNM contributes a larger fraction of the atomic mass and has a larger volume-filling factor relative to the CNM (Ferrière 2001; Murray et al. 2018a), and thus the extragalactic H I emission is usually assumed to be optically thin. This assumption is convenient for measuring the atomic ISM mass as the optically thin H I integrated intensity is (i) independent of the spin temperature and (ii) proportional to the column density. Since the contribution from opaque CNM emission is not accounted for, using the optically thin assumption will underestimate the atomic ISM mass. However, when the CNM structures are unresolved on these large scales, it is difficult to estimate what the missing opaque H I mass fraction is.

One potential solution to the unknown H I opacity is to explore H I emission spectra on  $<100$  pc scales where the largest CNM structures may be resolved (similar to giant molecular cloud scales) as observed in H I self-absorption (HISA) in the Milky Way (Gibson et al. 2005; Wang et al. 2020b). If H I emission is resolved (or at least fills the telescope beam) but is opaque, the spectrum is no longer Gaussian and instead is predicted to have a ‘top-hat’ shape, where the observed brightness temperature saturates at the H I spin temperature<sup>2</sup>  $T_s$  when the emission becomes optically thick (Rohlfs, Braunsfurth & Mebold 1972). Observing such features provides a potentially powerful method to identify the opaque CNM since the opaque ‘top-hat’ shape saturates at  $T_s$  and constrains the opacity  $\tau$ .

With current radio interferometers, resolving  $<100$  pc scales limits H I observations to the nearest galaxies, in particular, those in the Local Group ( $D < 1$  Mpc). Braun et al. (2009) explored this approach in M31 with  $\sim 100$  pc H I observations, where they noted the presence of ‘top-hat’ shaped spectra and modelled these spectra as arising from a single cold, opaque H I component. They further advocated that these features are arranged into  $\sim 100$  pc cold H I clouds. Braun (2012) extended this modelling to H I observations of the Large Magellanic Cloud (LMC) and M33 (Kim et al. 2003; Gratier et al. 2010). Together, they estimated that correcting for the H I opacity accounts for an additional  $\sim 35$  per cent of H I mass relative to the mass from the optically thin assumption. Braun (2012) then used this correction factor to constrain the opacity-corrected H I distribution function with redshift  $z = 0$ .

An implicit assumption in the Braun et al. (2009) and Braun (2012) modelling is that most lines of sight through these galaxies are dominated by a single H I component, and further, that that component is primarily cold ( $<100$  K) H I. This differs from the structure typically observed in the Milky Way and Magellanic Clouds, where few H I studies find ‘top-hat’ H I components (Lee et al. 2015; Jameson et al. 2019). Of particular note is the lack of top-

hat spectra from Stanimirović et al. (2014) on  $\sim 1$  pc scales towards the Perseus molecular clouds. Since the Braun et al. (2009) model identifies beam-filling opaque emission, the lack of such features in Perseus on much smaller scales raises the question of whether these conclusions are driven by the analysis approach or are unique to the environments observed.

Comparisons with other tracers of the neutral ISM column density also show different trends compared to the H I opacity-corrected column density from the Braun et al. (2009) method. For example, molecular clouds traced by bright CO emission have the highest neutral ISM column density and thus are likely places where opaque H I should be detected. However, there is not a strong relation between CO emission and the opaque H I regions from Braun et al. (2009, their figs 11, 12, and 16), in contrast to what is seen in the Milky Way, where the largest HISA structures are strongly correlated with CO emission and molecular gas (Gibson et al. 2005; Wang et al. 2020b).

Studies that do find top-hat H I components note the difficulty in distinguishing from two overlapping velocity components (Rohlfs et al. 1972). Some studies use spatial continuity to confirm such features are indeed from  $>1$  H I components overlapping in velocity (Peek et al. 2011). One limitation of the Braun et al. (2009) and Braun (2012) studies is the  $1.4\text{--}2.3$  km s<sup>−1</sup> velocity resolution of their observations. Assuming a spectrally resolved Gaussian component requires sampling with channels  $\sim 2\times$  the RMS line width (Koch, Rosolowsky & Leroy 2018a), and thus the minimum thermal temperatures for a resolved component with these channel widths are 1300–3600 K, far larger than the CNM temperatures. Our initial work with these new observations in M33 (Koch et al. 2018b) demonstrated that a finer spectral resolution can spectrally resolve features that may not have been evident in previous observations.

Given these open questions, we investigate the topic of H I opacity using new H I observations of M31 and M33 with  $0.42$  km s<sup>−1</sup> spectral resolution taken with the Karl G. Jansky Very Large Array (VLA). In particular, we address the question: do the observed H I line shapes in M31 and M33 imply a single opaque H I component, as suggested by Braun et al. (2009), or are the spectra better described by multiple components along the line of sight? Answering this question has broad implications for our understanding of the neutral ISM in the nearby Universe, since the ISM mass is dominated by the atomic phase (Saintonge et al. 2017; Walter et al. 2020), and constraints on the cosmic density of H I in galaxies at  $z = 0$  depend on our ability to accurately measure the atomic gas mass. In this work, the key observational improvement is the fine spectral resolution that can resolve the  $0.85$  km s<sup>−1</sup> thermal line width of 120 K cold H I gas while spatially resolving 60–300 pc (accounting for galaxy inclination), closely matching the spatial scales used in previous studies.

We introduce the new H I observations in Section 2 and the model fitting and metric for comparison in Section 3. In Section 4, we use the model comparison and present our main result: most H I spectra in M31 and M33 are best modelled with a multicomponent Gaussian model. We then compare the atomic surface density to tracers of dust in Section 5, where we find that including the inferred opaque H I correction would require large variations in dust properties to explain. From these results, we discuss in Section 6 the discrepancies with previous work, the uncertainty on the opacity-corrected atomic ISM mass, comparisons with H I absorption studies, and the future outlook on using H I emission surveys to identify cold atomic ISM. Finally, we summarize our conclusions in Section 7.

Throughout this work, we adopt distances of 744 kpc for M31 (Vilardell et al. 2010) and 840 kpc for M33 (Freedman et al. 2001).

<sup>2</sup>For the simplest case of a single  $T_s$  along the line of sight.

**Table 1.** Summary of 21-cm H I observations from this work and those from Braun et al. (2009) and Braun (2012, see their table 1). The M33 VLA observations used in Braun (2012) are presented in Gratier et al. (2010) (see also Thilker, Braun & Walterbos 2002). The linear resolution range accounts for the galaxy inclination.

Reference	Target	Resolution (arcsec)	Linear resolution (pc) Major $\times$ Minor axis	Channel Width (km s <sup>-1</sup> )	Noise per channel ( $\sigma_{\text{rms}}$ ; K)	$5\sigma_{\text{rms}}$ optically thin column density (cm <sup>-2</sup> )	Observatory
Braun et al. (2009)	M31	30	100 $\times$ 500	2.3	1.0	$2.0 \times 10^{19}$	WSRT
This work	M31	18	60 $\times$ 300	0.42	2.8	$9.8 \times 10^{18}$	VLA
Braun (2012)	M33	30	130 $\times$ 205	1.4	2.1	$2.6 \times 10^{19}$	VLA
This work	M33	19	80 $\times$ 130	0.42	2.0	$7.0 \times 10^{18}$	VLA

## 2 OBSERVATIONS

### 2.1 VLA and single-dish H I observations

We obtained new H I observations for M31 and utilize our recent H I data set for M33 obtained by the VLA. Table 1 presents basic properties of these VLA observations and those used in Braun et al. (2009) and Braun (2012) for comparison. The key property of our new observations is the fine spectral resolution (0.42 km s<sup>-1</sup>), which is a factor of 3–5 finer than in previous observations.

The M33 observations are presented and described in detail in Koch et al. (2018b); we briefly summarize the observations here. We observed a 13-pointing mosaic that covers the inner  $\sim 12$  kpc of M33 in C-configuration (project 14B-088) for a total of 52 h. The nearby source 3C 48 was used as the flux, delay, and gain calibrator. The H I line was observed with a 977 Hz resolution, corresponding to a spectral channel width of 0.206 km s<sup>-1</sup>, and the naturally weighted beam size is 19 arcsec ( $\sim 80$  pc). These VLA observations were combined with existing GBT data to provide short-spacing information (see Lockman, Free & Shields 2012). To match the M31 spectral resolution of 0.42 km s<sup>-1</sup> (see below), we spectrally downsample the 19 arcsec M33 H I cube from the original 0.206 km s<sup>-1</sup> from Koch et al. (2018b). The typical H I spectral widths identified in Koch et al. (2018b) and Koch et al. (2019) are larger than 0.42 km s<sup>-1</sup> and so we do not expect downsampling the spectral resolution by a factor of 2 will affect measurements of the spectral shapes.

The new M31 observations cover a 7-point mosaic in the northern half of the galaxy that overlaps with the *Hubble Space Telescope* observations from the PHAT survey (Dalcanton et al. 2012) and the CARMA CO(1-0) survey field (Schruba et al., in preparation; for a first presentation see Caldu-Primo & Schruba 2016). We combined observations taken in the VLA’s B, C, and D-configurations<sup>3</sup> (projects 14A-235 and 15A-175) and use a *uv*-taper when imaging to produce a mosaic with 18 arcsec beam size, roughly matching the C-configuration VLA mosaic of M33. The H I line was observed with 1.95 kHz spectral channels, corresponding to a velocity resolution of 0.42 km s<sup>-1</sup>. We reduced the data using the standard VLA pipeline with CASA v5.4.1, with minor modifications for spectral line data, including: (i) splitting the continuum and line spectral windows; (ii) disabling Hanning smoothing to retain the 0.42 km s<sup>-1</sup> velocity resolution for the H I data; and (iii) avoiding RFI flagging and reweighting (STATWT) over M31’s velocity range. Due to the overlap between the red-shifted side of M31 and Milky Way H I emission, we also flagged velocity ranges with H I absorption towards the

bandpass calibrator and interpolated across this gap. For each track, we iteratively reran the pipeline after checking the data and applying (typically) minor manual flagging. Most tracks required 2 pipeline runs to produce well-calibrated data.

We imaged the M31 H I data in 0.42 km s<sup>-1</sup> channels using a similar approach to that used for the M33 imaging described in Koch et al. (2018b). To include short-spacing information, we combined the VLA H I cube with the single-dish data from the Effelsberg-Bonn H I Survey (EBHIS; Winkel et al. 2016). In Appendix A, we describe the imaging approach and combination with the EBHIS data in detail.

Fig. 1 shows the H I integrated intensity maps for M31 and M33 at the resolution used in this work. The remainder of this section describes additional data handling required for this work.

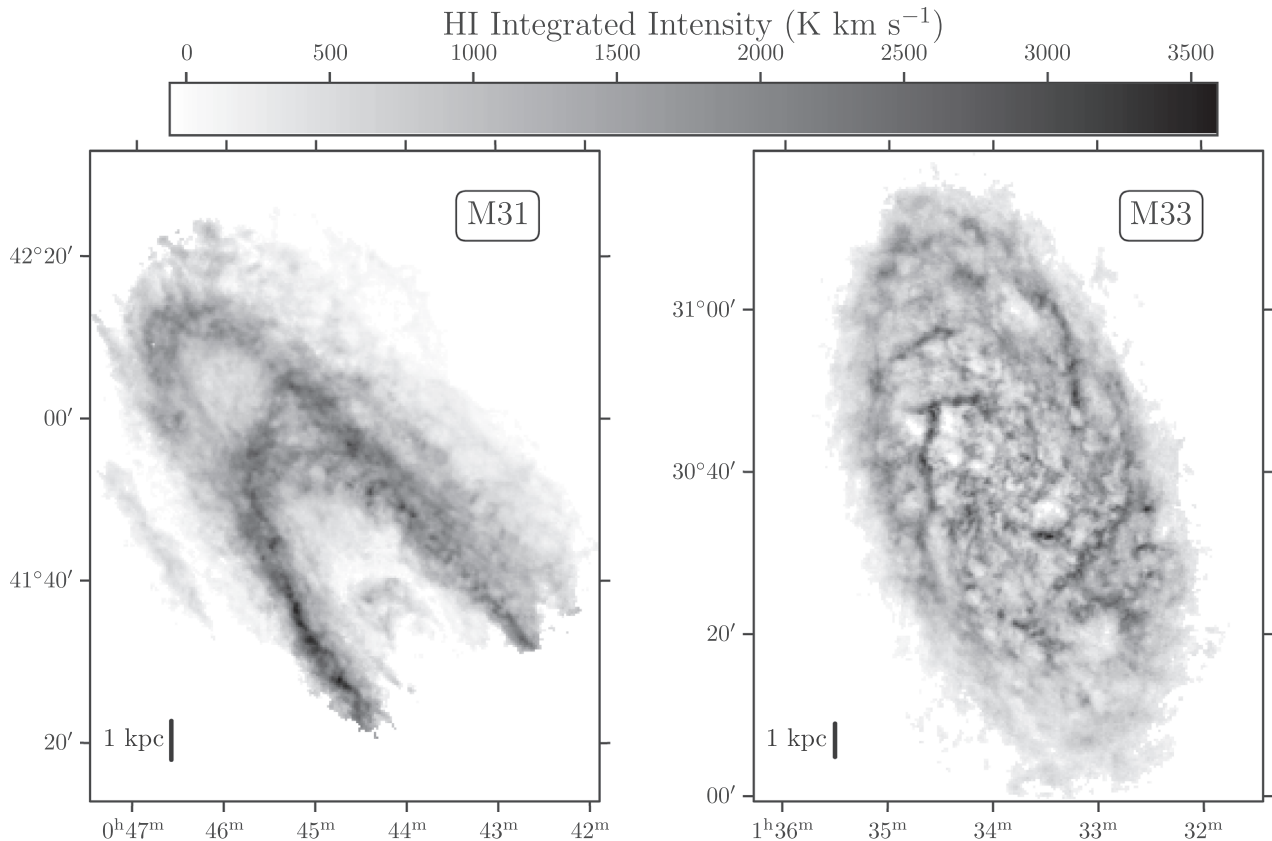
For the M31 map, the bright background radio source at ICRS 0<sup>h</sup>46<sup>m</sup>48<sup>s</sup> + 42°08′56″ falls within the field of view leading to strong H I absorption features from M31 (labelled as Source #497 in the catalogue from Braun 1990). These features persist after subtracting the continuum emission from the *uv* data prior to imaging. We mask the region around this source by defining a region 1.5  $\times$  the beam area to avoid fitting strong absorption features, for which the methods in Section 3 are not suited.

The per-channel rms noise is 2.8 K for M31 and 2.0 K for M33 in the spectrally matched data. Since strong emission continues to the edges of the observed area, particularly for M31 (Fig. 1), the noise varies with the effective primary beam for the mosaic. We weight the noise levels used for the spectral fitting in Section 3 by the primary beam, which increases the noise levels by up to  $\sim 5\times$  at the edges of the mosaic. Since the M31 mosaic only covers a part of the galaxy, H I signal is detected in some places along the edge of the mosaic and is affected by this increase in the noise.

Throughout the paper we use the M31 and M33 H I cubes at their respective angular resolutions. Because of the differences in distance and inclination between the two galaxies, the linear physical resolution overlaps at the respective original angular resolutions. The linear resolution for M31 is 60  $\times$  300 pc at 18 arcsec along the major and minor axes, respectively. This large range is due to M31’s large inclination angle of  $i = 78^\circ$  (Corbelli et al. 2010) that tilts the plane of sky in the observed frame. For M33, the linear resolution ranges from 80  $\times$  130 pc, which is smaller due to the lower inclination angle of  $i = 55.1^\circ$  (Koch et al. 2018b). We tested whether matching the angular resolution affects the model fits from Section 3 by fitting small regions of M31 and M33 at matched angular resolution. We found that the fits were consistent at the matched and original angular resolutions.

To convert the H I integrated intensity to atomic gas surface density, we use the optically thin factor 0.0196 M<sub>⊙</sub> pc<sup>-2</sup> (K km s<sup>-1</sup>)<sup>-1</sup> which includes a 1.36 factor for helium and metals.

<sup>3</sup>The D-configuration observations include a full mosaic of M31 that will be presented in Koch et al. (in preparation).



**Figure 1.** HI integrated intensity maps of M31 and M33 (see Table 1). Our new VLA HI map of M31 covers most of the Northern half of M31. The VLA HI map of M33 is presented and described in detail in Koch et al. (2018b).

## 2.2 Dust emission and extinction

Dust emission and dust extinction provide independent measurements of ISM column density since it is expected to be well mixed with gas on the  $>60$  pc scales we explore here. Here, we use both dust emission and extinction maps to compare with the gas column density inferred from HI.

First, we use the dust surface density maps of M31 and M33 from Utomo et al. (2019) with 167 pc resolution in the plane of sky. These surface density maps are derived from a modified blackbody model fit to the spectral energy distribution (SED) in five bands from the *Herschel Space Observatory* from 100 to 500  $\mu\text{m}$  using a broken power-law emissivity model, following the approach of Gordon et al. (2014) and Chiang et al. (2018).

Second, we use the dust extinction  $A_V$  map of M31 over the region observed by PHAT from Dalcanton et al. (2015). In this map, the dust extinction is modelled within a grid of 25 pc (6.5 arcsec) square regions by fitting the SEDs of  $\sim 10$ s of stars to a lognormal distribution in extinction  $A_V$ . From this modelling, we use the map of mean  $A_V$  values to compare with other tracers of the ISM column density.

## 3 MODELLING HI SPECTRA

The VLA HI observations we present have a similar sensitivity and spatial resolution as previous M31 and M33 observations. They differ, however, in their three to five times finer spectral resolution ( $0.42 \text{ km s}^{-1}$ ), which reveals ubiquitous spectral complexity in the HI spectra, as noted in Koch et al. (2018b) for M33.

Fig. 2 shows eight spectra – four from each galaxy – that demonstrate a range of HI line profiles. These different profiles present an interesting opportunity to compare models for HI spectra.

In this section, we explicitly test two models to describe the HI line shape: a single opaque HI component model, as advocated by Braun et al. (2009) and Braun (2012), and a multicomponent Gaussian model that implicitly assumes that profiles are due to a superposition of optically thin clouds along a line of sight. We first introduce these models, the fitting procedures, and the model selection used to compare between them.

### 3.1 Opaque HI model

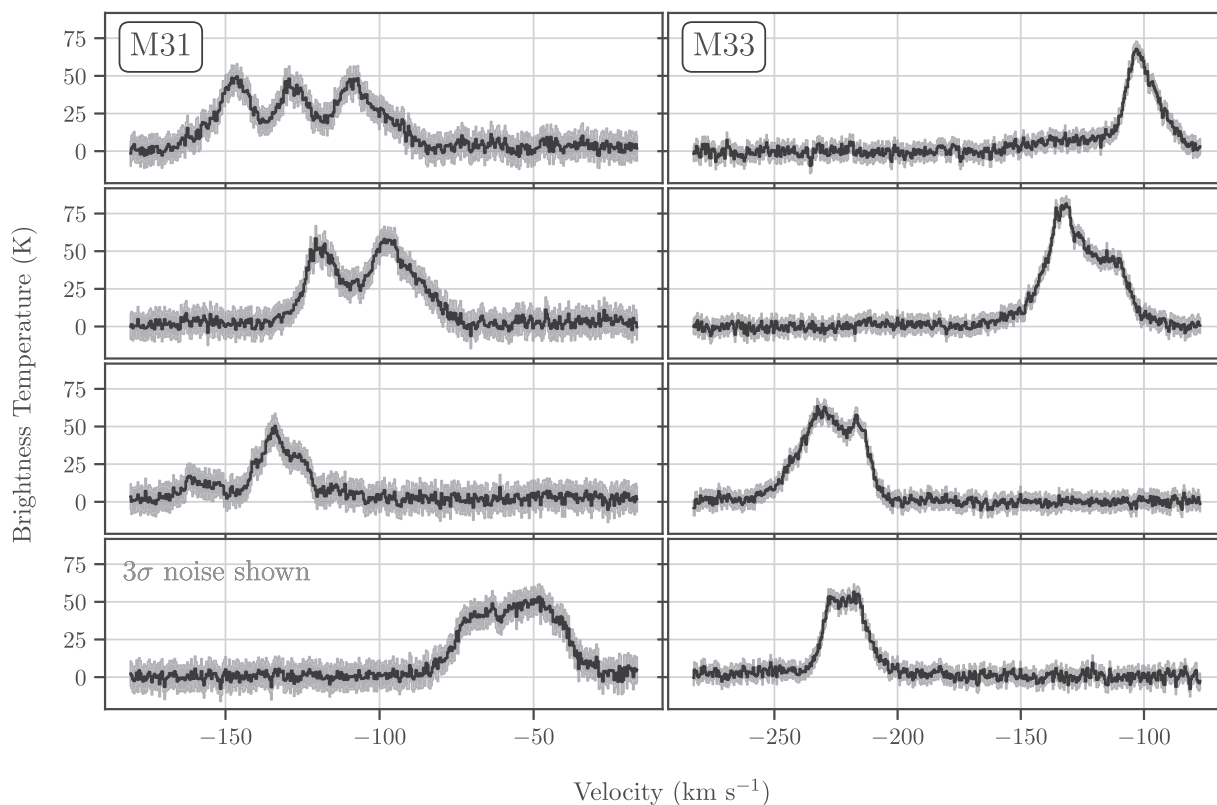
In the absence of a background continuum source, the general line profile for a single component of isothermal HI broadened by turbulence is

$$T_b(v) = T_s [1 - \exp[-\tau(v)]] \quad (1)$$

where  $v$  is the velocity,<sup>4</sup>  $T_b$  is the observed brightness temperature,  $T_s$  is the HI spin temperature, and  $\tau(v)$  is the optical depth profile. We also note that this assumes a diffuse continuum component is negligible (see equations 11 and 12 in Lee et al. 2015). Following Braun et al. (2009), we assume that the optical depth profile is a Gaussian with thermal and turbulent line broadening. The optical depth profile can then be expressed in terms of the HI column density

<sup>4</sup>Or an equivalent spectral quantity.





**Figure 2.** Example H I spectra in M31 (left) and M33 (right). The spectral channel width in all spectra is  $0.42 \text{ km s}^{-1}$ , and the  $3\sigma$  rms noise levels are shown in grey to emphasize that the spectral complexity is real. In both galaxies, we find distinct spectral features that consistently deviate from a single Gaussian profile. Most of these example spectra have multiple peaks, though notably, the spectra in the bottom row show the general ‘flat-top’ shape that may result from opaque H I. By consistently modelling the H I spectra with the multi-Gaussian and opaque models, we quantitatively distinguish between the appropriate model for the array of spectral shapes.

( $N_{\text{HI}}$ ;  $\text{cm}^{-2}$ ),  $T_s$  (K), and line width ( $\sigma$ ;  $\text{km s}^{-1}$ ) as

$$\tau(v) = \frac{5.49 \times 10^{-19} N_{\text{HI}}}{\sqrt{2\pi} T_s \sigma} \exp \left[ -\frac{(v - v_p)^2}{2\sigma^2} \right], \quad (2)$$

where  $v_p$  is the velocity at the peak in optical depth. The line width can be expressed in terms of a thermal and non-thermal component,  $\sigma = \sqrt{\sigma_t^2 + \sigma_{\text{nt}}^2}$ , where the non-thermal component is presumed to be dominated by turbulence. Similar to Braun et al. (2009), we assume the kinetic and spin temperatures are equal ( $T_k = T_s$ ) in the absence of additional constraints, which is correct for the CNM only. In the limit of high  $T_s$ , high  $\sigma$ , or low  $N_{\text{HI}}$ , we recover the optically thin limit of a Gaussian line shape since  $\exp[-\tau] \approx 1 - \tau$  for small  $\tau$ . Fig. 3 shows an example of the model shape from equations (1) and (2) where the opacity significantly affects the observed line shape that produces a ‘top-hat’ shape.

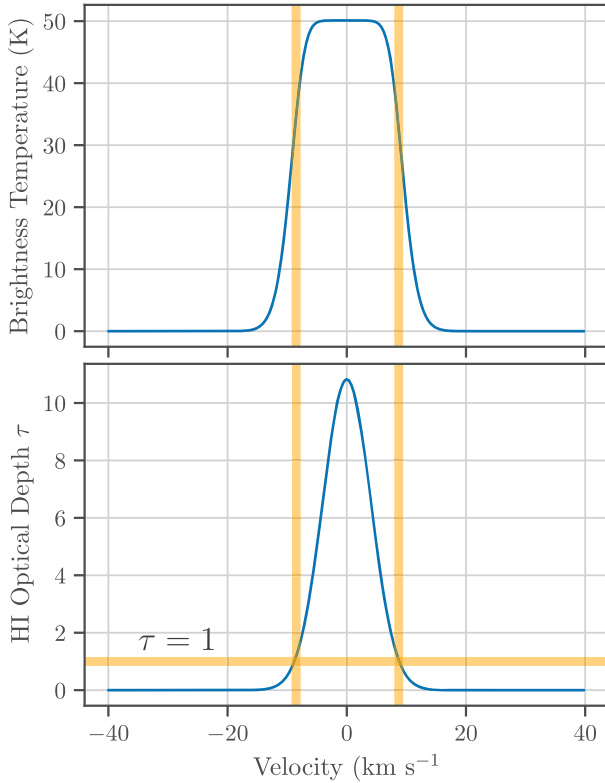
We stress that equation (1) describes the line profile for a single component of H I, though it is common in, for example, Milky Way H I observations for many components to be detected along a line of sight. The issue in extending this opacity-corrected H I model to multiple components is the strong degeneracy between the fit parameters. In particular, the ‘top-hat’ shape that constrains the line opacity must be spectrally resolved from other components, as blended structures can be fit by both a top-hat shape and a combination of Gaussians (see Fig. 3). For this reason, previous studies typically fit only a single component to each spectrum (Braun et al. 2009; Braun 2012), or at most two components (Peek et al. 2011). Here, we also fit only with a single opacity-

corrected component to compare with Braun et al. (2009) and Braun (2012).

While equation (1) assumes a single isothermal component, we expect that the H I measured within the beam will be a mixture of both WNM and CNM. For the H I opacity to be large enough to observe a top-hat shaped spectrum, the CNM must dominate by mass within the telescope beam. In our observations and those in Braun (2012), observing a top-hat shaped spectrum implies that cold atomic ‘clouds’ dominate the atomic ISM mass on  $\sim 100$  pc scales. If the WNM dominates instead, the observed spectrum will converge to the optically thin case, as is often assumed in studies of nearby galaxies with coarse spatial resolution.

Braun et al. (2009) and Braun (2012) fit H I spectra in M31, M33, and the LMC with  $\tau$  as defined in equation (2). They justified using an isothermal approximation to describe the H I based on the low filling fraction of cold H I clouds found by Braun (1997). This low filling fraction is inferred from the ‘high-brightness network’ of H I where Braun (1997) suggests the H I is most likely to be opaque. Braun et al. (2009) fit H I spectra by defining a grid of pre-computed models and minimizing the reduced  $\chi^2$  to identify the best-fitting model. To limit the range of possible models in the grid, the centre of the line was fixed to within  $\pm 5 \text{ km s}^{-1}$  of the velocity centroid, or about  $\pm 2$  channels in their observations. To summarize, the fitting method from Braun et al. (2009) has four free parameters:  $T_s$ ,  $N_{\text{HI}}$ ,  $\sigma_{\text{nt}}$ , and  $v_p$ .

Because three of the four free parameters contribute to the first term in the optical depth (equation 2), these parameters are strongly covariant in the fit. To measure and reduce the parameter covariance



**Figure 3.** A single opaque H I component model (equation 1; top) where the H I line opacity follows a Gaussian profile (equation 3; bottom). This example uses a peak temperature of  $T_p = 540$  K, a spin temperature of  $T_s = 50$  K, and a line width of  $\sigma = 4$  km s $^{-1}$ . The corresponding column density for this model is  $N_{\text{HI}} = 10^{22}$  cm $^{-2}$  ( $\Sigma_{\text{HI}} = 80$  M $_{\odot}$  pc $^{-2}$ ). The orange vertical lines show where the emission becomes opaque at  $\tau = 1$ . Within the opaque velocity region, the observed line profile has a distinct ‘top-hat’ shape that saturates at  $T_s$ . If the emission were to be treated as optically thin ( $\propto$  integrated intensity), the H I column density, and therefore mass, would be significantly underestimated.

from from equations (1) and (2), we make three changes to the fitting approach from Braun et al. (2009):

- (i) We fit the model using a Levenberg–Marquardt algorithm, as implemented in the SCIPY python package (Virtanen et al. 2020), instead of a grid search method. We then estimate the parameter uncertainty from the covariance matrix of the fit. Parameter covariances become important as the model approaches the optically thin limit, where  $\sigma$ ,  $N_{\text{HI}}$ , and  $T_s$  will be strongly covariant due to the lack of additional constraints on  $T_s$ .
- (ii) To minimize parameter covariance, we do not split the line width into a thermal and non-thermal component. This change is important when the profile approaches the optically thin limit and  $T_s$  is unconstrained, which leads to the thermal line width contribution being limited only by the upper limit of  $T_s$ . Here, we adopt an upper limit of  $T_s = 8000$  K for H I to match the range<sup>5</sup> used by Braun et al. (2009). When  $T_s$  is constrained, the non-thermal line width can be recovered by subtracting the thermal component in quadrature.
- (iii) We further decrease the parameter covariance in the model by simplifying the optical depth profile from equation (2). We define a

$T_p$  peak temperature such that<sup>6</sup>  $T_p = 5.49 \times 10^{-19} N_{\text{HI}} / \sqrt{2\pi} \sigma$ . Then the peak optical depth is simplified to be  $\tau_p = T_p / T_s$  and equation (2) becomes

$$\tau(v) = \left( \frac{T_p}{T_s} \right) \exp \left[ -\frac{(v - v_p)^2}{2\sigma^2} \right]. \quad (3)$$

For opaque H I, the observed peak temperature will converge to  $T_s$ , while for optically thin emission it will converge to  $T_p$ .

Using this change of variables removes the added covariance between  $T_s$  and  $\sigma$  from the thermal and non-thermal line width components, and introduces a parameter  $T_p$  with a similar range to  $T_s$ , replacing  $N_{\text{HI}}$  which varies over several orders of magnitude. The line opacity is then primarily set by the peak optical depth,  $\tau_p = T_s / T_p$ .

We keep similar constraints on the velocity of the line centre. Due to the finer spectral resolution of our data, a limit of  $\pm 5$  km s $^{-1}$  translates to  $\pm 12$  channels.

### 3.2 Multi-Gaussian model

We also fit the observed H I spectrum to a multicomponent model of Gaussians. Compared to the opaque H I model above, this multicomponent Gaussian model allows for multiple components along a line of sight, but restricts each component to a Gaussian shape, i.e. being optically thin. As we described below, overlapping Gaussians become degenerate and adding additional free parameters to account for opacity makes the fitting intractable, or at best, poorly constrained due to the strong covariance in parameters.

The general form for a multicomponent model comprised of  $N$  Gaussians is

$$T_b(v) = \sum_i^N T_{p,i}(v) = \sum_i^N T_{p,i} \exp \left[ -\frac{(v - v_{0,i})^2}{2\sigma_i^2} \right], \quad (4)$$

where each component is defined in terms of its peak temperature ( $T_{p,i}$ ), central velocity ( $v_{0,i}$ ), and line width ( $\sigma_i$ ).

Rohlfs et al. (1972) first point out that Milky Way H I spectra with flat-tops could also be explained by two or more Gaussian components blended together. The difficulty in distinguishing between a single optically thick component and multiple optically thin Gaussian components lies in the flexibility of a multi-Gaussian model: two or more Gaussians can be arranged to explain many spectral shapes and do not form an orthogonal basis set. In this section, we describe our approach for determining an appropriate number of Gaussian components to avoid overfitting, while accounting for spatial continuity to distinguish between one or more overlapping Gaussians.

Several recent studies present algorithmic approaches for (semi-)automated spectral line fitting of Gaussian components (e.g. Haud 2000; Lindner et al. 2015; Henshaw et al. 2016; Keown et al. 2019; Marchal et al. 2019; Sokolov et al. 2020). The method we use here combines elements from some of these studies to optimize the fitting for our particular data sets, though we highlight that we only include spatial information from nearest neighbours rather than fitting all neighbouring spectra together (Section 3.2.3). From our early testing of different methods, we note that most spectral line fitting algorithms

<sup>5</sup>We note that  $T_s = 8000$  K falls within the WNM range found by Murray et al. (2014) from deep H I absorption observations in the Milky Way.

<sup>6</sup>For reference, the parameter space that Braun et al. (2009) define allows for  $T_{p,\text{max}} = 3.47 \times 10^4$  K, where  $N_{\text{HI,max}} = 3.1 \times 10^{23}$  cm $^{-2}$  and  $\sigma_{\text{min}} = 2.0$  km s $^{-1}$ . We set an upper limit of  $T_{p,\text{max}} \leq 5 \max(T_B)$  for each spectrum; raising this limit did not affect our results.

will likely require at least small changes to be optimized for a particular data set or science goal. Fortunately, the publicly available codes from these projects enable specific optimization with relative ease.

Our multi-Gaussian fitting method has three stages: (1) identifying the number of components and initial parameter guesses; (2) an initial non-least squares optimization and internal model selection test; and (3) a nearest neighbour model selection test and final fit. Each stage is described in the following sections.

### 3.2.1 Stage 1: Number of components and parameter estimates

The biggest issue in fitting a multi-Gaussian model is choosing the appropriate number of components to fit the spectrum with. As mentioned above, part of this difficulty is the correlation between different components. To overcome this issue, the number of components and initial estimates for the components can be recovered by using information from the spectrum shape.

Here, we use the ‘derivative spectroscopy’ implemented in GAUSSPY (Lindner et al. 2015), which we briefly describe below. Estimates of the derivative from the finite-difference between channels are strongly susceptible to noise. To overcome this, smoothing can include the information over several channels at each point in the spectrum. Common approaches for this smoothing include Gaussian or median smoothing (Riener et al. 2020) or total-variation regularization (Lindner et al. 2015). From this smoothing, derivatives provide information to locate and estimate components. Peaks in a spectrum are identified based on (i) a negative second derivative minimum, (ii) a zero crossing in the third derivative, and (iii) a positive fourth derivative (Lindner et al. 2015).

Previous works using GAUSSPY identified Gaussian components in two steps: first by identifying narrow components with a small smoothing length, followed by identifying wide components with a large smoothing length. Previous studies using this method identified optimal smoothing lengths using gradient descent from a training set of spectra (Lindner et al. 2015; Murray et al. 2018a; Riener et al. 2020). In this work, we assess the performance of a two-stage identification by-eye on a small set of spectra from our data sets while varying both the narrow and wide smoothing lengths. Our data contain extremely wide ( $\sigma > 30 \text{ km s}^{-1}$ ) features, particularly in M31 where the line-of-sight depth through the disc is large, and we find that the two-stage approach used in previous works does not consistently identify spectral features across this large range in line widths. Here, we explore a more ‘brute-force’ component identification method.

To account for the wide range in spectral shapes in our data, we modify the identification procedure in GAUSSPY to consider a range of smoothing lengths. Following the smoothing and component identification scheme from GAUSSPY, we identify and estimate the component parameters from small to large smoothing scales. At each smoothing scale, the component properties are estimated and subtracted from the spectrum. Then, at progressively larger smoothing scales, wider components are added from the residual in the spectrum. This procedure closely follows GAUSSPY, just with  $>2$  smoothing lengths considered.

### 3.2.2 Stage 2: Initial fit and internal model selection

The parameter estimates from the first step are crucial for the multicomponent Gaussian fit to converge to a reasonable solution.

The hope is that the initial parameter estimates start the minimization algorithm close enough to the global minimum so that it will converge quickly and not fall into a local minimum that sometimes results from the correlations between Gaussian components. However, the approximations in step one to get parameter estimates are not always robust and can lead to (i) fits converging to vastly different parameters from the starting point, or (ii) incorrect number of components or their placements. To account for this, we first fit the spectrum using the estimates from stage one, and then perform an internal model selection to ensure the resulting fit is valid. Similar to GAUSSPY (Lindner et al. 2015), we fit spectra using the Levenberg–Marquardt method implemented in the python package LMFIT.<sup>7</sup>

The model selection used here and in the following sections relies on the Bayesian information criterion (BIC) fit statistic (Schwarz 1978). The BIC is defined as a likelihood plus penalization term, where the latter increases in value with the number of free parameters to avoid overfitting

$$\text{BIC} = \ln(m)k - 2 \ln(\hat{L}), \quad (5)$$

where  $m$  is the number of data points (velocity channels),  $k$  is the number of free parameters, and  $\hat{L}$  is the likelihood function. An optimal model fit minimizes the BIC statistic, as the penalization term in equation (5) is positive.

For the multicomponent Gaussian model, we assume the data uncertainties ( $\sigma_{\text{rms}}$ ) are independent and normal, and so the log-likelihood function  $\ln(\hat{L})$  for fitting a single spectrum has the standard form of

$$\ln(\hat{L}) = -\frac{1}{2\sigma_{\text{rms}}^2} \sum_{j=1}^m [y_j - T_b(v_j)]^2 + C(\sigma_{\text{rms}}), \quad (6)$$

where  $m$  is the number of velocity channels, and  $y_j$  is the observed brightness temperature at velocity  $v_j$  and  $T_b(v_j)$  is the model from equation (4). The  $C(\sigma_{\text{rms}})$  term is a constant that depends only on the noise, which is a constant for all spectral channel and therefore does not change between models. The preferred model minimizes the BIC. We choose the BIC since it penalizes additional free parameters more strongly than other common statistics (e.g. the Akaike information criterion).

There are two parts to the internal model selection:

(i) The integral of each component must exceed  $5\sigma$  uncertainty integrated over a resolved Gaussian line. This restriction removes spurious narrow features due to noise. Following Koch et al. (2018a), we define a resolved line as having  $>5$  channels (i.e.  $2.1 \text{ km s}^{-1}$ ) across the full width half-maximum (FWHM). This restriction requires a minimal integrated intensity of  $\sqrt{\pi/(4 \ln 2)}(2.1 \text{ km s}^{-1})(5\sigma_{\text{rms}}) = 31 \text{ K km s}^{-1}$  for M31 and  $22 \text{ K km s}^{-1}$  for M33. Components below these limits are removed and the spectrum is refit with fewer components.

(ii) The above procedure is well suited to identify most spectral features, but components near the noise level may be spurious noise features or may be split into multiple components. To ensure the minimum number of components is used in the fit, we next test whether the removal of the smallest integrated intensity substantially changes the BIC of the fit. Following Kass & Raftery (1995), we consider changes of  $\Delta \text{BIC}_{N_i, N_f} > 10$  to indicate a strong preference for the initial model with the initial  $N_i$  number of components relative

<sup>7</sup>LMFIT wraps and extends the optimization algorithm from the SCIPY library (Newville et al. 2020).

to the simplified model with  $N_f$  components. Since the test is always with fewer components than the initial fit,  $N_i > N_f$ . We repeat this procedure by iteratively removing components from the model until the BIC of the fits converge to  $\Delta\text{BIC}_{N_i, N_f} = \text{BIC}_{N_i} - \text{BIC}_{N_f} < 10$ . The components are removed in order from lowest to highest integrated intensity.

One key aspect of this procedure is testing the null hypothesis of whether the spectrum is best modelled by *any* Gaussian components. We continue the process of iteratively removing components, if needed, until  $N_f = 0$ .

The updated multi-Gaussian model is fit using the fewest components that passes both selection tests. In Appendix B, we explore how well this fitting method recovers Gaussian components from synthetic multicomponent Gaussian spectra.

### 3.2.3 Stage 3: Nearest neighbour model selection and final fit

We introduce one additional step to produce the models used later in this paper. Because we fit each spectrum independently of its neighbours, the fits do not account for the correlation of nearby pixels and models of neighbouring pixels may differ. These small-scale differences are particularly an issue when different numbers of Gaussian components are found, and when strongly overlapping components converge to different solutions due to the large covariance in their parameters.

We account for these differences by comparing each spectrum to the properties of its nearest neighbours. We check and attempt to correct for differences by:

- (i) We refit the spectrum using the model with the lowest BIC from the nearest 8 pixels. We replace the spectrum's fit when the neighbouring model at pixel  $k$  decreases the BIC significantly compared to the original model at pixel  $l$ , where we define 'significant' as  $\Delta\text{BIC}_{k,l} > 10$  (Kass & Raftery 1995).
- (ii) We also refit and compare the spectrum's fit when any of the neighbouring 8 pixels has a different number of Gaussian components. In this case, we refit the spectrum using the models with the lowest and highest number of components in the neighbourhood, unless either is the spectrum's original model. We use this procedure for any difference in the number of components. We then reimpose the selection criteria from Section 3.2.2 and replace the solution until the difference in BIC of the fits converges, i.e.  $\Delta\text{BIC}_{N_i, N_f} < 10$ . Where applicable, both steps are applied.

We apply these checks in a forward and reverse direction, looping through the spatial positions of valid fits along rows of Right Ascensions, followed by column in Declination.

We find that this procedure encourages coherent spatial solutions with a consistent number of components in the model. This procedure is particularly important for spectra with weak signal relative to the noise, where the initial component guesses from Section 3.2.1 are more susceptible to noise. Though we do not make use of this product here (Section 3.3), the models after this neighbour comparison produce a far more complete and coherent model of the Milky Way H I foreground in the M31 field. Similar neighbour comparisons have been used in other multicomponent Gaussian modelling (Haud 2000), in particular, we highlight that our fitting procedure has many similarities to the GAUSSPYPLUS algorithm (Riener et al. 2020).

Finally, we note that while this approach produces more spatially coherent solutions, the spectra are still fit independently of neighbouring pixels and so the spatial continuity is not explicit

in the fitting. Correctly accounting for these differences requires the model to explicitly include its nearest neighbours and be fit simultaneously (Henshaw et al. 2016; Marchal et al. 2019). This simultaneous spatio-spectral modelling is critical when modelling blended components. However, the large data cubes we use here would require significant computational time to model in this manner. We will explore more coherent fitting methods in future work, where the spatial info becomes crucial for constraining blended components.

The multi-Gaussian models following the neighbour comparisons are the final models we use in our model comparisons (Section 4).

### 3.3 Excluding foreground and off-rotation components

M33 and, primarily, M31 overlap spectrally with Milky Way H I emission and some 'off-rotation' emission components in both galaxies (e.g. high-velocity clouds; Kam et al. 2017; Koch et al. 2018b). These features are contaminants for the model comparison we propose here since only the multi-Gaussian model can account for them.

We remove Milky Way foreground and off-rotation H I features in two ways. First, we assume the multi-Gaussian model correctly accounts for all detected components, regardless of their origin. We then remove components from the multi-Gaussian model which deviate from the centroid velocity by a velocity offset  $\Delta v$ . Due to the different galaxy inclinations, we choose different  $\Delta v$  for our targets:  $\Delta v = 50 \text{ km s}^{-1}$  for M33 (lower inclination) and  $\Delta v = 80 \text{ km s}^{-1}$  for M31 (higher inclination). We find that these choices appropriately remove off-rotation features in M33 (Koch et al. 2018b) and remove the majority of Milky Way foreground in M31.

Second, we impose an additional masking to M31 where the red-shifted edge of the disc clearly blends with the Milky Way foreground and is not distinguished by the velocity range cut applied in the first step. We use the interactively selected M31 mask described in Section 2 to remove components whose velocity centre is outside of the mask. Our results in Section 4 do not show a systematic trend near the red-shifted edge of M31's emission suggesting this separation criterion effectively excludes foreground H I. The spatial footprint of our M33 observations do not include spectra that are strongly blended with Milky Way H I, and so this second step is not applied to the M33 fits.

This separation is imposed on all results in Section 4. After this separation, we recompute the BIC statistic for the multi-Gaussian fits without the foreground/off-rotation components included in the model. These recomputed BIC values are used in all of our results presented in Section 4.

We do not, however, recompute the BIC values for the opaque H I fits. In the vast majority of the fits, the BIC statistic did not vary significantly. The lack of change is due to the original fits not accounting for off-rotation features, since the opaque model is limited to a single component. Because the revised multi-Gaussian models similarly do not include these features, the BIC values are similarly biased.

In the following section, we adopt the BIC *difference* ( $\Delta\text{BIC}$ ) as the metric for comparison between the models. The removal of foreground/off-rotation components in the multi-Gaussian model removes a source of bias when comparing the BIC with the opaque H I model. Because the opaque H I model is limited to a single component in the fit, the model is never able to account for these off-rotation components far from the bright H I emission. Thus, the removal of these foreground/off-rotation components produces  $\Delta\text{BIC}$  a fairer comparison for the opaque H I model.



## 4 WHICH MODEL IS PREFERRED FOR H I SPECTRA?

In this section, we describe the model selection test we use to compare fits between the multi-Gaussian and opaque models, synthetic tests of the model selection test, and the results of the test.

### 4.1 Model comparison using BIC

We compare the opaque H I and multi-Gaussian models by examining the difference in the BIC statistic (equation 5;  $\Delta\text{BIC}$ ) of both fits to each spectrum. Specifically, we define the BIC difference as the multi-Gaussian BIC subtracted from the opaque BIC:

$$\Delta\text{BIC} = \text{BIC}_{\text{Gauss}} - \text{BIC}_{\text{Opaque}}. \quad (7)$$

Since a minimum BIC is optimal (equation 5),  $\Delta\text{BIC} > 0$  indicates a preference for the opaque model and  $\Delta\text{BIC} < 0$  shows a preference for the multi-Gaussian model. Following Section 3.3, the BIC for the multi-Gaussian models excludes foreground and off-rotation components.

Similar to selecting the number of components for the multi-Gaussian model (Section 3.2), we choose the BIC statistic for comparing the models since it penalizes additional free parameters more strongly than other similar statistics (e.g. the Akaike Information Criteria; AIC). Specifically, additional free parameters increase the BIC value in the  $\ln(m)k$  term from equation (5), where  $m$  is the number of velocity channels and  $k$  is the number of free parameters in the model. Because each Gaussian component adds an additional three free parameters, the BIC is more likely to prefer a simpler model than the AIC, though the statistics will likely prefer a similar model in many cases.

Following Kass & Raftery (1995), we consider  $|\Delta\text{BIC}| > 10$  to be strong evidence for a particular model, with smaller values indicating a weak preference that may be spurious. Many of the comparisons exceed  $|\Delta\text{BIC}| > 10$  and so our results are not affected by the chosen threshold.

### 4.2 Producing a sample of synthetic spectra

The model comparison we outline in the previous section is a relative comparison between two specific models. While the multicomponent Gaussian model can reproduce a wide range of spectral shapes, the single component in the opaque H I model (equation 1) is far more limited. A critical check for our model selection test is then to ensure that a spectrum drawn from the opaque H I model is preferred by the model selection test. We expect this to be true since a ‘top-hat’ opaque H I model has only four free parameters, while the multicomponent Gaussian model would require at least two components, and therefore  $>6$  free parameters. However, the noise level in the data results in scatter of the BIC statistics.

To ensure our model selection test is sensitive enough to prefer true opaque H I spectra, we produce a population of 20 000 synthetic spectra using the opaque model (equation 1) with randomly drawn parameters within the parameter range used for the opaque H I observational fits that ensure a minimum peak signal-to-noise ratio (S/N) of  $>3$ . For each random draw of parameters, we evaluate the model and fit the synthetic spectrum in the following steps:

(i) We produce the true model with equation (1), parametrizing the optical depth with equation (3), evaluated over a spectral axis from  $-200$  to  $+200 \text{ km s}^{-1}$  with  $0.1 \text{ km s}^{-1}$  channels. All synthetic spectra are centred at  $0 \text{ km s}^{-1}$  since the line centre does not affect the model comparison.

(ii) We then downsample the model by averaging over  $0.42 \text{ km s}^{-1}$  to match the observations, which produces an idealized observed spectrum (Koch et al. 2018a).

(iii) We add Gaussian noise to the downsampled model with  $\sigma_{\text{rms}} = 2.8 \text{ K}$ , matching the noise in the M31 H I cube. This produces the equivalent observed spectrum to be fit.

(iv) Both models are fit to the synthetic observed spectrum using the same parameters and limits used for the observational fits.

Following the BIC model comparison from Section 4.1, we calculate the fraction of synthetic spectrum fits that strongly prefer one model ( $|\Delta\text{BIC}| > 10$ ). Since the models are drawn only from the opaque model (equation 1), the model selection test will have sufficient power if the opaque model is preferred.

Table 2 shows the break-down of model preference for the synthetic spectra. We find that 49.0 per cent correctly prefer the opaque model ( $\Delta\text{BIC} > 10$ ), and just 1.1 per cent of the fits strongly prefer the multi-Gaussian model ( $\Delta\text{BIC} < -10$ ). The remaining 49.9 per cent of the synthetic spectrum fits fall into the ambiguous region with  $-10 < \Delta\text{BIC} < 10$  where the fits are roughly equivalent (e.g. both models converge to a single optically thin Gaussian). A large fraction of the model comparisons remain in this ambiguous region because the opaque model requires a sufficient number of channels across the opaque or ‘tophat’ shape to strongly constrain the fits. The strongest model preference for the opaque model we find results from a combination of both high S/N and a large line width ( $\sigma > 15 \text{ km s}^{-1}$ ). For spectra with these wide line widths, the multi-Gaussian model typically infers  $>1$  components. If we consider only the synthetic spectra fit with  $>1$  components in the multi-Gaussian model, we find that 80.4 per cent of the spectra prefer the opaque model ( $\Delta\text{BIC} > 10$ ), 18.0 per cent are ambiguous ( $-10 < \Delta\text{BIC} < 10$ ) and 1.5 per cent prefer the multi-Gaussian model ( $\Delta\text{BIC} < -10$ ). This demonstrates that the preference becomes much stronger for the opaque model fit when multiple Gaussian components are inferred.

These results confirm that our fitting approach and model selection test has sufficient power to distinguish between the opaque and multi-Gaussian models in our H I observations, and that that power diminishes as the S/N in the spectrum decreases, as expected. In the following sections, we use these results from the synthetic spectra to compare and contrast with the model preference from the observed H I spectra.

### 4.3 A strong preference for multi-Gaussian models

Using the comparison framework from Section 4.1, we examine  $\Delta\text{BIC}$  for the population of H I spectra in M31 and M33. Our comparisons in this section demonstrate a strong preference for a multi-Gaussian model.

We first visually examine the fits for a few spectra, chosen specifically to highlight a range in the inferred peak optical depth from the opaque model. Fig. 4 shows four example spectra with fits to both models (chosen from a subset of the spectra in Fig. 2). Each panel in the figure shows the key fit results, including the number of Gaussians ( $N_{\text{Gauss}}$ ), inferred peak optical depth ( $\tau_p$ ), and the BIC values are stated for each fit. Following visual inspection, the spectra in panels (a) and (c) clearly demonstrate multiple peaks and strongly prefer the multi-Gaussian model based on their smaller BIC values ( $\Delta\text{BIC} < 0$ ). The spectra in panels (b) and (d) both have a single bright ‘pedestal,’ which is qualitatively the flat-top shape expected for the opaque H I model over the velocity range where  $\tau > 1$  (Braun et al. 2009). Only spectrum (b) prefers the opaque model ( $\Delta\text{BIC} >$

**Table 2.** Percentage of spectra for different ranges in  $\Delta\text{BIC}$  in the M31, M33, and synthetic spectra fit distributions. We define three ranges: (i) a strong preference for the multi-Gaussian model ( $\Delta\text{BIC} < -10$ ); (ii) spectra without a strong preference for either model ( $-10 < \Delta\text{BIC} < 10$ ); and (iii) a strong preference for the opaque model ( $\Delta\text{BIC} > 10$ ). We compute the percentage in the ranges for the three tests including all fits and only those with  $\tau_p > 1$  (Fig. 6), and the BIC computed only where  $\tau(v) > 0.5$  (Fig. 7). Our synthetic spectrum fit distribution correctly shows a preference for the opaque model, from which the spectra are drawn from. However, the vast majority of observational fits in M31 and M33 prefer the multi-Gaussian model for all three comparisons.

		Multi-Gaussian preferred per cent with $\Delta\text{BIC} < -10$	No strong preference per cent with $-10 < \Delta\text{BIC} < 10$	Opaque preferred per cent with $\Delta\text{BIC} > 10$
M31	All (Fig. 6; top)	82.0	17.5	0.5
	$\tau_p > 1$ (Fig. 6; bottom)	88.7	10.5	0.8
	$\tau(v) > 0.5$ (Fig. 7)	83.7	15.4	0.9
M33	All (Fig. 6; top)	80.2	19.6	0.2
	$\tau_p > 1$ (Fig. 6; bottom)	83.2	16.0	0.7
	$\tau(v) > 0.5$ (Fig. 7)	74.4	25.0	0.6
Synth.	All (Fig. 6; top)	1.1	49.9	49.0
	$\tau_p > 1$ (Fig. 6; bottom)	1.4	35.0	63.6
	$\tau(v) > 0.5$ (Fig. 7)	0.6	78.5	20.9

0), however. In panel (d) the multi-Gaussian model fit has a much smaller BIC value and is therefore preferred.

To test this apparent preference for the multi-Gaussian model, we create  $\Delta\text{BIC}$  maps of both galaxies. Fig. 5 shows  $\Delta\text{BIC}$  in M31 and M33 where valid fits are found (Section 3). The colour is centred such that  $\Delta\text{BIC} = 0$  is shown in grey. Throughout the maps, we find a strong preference for the multi-Gaussian model ( $\Delta\text{BIC} < 0$ ). Most locations where the opaque model is preferred ( $\Delta\text{BIC} > 0$ ) correspond to regions with faint H I emission (Fig. 1). This result is in contrast with the findings by Braun et al. (2009) and Braun (2012), where opaque H I is found in regions of bright H I (e.g. Braun 1997). While we find that the largest opacity corrections correspond to regions with bright H I, as shown by the black contour in Fig. 5, these are also the spectra that tend to have the strongest preference for the multi-Gaussian model.

We see a general trend towards  $\Delta\text{BIC} \approx 0$  near the edge of H I detections in both maps, where the noise increases (e.g. in M31) and the H I emission is fainter. At lower S/N, the multi-Gaussian model tends to have a single component, and therefore is similar to the opaque model in the optically thin limit. In these regions, there is a small tendency for  $\Delta\text{BIC} < 0$  since a single Gaussian component has three free parameters while the optically thin limit of the opaque model includes a fourth unconstrained parameter,  $T_s$ , and so the former model is preferred.

Table 2 shows the fraction of spectra prefer each model, split into categories of  $\Delta\text{BIC} < -10$ ,  $-10 < \Delta\text{BIC} < 10$ , and  $\Delta\text{BIC} > 10$ , as defined in Section 4.1. As shown in the maps of  $\Delta\text{BIC}$  in Fig. 5, we find that most (82.0 per cent and 80.2 per cent) of the spectra prefer the multi-Gaussian model, while just 0.5 per cent and 0.2 per cent prefer the opaque model in M31 and M33, respectively. These fractions differ significantly from the synthetic sample of opaque fits, where 49.0 per cent prefer the opaque model.

To visualize which spectra most strongly prefer the multi-Gaussian model, we compare the model preference  $\Delta\text{BIC}$  to how much the opaque H I model deviates from a single Gaussian (i.e. the optically thin limit). This deviation of the opaque H I model from a Gaussian is difficult to parametrize in terms of a single fit parameter, or, for example, the peak optical depth. The reason for this is because the three parameters that determine the profile shape –  $T_p$ ,  $T_s$ , and  $\sigma$  – are highly correlated. Instead, we compare  $\Delta\text{BIC}$  to the apparent

‘dark’ H I intensity,  $I_{\text{dark}}$ , which is the integrated difference of the opaque model fit (equation 1) if the medium were optically thin and the model itself:

$$I_{\text{dark}} = \int \left[ T_p \exp\left(-\frac{(v - v_0)^2}{2\sigma_p^2}\right) - T_s (1 - \exp[-\tau(v)]) \right] dv, \quad (8)$$

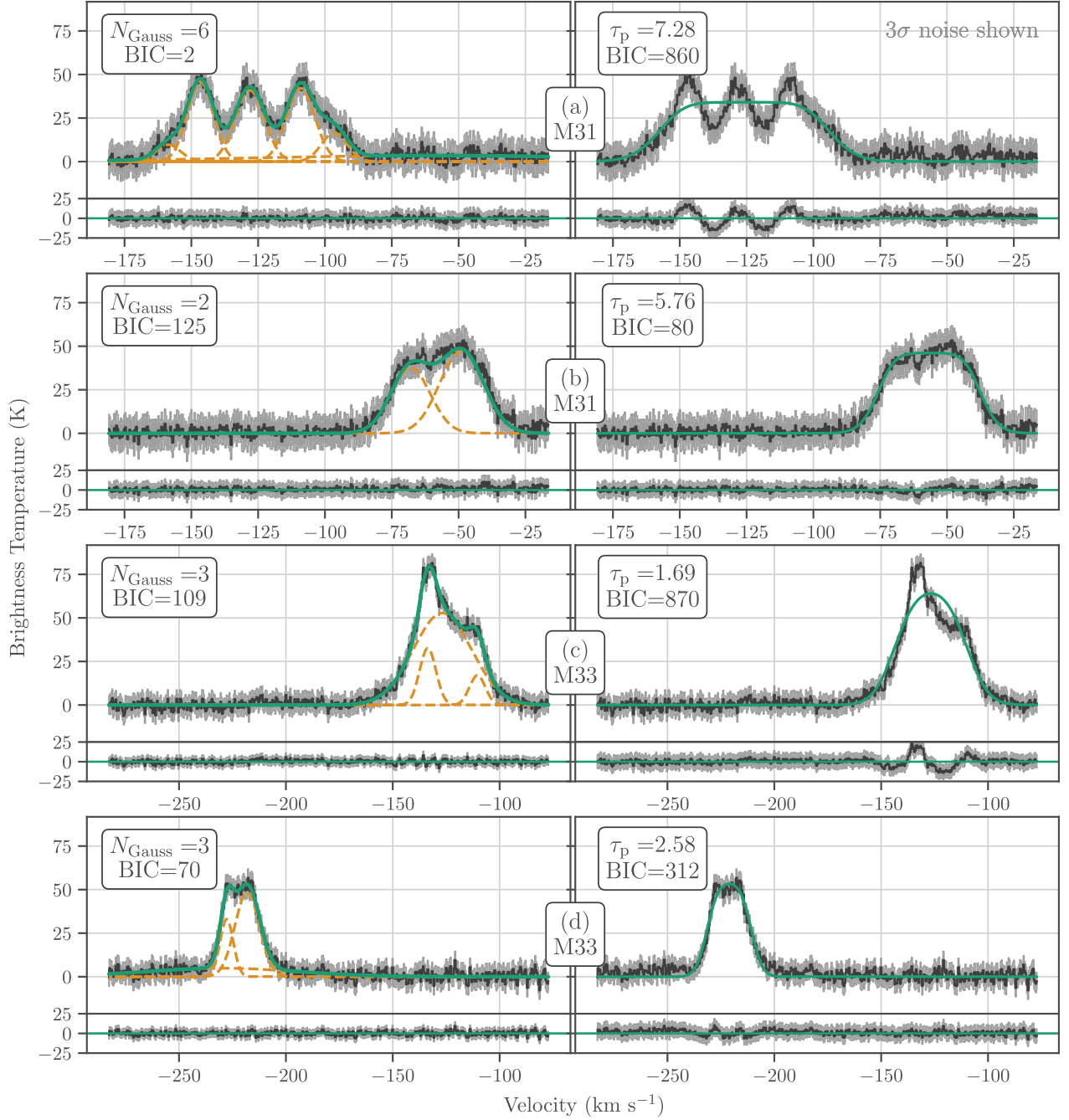
where  $\tau(v)$  is given by equation (3). When the model fit infers the spectrum is optically thin,  $I_{\text{dark}}$  converges to 0. We calculate the  $1\sigma$  uncertainty on  $I_{\text{dark}}$  using the `UNCERTAINTIES` package<sup>8</sup> which propagates the parameter uncertainties from the opaque model fit (equation 3) in equation (8). We note that these uncertainties, following the fit parameter uncertainties, assume they are normally distributed. This assumption will break-down when  $I_{\text{dark}}$  approaches the optically thin limit because  $I_{\text{dark}}$  cannot be negative.

We stress that interpreting  $I_{\text{dark}}$  as a physical quantity depends on the opaque model fitting a spectrum well. As we show above, <1 per cent of the observed spectra prefer the opaque model, and so the opaque fit parameters should be considered carefully before assigning a physical meaning to what may be a poor fit. Because of this, we consider  $I_{\text{dark}}$  as a useful *metric* to compare with, rather than a physical quantity.

Fig. 6 shows the  $\Delta\text{BIC}$  statistic distributions for M31, M33, and the synthetic spectrum fits plotted against  $I_{\text{dark}}$ . The top row shows the distribution of all fitted spectra. Consistent with the maps in Fig. 5, the H I spectra in M31 and M33 with the strongest preference for the multi-Gaussian tend to have a larger  $I_{\text{dark}}$  inferred from the opaque model. This trend is the opposite of the trend for our fits to the synthetic spectra of the opaque model (Section 4.2; middle panel in Fig. 6), where the preference for the opaque model in several fits increases at large  $I_{\text{dark}}$ .

A common trend of all three samples (M31, M33, and the synthetic spectra) in Fig. 6 is the broad distribution of  $\Delta\text{BIC}$  values when  $I_{\text{dark}}$  approaches zero. This broad range shows where the opaque model approaches the optically thin solution, where some spectra are better modelled with multiple components, or where the S/N is low and both models approach a similar fit, as we see at the edges of the maps in Fig. 5.

<sup>8</sup>v3.1.5; <https://pythonhosted.org/uncertainties/>.

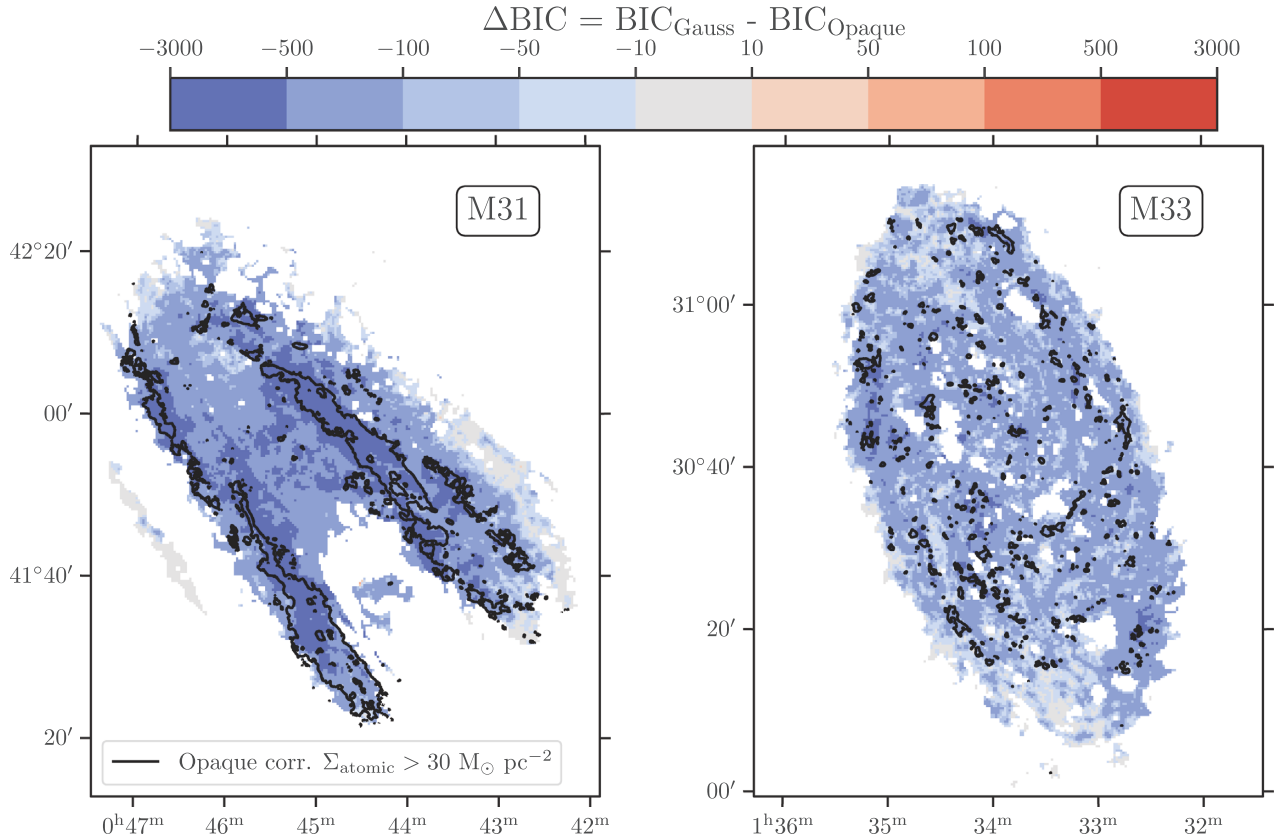


**Figure 4.** Four of the H I spectra shown in Fig. 2 with their multi-Gaussian (left) and opaque (right) model fits. The fit residuals are shown in the panel below each spectrum and fit, with the  $3\sigma$  rms noise shown in grey. The number of Gaussians and the peak opacity ( $\tau_p$ ) are stated for each model, respectively, as well as the BIC statistic for each fit. In three of the examples (a, c, and d), the BIC statistic is smaller for the multi-Gaussian model and is therefore the preferred model. The opaque model infers  $\tau_p > 1$  in all three cases, but the fit quality is noticeably worse, especially for (a) and (c) where there are clearly multiple components. Panel (b) shows a case where the opaque and multi-Gaussian models provide a similar fit to the data (i.e. the fit residuals are similar), but the opaque model has fewer free parameters and the smaller BIC values indicates it is preferred. We assess the validity of this visual inspection by comparing the BIC statistics for the entire population of fitted spectra.

We next examine whether the preference for a multi-Gaussian model persists if we consider only the spectra where opaque H I emission is inferred from the opaque model, or equivalently where the peak opacity is  $\tau_p > 1$ . After applying this criterion, we find that the same general trend persists: most spectra prefer the multi-Gaussian model and that preference becomes stronger for the largest inferred  $I_{\text{dark}}$ . We find that fraction of spectra that prefer the multi-Gaussian

model increase to 88.7 per cent for M31 and 83.2 per cent for M33. This comparison is shown in the bottom row of Fig. 6. Because only a subset of the  $\Delta\text{BIC}$  are shown, the percentiles defining the contours in the plots are changed, leading to some difference in the observed distributions.

This trend of a stronger preference for the multi-Gaussian model where opaque H I is inferred from the opaque model (equation 1)



**Figure 5.**  $\Delta\text{BIC}$  values comparing the multi-Gaussian and opaque models. Negative values (blue) indicate a preference for the multi-Gaussian model, while positive values (red) show where the optically thick model is preferred. The grey regions in the map indicate where neither model has a strong model preference ( $|\Delta\text{BIC}| < 10$ ). The black contours show where  $\Sigma_{\text{atomic}} > 30 \text{ M}_{\odot} \text{ pc}^{-2}$  with the opacity correction from equation (1). This surface density exceeds all values assuming optically thin emission. Multi-Gaussian models are preferred in the majority of spectra. Further, as the contour regions indicate, regions with the strongest preference for multi-Gaussian (largest negative  $\Delta\text{BIC}$ ) correlate with the largest inferred opacity corrections from the opaque model.

is opposite of the trend with the synthetic spectra. In the synthetic sample, we find that the fraction of spectra that prefer the opaque model increases from 49.0 per cent to 63.6 per cent when considering only spectra where opaque H I is inferred from the opaque model (Table 2). This follows our expectation for the opaque model since the spin temperature  $T_s$  is only constrained when the line shape approaches a ‘top-hat’ shape (Fig. 3).

Finally, we include one further test that explicitly accounts for the single-component nature of the opaque model (equation 1). The spectrum in panel (d) in Fig. 4 demonstrates the limits of the opaque model well. The multi-Gaussian model correctly includes a wide component that accounts for faint ‘tails’ in the emission. However, the opaque model cannot account such features and is therefore more likely to have a larger BIC, even if the model has fewer parameters.

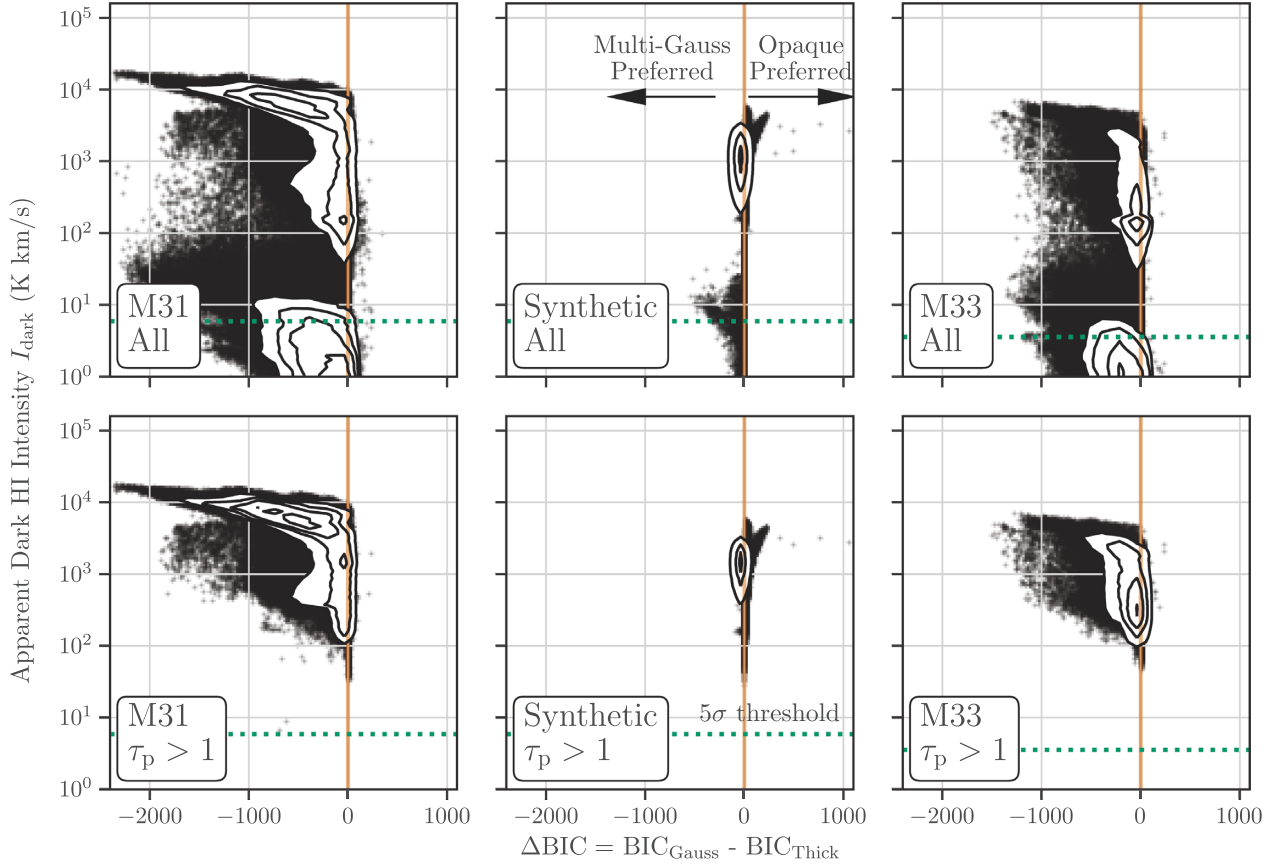
To address this limitation, we recalculate the BIC statistic for both models limited to the channels where  $\tau(\nu) > 0.5$ . In terms of the opaque model, this limits the model selection to comparing *only* the spectral range inferred to have opaque H I emission (see the example opaque model in Fig. 3) or where the profile will have a ‘top-hat’ shape. This limit is chosen based on our synthetic spectrum fits where the opaque model measurably deviates from a Gaussian line shape given the noise level and spectral resolution in our observations. For the multi-Gaussian model, we consider only the components where  $>75$  per cent of their intensity is within range of  $\tau(\nu) > 0.5$ .

Fig. 7 shows the resulting distribution of  $\Delta\text{BIC}$  limited to  $\tau(\nu) > 0.5$  and the model preference fractions are given in Table 2. As expected, the synthetic spectrum fit distribution shows a preference for the opaque model, with 20.9 per cent spectra with  $\Delta\text{BIC} > 10$  and just 0.6 per cent with  $\Delta\text{BIC} < -10$  preferring the multi-Gaussian model. These fractions are nearly the opposite of what we find for the observed spectra, with 0.9 per cent and 0.6 per cent of spectra with  $\Delta\text{BIC} > 10$ , while 83.7 per cent and 74.4 per cent have  $\Delta\text{BIC} < -10$  for M31 and M33, respectively.

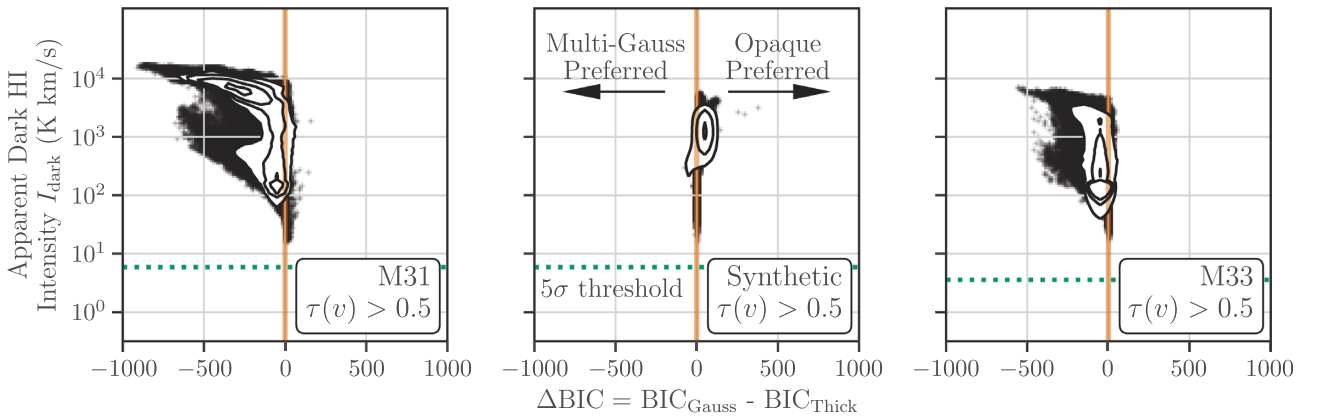
In all three distributions, a larger fraction of spectra do not show a strong preference for one model or another ( $-10 < \Delta\text{BIC} < 10$ ), particularly for the synthetic spectrum fit distribution which has 78.5 per cent in this ambiguous region. This trend occurs because this test limits the number of spectral channels that are used for the model comparison, in effect, limiting the amount of information available and increasing the scatter in  $\Delta\text{BIC}$  values. Despite the fact that the  $\tau(\nu) > 0.5$  restriction makes  $\Delta\text{BIC}$  a noisier metric, we still find a strong preference for multi-Gaussian models in both M31 and M33. This result demonstrates that the preference for multi-Gaussian models does not result from the opaque model being limited to a single component (equation 1).

In summary, we identify a strong preference for the multi-Gaussian model in observed H I spectra when compared to a single opaque H I component model. This result indicates that most H I are best described as emission from multiple atomic ISM components along the line of sight.

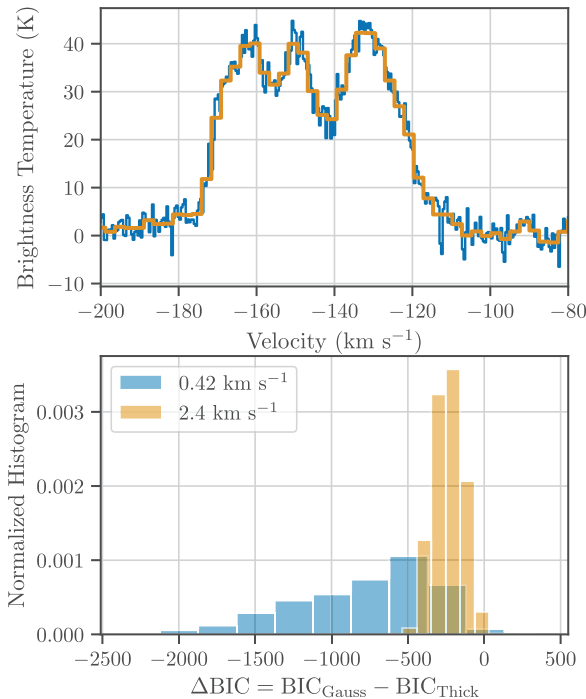




**Figure 6.** Apparent dark H I integrated intensity from the opaque model (equation 8; i.e. optically thin equivalent line subtracted by the optical depth-corrected profile) plotted against  $\Delta\text{BIC}$ .  $\Delta\text{BIC} < 0$  indicates a preference for the multi-Gaussian model, while  $\Delta\text{BIC} > 0$  are fits with the opaque model preferred. The panels show the plots for M31 (left), synthetic spectra generated from equation (1) (centre), and M33 (right). The top row shows the results for all spectra, while the second row shows fits where the peak optical depth from the opaque model fit is  $\tau_p > 1$ . The contours in each panel show the density of fits contained at the  $1-4\sigma_{\text{rms}}$  levels, while individual data points show individual fits that fall outside of the  $4\sigma_{\text{rms}}$  contour. The narrow vertical shaded region indicates where  $-10 < \Delta\text{BIC} < 10$  where there is no strong preference for either model. The horizontal dotted line is the  $5\sigma_{\text{rms}}$  sensitivity in integrated intensity for a single  $0.42 \text{ km s}^{-1}$  spectral channel. The observed fit results strongly prefer a multi-Gaussian model ( $\Delta\text{BIC} < 0$ ), while the synthetic sample correctly prefers the opaque model used to generate them ( $\Delta\text{BIC} > 0$ ). These differences become even more apparent when *only* opaque model fits are shown in the second row, demonstrating the opposite trend expected for fits to the opaque model. Despite having many more free parameters, the multi-Gaussian model is consistently preferred.



**Figure 7.**  $\Delta\text{BIC}$  plotted against the apparent dark H I integrated intensity from the opaque model, similar to Fig. 6. Here, we show only the fit statistics calculated where  $\tau(v) > 0.5$  from equation (3). This comparison controls for the single-component limitations of the opaque model, which cannot account for multiple peaks in a spectrum. By computing the BIC statistics only where  $\tau(v) > 0.5$ , this comparison solely compares the models where the ‘flat-top’ is measurable in the spectra. However, these results are similar to the  $\Delta\text{BIC}$  over the whole spectra (Fig. 6). The synthetic spectra, drawn from equation (1) with observational noise added, confirm that our observations are sensitive to opaque line profiles and the fitting procedures correctly find  $\Delta\text{BIC} > 0$  on average. The observed spectra in M31 (left) and M33 (right), however, have  $\Delta\text{BIC} < 0$ , indicating a continued strong preference for the multi-Gaussian model.



**Figure 8.** Effect of spectral resolution shown with an example H I spectral shape (top) and distributions of  $\Delta\text{BIC}$  of a sub-region of M31 (bottom). We demonstrate here that a coarser spectral resolution removes spectral shape information from the original spectra, making some components appear more blended. In this case, we plot the spectrum with  $2.4 \text{ km s}^{-1}$  spectral channels, similar to the resolution of the M31 H I data from Braun et al. (2009). The bottom panel shows the model selection comparison with  $\Delta\text{BIC}$  of a  $200 \times 200$  pixel region in our M31 H I data. Though the multi-Gaussian model is preferred in both cases, we find that fits to the coarser resolution data increase most  $\Delta\text{BIC}$  values, indicating that the model preference is *weaker* for spectra with coarser spectral resolution.

#### 4.4 Effect of spectral resolution on model preference

A key difference in the analysis we perform here relative to that of Braun et al. (2009) is the improved spectral resolution of the H I observations. Here, we test how changing the spectral resolution affects the model preference by selecting a  $1'1 \times 1'1$  ( $200 \times 200$  pixel) region in M31 centred at  $0^{\text{h}}44^{\text{m}}34^{\text{s}} + 41^{\text{d}}55^{\text{m}}13^{\text{s}}$ . This small region spans across where both large and small opaque regions are inferred from the opaque model fit (see Fig. 5). We spectrally smooth and then downsample the H I spectra to have spectral channels of  $2.4 \text{ km s}^{-1}$ , similar to the  $2.3 \text{ km s}^{-1}$  spectral channels in the M31 H I observations from Braun et al. (2009). The top panel in Fig. 8 shows an example spectrum at the original and coarser spectral resolution.

We then applied the same fitting approach to the spectrally downsampled H I spectra, as described in Section 3, and calculate the  $\Delta\text{BIC}$  between the two model fits. In the bottom panel of Fig. 8, we plot histograms of  $\Delta\text{BIC}$  for fits at the original  $0.42 \text{ km s}^{-1}$  and coarser  $2.4 \text{ km s}^{-1}$  spectral resolutions. We find that, while both distributions indicate a preference for the multi-Gaussian model ( $\Delta\text{BIC} < -10$ ), the preference is weaker for fits at coarser spectral resolution. This result demonstrates that the model selection test becomes weaker as information is lost by downsampling the spectra. For example, the example spectrum in Fig. 8 has two blended peaks that become more difficult to distinguish at coarser spectral resolution.

We additionally highlight that, because we spectrally smooth the data when downsampling, the S/N is also improved relative to the original data. If the noise levels were higher – comparable to the H I data used by Braun et al. (2009) and Braun (2012) (Table 1) – the  $\Delta\text{BIC}$  would further increase and trend towards  $\Delta\text{BIC} \sim 0$ . Including additional noise will lead to poorer discrimination between the models.

The test we present here demonstrates a key advantage for finer spectral resolution H I observations when the S/N is sufficient to constrain the spectral shape. When the spectral resolution is too coarse, leading to highly blended H I components, it becomes more difficult to identify a model preference. This example highlights how the trade-off between using coarser spectral resolution, to maximize S/N, versus finer spectral resolution, to best recover the spectral shape, affects our ability to model the H I spectra.

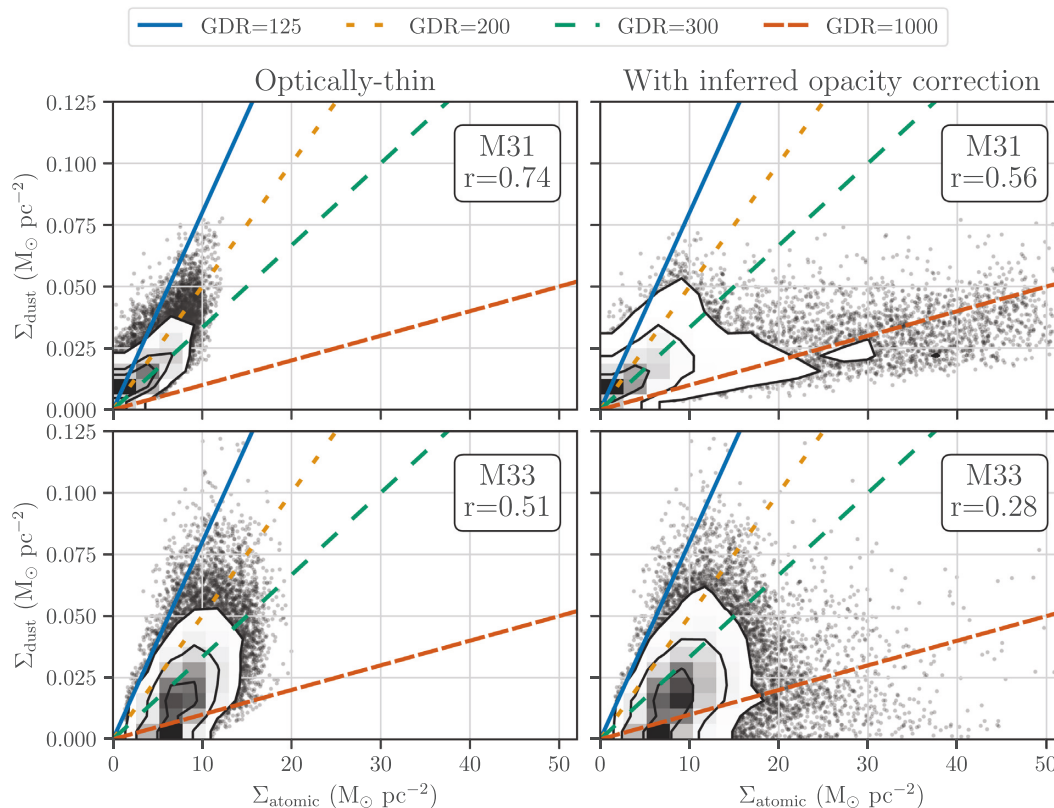
## 5 COMPARISON WITH DUST TRACERS

Dust emission and dust extinction provide independent measurements of ISM column density that can be compared with the inferred H I opaque corrections from the opaque model (equation 1). In this section, we compare the atomic gas surface density and inferred opacity corrections to the dust surface density from modelling the IR SED in M31 and M33 from Utomo et al. (2019), and the mean visual extinction  $A_V$  in M31 from Dalcanton et al. (2015) over the region observed by PHAT. We additionally correct the surface densities for disc inclination in both galaxies (see Section 2.1). Finally, to emphasize the atomic gas to dust comparison, we exclude regions where the molecular gas dominates the neutral ISM. Specifically, we remove regions with  $\Sigma_{\text{mol}} > 1.5 \text{ M}_{\odot} \text{ pc}^{-2}$  based on CO emission using CO(1-0) in M31 (Nieten et al. 2006) and CO(2-1) in M33 (Gratier et al. 2012; Druard et al. 2014), where we correct for the galaxy inclination angle and use a conversion factor of  $\alpha_{\text{CO}10} = 4.3 \text{ M}_{\odot} \text{ pc}^{-2} / \text{K km s}^{-1}$  with a line ratio of  $R_{21} = 0.7$  for the CO(2-1) line. The surface density needed for  $\text{H}_2$  formation is  $\sim 10 \text{ M}_{\odot} \text{ pc}^{-2}$ , and increases with decreasing metallicity (e.g. Krumholz, McKee & Tumlinson 2009; Schruba, Bialy & Sternberg 2018), so our chosen threshold should keep only regions where the atomic gas is the primary neutral ISM component.

By considering only regions where the atomic gas dominates the neutral ISM, we expect the relation between dust and atomic gas surface density to approximately be  $\Sigma_{\text{dust}} \approx \text{GDR}^{-1} \Sigma_{\text{atomic}}$ , where GDR is the gas-to-dust ratio. If a significant part of the H I along a line of sight is opaque, following equation (1), we expect to find two features in the relation between  $\Sigma_{\text{atomic}}$  and  $\Sigma_{\text{dust}}$ : (i) additional scatter when using the optically thin assumption for H I, with a population of lines of sight with a low apparent GDR where  $\Sigma_{\text{atomic}}$  is underestimated;<sup>9</sup> and (ii) reduced scatter after applying the inferred H I opacity correction as this correction accounts for a missing component of  $\Sigma_{\text{atomic}}$  that is traced by the dust.

Fig. 9 shows the atomic gas versus dust surface density in M31 and M33, with and without the H I opacity correction. We smooth the H I surface density maps to match the 167 pc resolution of the dust surface density from Utomo et al. (2019), which corresponds to  $46'3$  for M31 and  $41$  arcsec and for M33. Using the optically thin H I assumption, we find a strong correlation between  $\Sigma_{\text{atomic}}$  and

<sup>9</sup>Despite removing lines of sight where molecular gas dominates, there remain many spectra with a large inferred opacity based on the opaque model fit (equation 1). Thus, we would still expect to find low apparent GDR lines of sight if the inferred H I opacity is accurate.



**Figure 9.** Atomic gas surface density versus dust surface density without (left column) and with (right column) H I opacity corrections from equation (1). All values are corrected for the disc inclination along the line of sight and are convolved to a common physical resolution of 167 pc, matching the dust surface density maps from Utomo et al. (2019). In each panel, regions with  $\Sigma_{\text{mol}} > 1.5 \text{ M}_{\odot} \text{ pc}^{-2}$  (i.e. non-negligible molecular gas mass) are removed, so that to first order the atomic and dust surface densities are related only by the gas-to-dust ratio (GDR; e.g. Sandstrom et al. 2013), and so we report the linear correlation coefficient  $r$  in each panel. A range of GDRs are shown with overlaid lines. The inferred opaque H I corrections from the opaque model (equation 1) significantly increase the scatter in this relation for both galaxies, and further require a factor of  $\sim 10$  range in the GDR to explain. Consistent with our spectral model comparisons, these results suggest that the opaque model overestimates or has highly uncertain opaque H I corrections.

$\Sigma_{\text{dust}}$  consistent with  $\text{GDR} \approx 100\text{--}300$ . This range in GDR broadly agrees with previous findings in M31 and M33 (e.g. Nieten et al. 2006; Leroy et al. 2011; Gratier et al. 2017; Williams, Gear & Smith 2019). We note that most studies also include molecular gas as traced by CO and so do not remove molecular gas-dominated lines of sight.

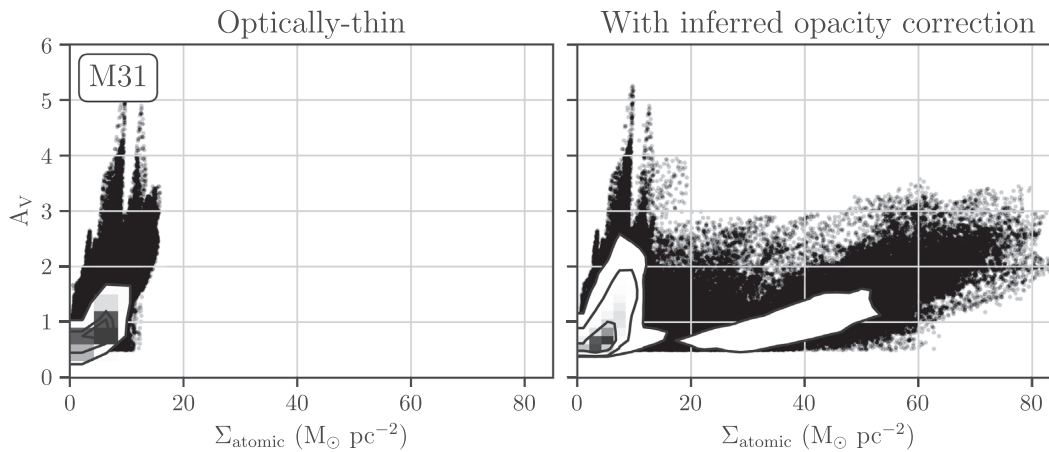
Including the inferred H I opacity corrections from the opaque model (equation 1), the correlation between  $\Sigma_{\text{dust}}$  and  $\Sigma_{\text{atomic}}$  is reduced and the scatter in the relation significantly increases. In both galaxies, lines of sight with large H I opacity corrections require a much larger GDR to explain the observed relations. Particularly in M31, regions with large H I opacity corrections require  $\text{GDR} \approx 1000$ . Higher values of the GDR are expected in low-metallicity galaxies, such as the SMC (Leroy et al. 2011; Roman-Duval et al. 2014), but are otherwise not found in higher metallicity galaxies in the Local Group (Leroy et al. 2011) or other nearby galaxies (Sandstrom et al. 2013).

We can further compare the H I surface density with the dust extinction  $A_V$  in M31 (Dalcanton et al. 2015), which provides another useful measure of the dust with independent systematics from the IR SED fitting (e.g. Whitworth et al. 2019). The dust extinction provides two key advantages relative to the dust surface density from fitting the dust emission SED. First,  $A_V$  is not dependent on the dust temperature. This overcomes ambiguity when comparing with the dust emission SED, where the dust surface density and temperature are correlated. This correlation could explain the large range in DGR

since the SED fit is a mixture of warm and cold dust along the line of sight and the warm component, which radiates more strongly per unit mass, biases the inferred dust temperature (Sandstrom et al. 2013). Second, the IR bands, where the dust SED is constrained, have a coarser spatial resolution compared to our H I data by a factor of  $\sim 2$ . The M31  $A_V$  map can be compared at the same spatial resolution as the H I maps and thereby tests whether compact high surface density regions are smoothed out at the coarser resolution of the dust emission maps. We note, however, that we lack matched spatial resolution CO data covering the entire H I mosaic (Fig. 1). Because of this, we do not exclude molecular dominated regions in this comparison.

Fig. 10 shows the relation between  $\Sigma_{\text{atomic}}$  and  $A_V$  with and without the inferred H I opacity correction from the opaque model (equation 1). Similar to Fig. 9, the inferred H I opacity correction increases the scatter in the relation of  $\Sigma_{\text{atomic}}$  and  $A_V$ . In particular, lines of sight with large opaque H I corrections require a population where  $A_V$  remains constant but  $\Sigma_{\text{atomic}}$  varies by a factor of 4. This difference appears hard to explain, since the inclusion of other ISM components (e.g. ionized or molecular gas) that are not accounted for creates a larger discrepancy.

One potential explanation for this inferred high  $\Sigma_{\text{atomic}}$  at low  $A_V$  is if the dust emission is primarily from compact and unresolved sources. However, these features would be hard to explain given the dust surface density remains low in Fig. 9. This is further disfavoured based on a recent comparison of dust emission and extinction on



**Figure 10.** Atomic gas surface density versus mean visual extinction  $A_V$  from Dalcanton et al. (2015) without (left) and with (right) H I opacity corrections from equation (1). The data are convolved to a common resolution of  $\sim 60$  pc, matching the original 18 arcsec resolution of our H I data for M31 (Section 2). Similar to the dust surface density from IR emission (Fig. 9), we find a roughly linear relation assuming optically thin H I emission and a large deviation when the inferred H I opacity corrections from the opaque model (equation 1) are included. To be consistent with the inferred opacity corrections, we expect to find compact regions with large  $A_V$ . The lack of such features is consistent with the opaque model overestimating opaque H I corrections and previous work (Whitworth et al. 2019).

$\sim 30$  pc scales by Whitworth et al. (2019). They find that dust emission from compact sources would require far wider distributions in  $A_V$  than those found in Dalcanton et al. (2015). The nature of these ‘compact sources’ and their effect on the relation between atomic gas and dust would need to explain these apparent contradictions in the observed relations.

The comparison of atomic gas surface densities with and without H I opacity correction with both dust emission and dust extinction maps brings us to the conclusion that the opaque H I corrections inferred from the opaque model fits (equation 8) are likely overestimated or highly uncertain. Because of this, the scatter is significantly increased in the relation between  $\Sigma_{\text{atomic}}$  and  $\Sigma_{\text{dust}}$  when the inferred opaque H I corrections are included, which is the opposite trend expected for significant opacity corrections. These results qualitatively agree with the lack of evidence for the opaque H I model fits based on the spectral fitting and model selection in Section 4.

## 6 DISCUSSION

Our results suggest a vastly different picture than the ubiquity of opaque H I in M31, M33, and the LMC proposed by Braun et al. (2009) and Braun (2012). Here, we discuss reasons for this discrepancy (Section 6.1), explore differences in the opaque H I mass correction factors that we determine (Section 6.2), and compare the opaque model fits with  $T_s$  values derived from H I absorption studies (Section 6.3). We also state shortcomings of the models we fit (Section 6.4) and identify additional approaches to find opaque H I using emission spectra (Section 6.5).

### 6.1 Discrepancies with the Braun et al. interpretation

Our results are clearly discrepant from those in Braun et al. (2009) and Braun (2012). These discrepancies result from (i) improved spectral resolution and a higher sensitivity on smaller spatial scales, and (ii) improvements in the modelling methodology.

Table 1 summarizes the differences in our observations versus those from Braun et al. (2009) and Braun (2012). Our observations have a similar sensitivity at moderately higher spatial resolution

but with a finer spectral resolution. In Section 4.4, we demonstrate how coarser spectral resolution ( $2.4 \text{ km s}^{-1}$ ) leads to a larger  $\Delta\text{BIC}$  when comparing the fits of the opaque and multi-Gaussian fits, though we find that the latter is still preferred in the majority of the spectra ( $\Delta\text{BIC} < 0$ ). This comparison shows the importance of spectral resolution for resolving blended spectral components that may otherwise appear as a ‘top-hat’ like shape.

Another source for discrepancy in our results stems from the improvements in methodology used for fitting. Braun et al. (2009) do not explicitly model multiple components, and additional residuals after fitting equation (1) are assumed to be optically thin emission. Further, the model selection criteria they use to distinguish between multiple components and a flat-top are more coarse than our comparison. They define spectra with reduced chi-square values of  $\chi_r^2 < 25$ , based on visual inspection, to be a valid opaque H I fit. However, given there are faint features not modelled in their fits (see their fig. 14), it is not clear whether this threshold is sufficient to distinguish between multiple components and faint emission not included in the fit. Braun (2012) rejects 4 per cent of their fits in M31 and  $< 1$  per cent in M33 based on this threshold.

For comparison, if we were to apply the same restriction of  $\chi_r^2 < 25$  to our opaque model fits then we would find that *none* of the opaque model fits to be excluded. This difference is likely the result of different systematics in the observations, combined with the effect of spectral resolution as mentioned above. This demonstrates that an absolute fit statistic value, like  $\chi_r^2$ , produces different effects when applied to different data sets. This makes it difficult to directly compare the fit quality.

### 6.2 Correction factor for optically thick H I column density

The key outcome from previous studies using the opaque model is the ability to account for ‘dark’ H I over a large spatial area, overcoming limitations of other methods that require H I absorption against background sources. With a constraint on the opacity, the cold opaque atomic gas content can be estimated, yielding a mass correction factor for the atomic ISM relative to the optically thin mass. In Braun et al. (2009) and Braun (2012), they infer that opaque H I accounts for an additional  $\sim 35$  per cent of atomic ISM mass



**Table 3.** Dark atomic ISM mass in M31 and M33 estimated with (i) no selection criteria, (ii) only spectra with peak optical depths of  $\tau_p < 5$  based on the opaque model, and (iii) spectra where the opaque model is preferred ( $\Delta\text{BIC} > 0$ ; equation (1)). We highlight that the dark atomic mass converted from  $I_{\text{dark}}$  (equation 8) must be  $> 0$  and our assumption of normal-distributed uncertainties breaks down when the uncertainty is large, which we indicate in the table with a  $^\dagger$ . We note that all of our fits satisfy the  $\chi_r^2 < 25$  criterion from Braun et al. (2009), and therefore their selection criterion is equivalent to using all of the opaque H I fits. The table also shows the percent of spectra after the selection criterion is applied, and the percent mass difference compared to the optically thin mass. The optically thin atomic ISM mass is calculated from the integrated intensity maps excluding pixels without valid spectral fits. We find large differences in the mass correction factors that depend strongly on the sample selection, and similarly the goodness of fit. This strong dependence on sample selection suggests that the dark H I mass from the opaque model (equation 1) is highly uncertain.

	Sample selection	M31	M33
Opt. thin atomic mass* ( $M_\odot$ )	–	$1.7 \times 10^9$	$1.8 \times 10^9$
Dark atomic mass* ( $M_\odot$ )	All	$(2.0 \pm 0.8) \times 10^9$	$(3.7 \pm 2.7) \times 10^8$
Per cent of spectra		100 per cent	100 per cent
Per cent mass difference		$118 \pm 47$ per cent	$21 \pm 15$ per cent
Dark atomic mass* ( $M_\odot$ )	$\tau_p < 5$	$(1.8 \pm 2.3) \times 10^8$ $^\dagger$	$(1.8 \pm 1.9) \times 10^8$ $^\dagger$
Per cent of spectra		85.7 per cent	97.3 per cent
Per cent mass difference		$11 \pm 14$ per cent $^\dagger$	$11 \pm 11$ per cent $^\dagger$
Dark atomic mass* ( $M_\odot$ )	$\Delta\text{BIC} > 0$	$(2.1 \pm 1.2) \times 10^7$	$(8.8 \pm 3.4) \times 10^6$
Per cent of spectra		1.6 per cent	0.8 per cent
Per cent mass difference		$1.3 \pm 0.7$ per cent	$0.5 \pm 0.2$ per cent

Note. \* Including 1.36 factor for He and metals.

in M31 and M33. When using the opaque model (equation 1), we calculate similar inferred opaque correction factors from the dark H I intensity (equation 8) and the optically thin atomic mass conversion described in Section 5. However, in this section, we expand these estimates to demonstrate the effect of (i) model selection criteria and (ii) the inherent uncertainty of the opaque model fits.

When applying different choices of the model selection criteria, we find that the dark H I mass estimate varies drastically. To show this, we calculate the inferred mass correction factors from the opaque model when (i) no model selection is applied (i.e. all opaque model fits), (ii) fits where the inferred peak opacity is  $\tau_p < 5$  (i.e. to test how sensitive the mass correction factor is to extreme, but allowed, highly opaque model fits), and (iii) where the opaque model is preferred (i.e. fits with  $\Delta\text{BIC} > 0$ ). Table 3 shows these mass correction factors for the different selections and the fraction of the fit spectra included. In all three cases, we find that the mass correction factors have large uncertainties.

First, we calculate the mass correction factors with no model selection ( $\Delta\text{BIC}$ ) applied, and so all of the opaque fits are included. We further note that this is the closest comparison to the mass correction factors computed (Braun et al. 2009; Braun 2012), since their restriction of  $\chi_r^2 < 25$  removes none of our fits (Section 6.1). For M33, we recover a correction factor of  $21 \pm 15$  per cent, which is smaller than the  $\sim 36$  per cent found by Braun (2012), but is consistent within the large uncertainty. For M31, we find a much larger correction factor of  $118 \pm 47$  per cent compared to M33. Despite the large uncertainty, this correction factor is not consistent within the  $1\sigma$  uncertainty to the 30 per cent factor found by Braun et al. (2009). The reason for finding such a large correction factor is due to including opaque fits to equation (1) with very large ( $\tau_p > 5$ ) inferred peak opacities, which correspond to large dark H I masses that often exceed the corresponding optically thin H I mass. These lines of sight are evident in Figs 6 and 7 as the contours with large negative  $\Delta\text{BIC}$  values with  $I_{\text{dark}} \sim 10^4$  K km s $^{-1}$ . The preference for  $\Delta\text{BIC} \ll -10$  indicates that the fits to the opaque model are

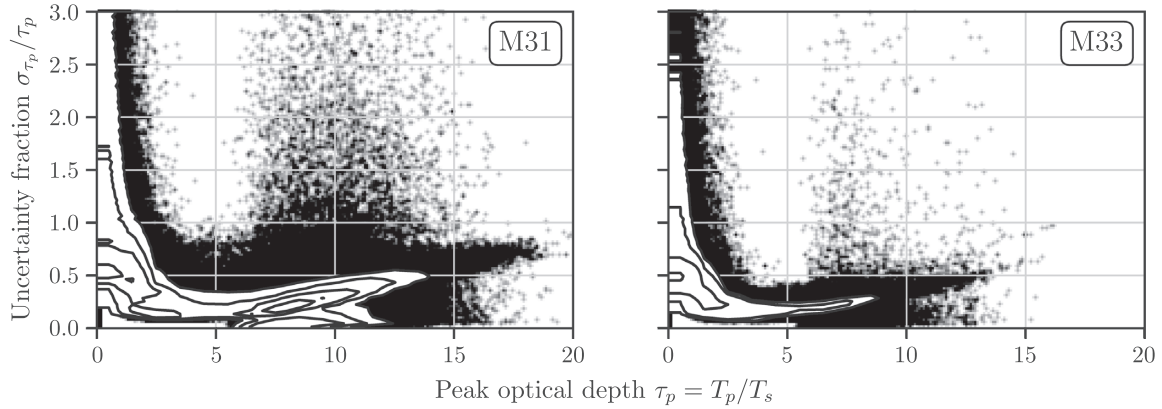
poor. In particular, we note that these are likely the lines of sight that were excluded based on the  $\chi_r^2 < 25$  restriction used in Braun et al. (2009) and Braun (2012). Here, however, we do not find that this criterion removes these cases of poor fits, demonstrating that an absolute threshold for a fit statistic is dependent on the data that are fit.

While we expect these lines of sight with a large inferred opacity from equation (1) to be poor fits, strictly speaking, these large  $\tau$  values are acceptable in the model. Additionally, Milky Way studies towards extreme star-forming regions, like W43, do infer lines of sight with a 240 per cent increase in atomic gas mass compared to the optically thin mass (Bihr et al. 2015). However, such large correction factors seem to be rare in the Milky Way (Wang et al. 2020a) and may result from many cold and warm atomic gas layers towards these lines of sight through the Galactic plane.

We control for the possibility of very large  $\tau_p$  being poor opaque model fits by excluding fits with  $\tau_p > 5$ . This threshold is based on Fig. 11, where we show that the uncertainty increases for opaque model fits with approximately<sup>10</sup>  $\tau_p > 5$ . Using this threshold, we compute mass correction factors of  $11 \pm 14$  per cent for M31 and  $11 \pm 11$  per cent for M33. We note that negative correction factors are not defined based on equation (8) and that the large uncertainty range only reflects the break-down of our assumption of normally distributed uncertainties near a mass correction 0 per cent, or the optically thin limit. Removing these likely poor fits with large  $\tau_p$  significantly decreases the correction factors, however, large uncertainty ranges persist.

In the final comparison, we use the results from our model comparison test in Section 4 to only include fits where the opaque model is preferred ( $\Delta\text{BIC} > 0$ ). Because the opaque model is preferred in only  $\sim 1$  per cent of the spectra in both galaxies, the corresponding

<sup>10</sup>The exact threshold on  $\tau_p$  is not critical as we impose this cut to demonstrate how it affects the mass correction factor.



**Figure 11.** Peak optical depth ( $\tau_p = T_p/T_s$ ) versus its uncertainty fraction ( $\sigma_{\tau_p}/\tau_p$ ) for M31 (right) and M33 (left). Contours and outliers levels are defined in Fig. 6. In the optically thin limit, the uncertainty on  $\tau_p$  is large as  $T_s$  is unconstrained. However, the uncertainty also becomes larger for  $\tau_p > 5$ , particularly in M31. Large uncertainties for large  $\tau_p$  produce large uncertainties in the inferred dark H I mass based on the opaque model fits (equation 1).

mass correction factors are also small. We find correction factors of  $1.3 \pm 0.7$  per cent for M31 and  $0.5 \pm 0.2$  per cent for M33.

Together, these three approaches for calculating the opaque mass correction factor demonstrate its sensitivity to (i) the parameter space that is allowed for the model, and (ii) the ‘goodness-of-fit’ criteria. These factors make it difficult to robustly determine the opaque mass correction factor from fits to the single opaque component model (equation 1).

In each of the cases presented above, the uncertainty of opaque mass correction factor is large. This uncertainty results from the non-linear relation between the opaque H I intensity  $I_{\text{dark}}$  on  $\tau_p = T_{\text{peak}}/T_{\text{spin}}$ . The fit parameters  $T_p$  and  $T_s$  have appreciable uncertainties from the fits, and this is reflected in the uncertainty on  $\tau_p$  (see Fig. 11). The large uncertainty on  $I_{\text{dark}}$ , and consequently on deriving the dark atomic mass, adds to the difficulty in robustly measuring the opaque mass correction factor. We note that this uncertainty was not explored in Braun et al. (2009) and Braun (2012).

Finally, we highlight that finding a  $\sim 1$  per cent correction factor does *not* imply a lack of opaque or cold H I in either M31 and M33. Our results only demonstrate that H I spectra are not well described by a single opaque H I component. We discuss the limitations to this model and our comparisons further in Section 6.4.

### 6.3 Comparisons with H I absorption studies of M31 and M33

Observations of H I absorption provide strong constraints on the H I opacity that can be compared with the single component opaque model fit to the H I emission. Dickey & Brinks (1993) present H I absorption measurements against 11 and 10 background continuum sources ( $>5$  mJy) for M31 and M33, respectively. By matching the measured absorption to the H I emission on larger scales, they infer the fraction of H I mass in the CNM,<sup>11</sup>  $f_{\text{CNM}}$ , along these lines of sight and find average values of  $f_{\text{CNM}} \approx 0.4$  in M31 and 0.15 in M33. Though measured towards only a small number of lines of sight, these  $f_{\text{CNM}}$  values fall within a similar range to those measured in the Milky Way ( $f_{\text{CNM}} \approx 0.3$ ; Stanimirović et al. 2014; Murray et al. 2018a), LMC ( $f_{\text{CNM}} \approx 0.3$ ; Marx-Zimmer et al. 2000), and SMC ( $f_{\text{CNM}} \approx 0.2$ ; Jameson et al. 2019).

<sup>11</sup>Dickey & Brinks (1993) assume that  $T_{\text{spin}} = 60$  K. Lowering  $T_s$  will decrease  $f_{\text{CNM}}$ .

To compare with  $f_{\text{CNM}}$  values from H I absorption studies, we use the opaque model fits to estimate  $f_{\text{CNM}}$  where the inferred spin temperature is constrained and implies the presence of CNM ( $T_{\text{spin}} < 250$  K; Murray et al. 2018a). Following Braun et al. (2009), we compute the effective total H I intensity from the opaque model including the ‘dark’ component ( $I_{\text{op. model}}$ ), plus the difference of the observed integrated intensity (i.e. the zeroth moment) minus the opaque model fit:

$$I_{\text{total}} = I_{\text{op. model}} + \left( I_{\text{int}} - \int T_s [1 - \exp[-\tau(v)]] dv \right), \quad (9)$$

where  $I_{\text{int}}$  is the zeroth moment integrated intensity, and the second term is assumed to be the optically thin component. Since we compare only the fits where  $T_{\text{spin}} < 250$  K,  $I_{\text{CNM}} = I_{\text{op. model}}$ , and the CNM fraction can be estimated as

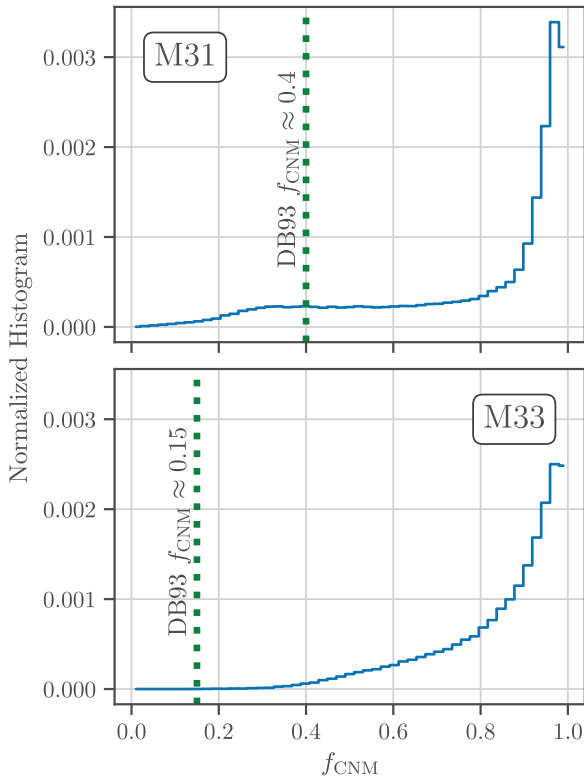
$$f_{\text{CNM}} = \frac{I_{\text{CNM}}}{I_{\text{total}}}. \quad (10)$$

Fig. 12 shows that the distributions of  $f_{\text{CNM}}$  from the opaque model tend to be quite large, with a peak near  $f_{\text{CNM}} \sim 1$  in both galaxies. Such large  $f_{\text{CNM}}$  values are difficult to reconcile with the  $f_{\text{CNM}} < 0.4$  found from H I absorption studies in the Milky Way and other Local Group galaxies, as discussed above. However, the lines of sight with  $f_{\text{CNM}} \sim 1$  correspond to those with the largest  $\tau$ , lowest  $T_s$ , and largest  $I_{\text{dark}}$  (equation 8). As we show in Section 4.3, these lines of sight are also those with the *strongest* preference for the multicomponent Gaussian model (Figs 6 and 7). Therefore, the discrepancy in  $f_{\text{CNM}}$  with H I absorption studies is consistent with the single opaque model poorly representing the observed H I spectra.

### 6.4 Limitations of the H I models

We compare two models to describe H I in this work and each has its own limitations. We describe these limitations here and how this limits making general conclusions about the H I opacity in M31 and M33.

The opaque model that we use, similar to the model used in Braun et al. (2009) and Braun (2012), is primarily limited by having a single component. The example spectra in Fig. 11 demonstrate this limitation clearly as many show  $>1$  distinct components. This is what drives our model comparison results (Section 4). However, it does not test whether those individual components are better modelled by this opaque model.



**Figure 12.** Histograms of  $f_{\text{CNM}}$  measured from the single opaque component model where the fits are consistent with CNM temperatures ( $T_{\text{spin}} < 250$  K; equation 10). The average  $f_{\text{CNM}}$  derived from H I absorption studies by Dickey & Brinks (1993, DB93) are shown with the vertical green dotted lines. Though the H I absorption measurements are based on a small number of lines of sight, the single opaque component model infers much higher  $f_{\text{CNM}}$ . In particular, the histograms peak near  $f_{\text{CNM}} \sim 1$ , which would imply a population of lines of sight with *only* CNM, inconsistent with H I absorption studies of the Milky Way and other Local Group galaxies.

The limitation in applying the opaque model to individual H I components is that it is difficult to distinguish independent components versus those that are  $>1$  Gaussians blended together (Rohlfs et al. 1972; Peek et al. 2011). One way to overcome this limitation is to identify components that are most likely to have opaque H I based on, for example, the H I components associated with tracers of molecular gas.

The key limitation in the multicomponent Gaussian model is the inherent uncertainty when components overlap because Gaussians do not form an orthogonal basis set. This means that the properties of individual components from the multi-Gaussian fit models (Fig. 4) should be treated carefully for physical interpretations (Rohlfs et al. 1972; Haud 2000). Our fitting approach uses spatial similarities to improve spectral fitting, which encourages spatial coherence, but remains uncertain where blended components are correlated. Improving how spatial information is included in the spectral fits can minimize some of this uncertainty (Marchal et al. 2019; Riemer et al. 2020).

There are additional H I features that are also not accounted for in either model. For example, we do not include the possibility for HISA features, which can be distinguished from the opaque H I model as a narrow absorption ‘dip’, e.g. as those frequently seen in Milky Way H I spectra (Gibson et al. 2005). While self-absorption may be detected towards some lines of sight, we do not expect our observations to be sensitive to strong HISA features due to the  $>60$  pc

linear resolution. In agreement with this, Koch et al. (2019) find the velocity at peak H I and CO brightness temperature to be typically consistent within the  $2.6 \text{ km s}^{-1}$  CO velocity channel size, rather than additional scatter or offsets in this relation if the CO peaked at a strong self-absorption feature. In qualitative agreement with these results, Liu et al. (2019) use CO components to derive HISA in the LMC and find relatively small corrections on 15 pc scales.

### 6.5 Future directions for identifying opaque H I in nearby galaxies

Our results show that very few lines of sight in M31 and M33 are well described by a single opaque H I component and that a multicomponent model is strongly preferred. This indicates that we are likely recovering a mixture of atomic ISM phases, in addition to multiple kinematic components reflecting galactic structure. For example, our M33 H I observations (Koch et al. 2018b) detect a lagging thick H I disc in M33 (see also Kam et al. 2017). Disentangling this spectral complexity and linking it to the underlying atomic ISM structure is an obvious next step in exploring these observations. One crucial link in this next step will be to determine whether components of the H I emission spectra *can* identify cold opaque H I, thereby providing an improved estimate of the ‘dark’ opaque H I mass.

With current observations, we suggest three approaches to further search for opaque H I without requiring H I absorption observations against bright background sources:

(i) The line width of H I components can distinguish between emission from different atomic ISM phases (e.g. Haud 2000; Marchal et al. 2019). However, the degeneracy of line broadening from turbulence limits line width measurements to an upper limit of the H I temperature. In the simplest case of a constant WNM temperature of 6000 K, measured line widths  $<6 \text{ km s}^{-1}$  must be non-WNM. For much narrower line widths ( $\leq 2.5 \text{ km s}^{-1}$ ), the equivalent temperature is  $<1000$  K without including turbulent broadening, making it more likely that these features are indeed from the CNM. We identify few distinct narrow components from the multicomponent Gaussian fits, however, there remains ambiguity about the completeness of these narrow components given the issues with blended components (see Appendix B1). This issue will be explored in Koch et al. (in preparation).

Resolving finer spatial scales will also assist in resolving some of these spectral blending issues, since the blending will be minimized when the associated spatial extent of these components becomes resolved. We have in-hand additional VLA H I observations resolving  $10 \text{ arcsec} \approx 35 \text{ pc}$  scales in M31, and  $8 \text{ arcsec} \approx 30 \text{ pc}$  in M33 that we will use to explore this avenue (Koch et al. in preparation).

(ii) The degeneracy between spectral components is a key limitation of the multicomponent Gaussian modelling. As we note in Section 6.4, using constraints from other spectral lines can assist in resolving between this degeneracy by restricting where we search for signatures of opaque H I. Additionally, these comparisons could lead to HISA detections (e.g. Liu et al. 2019; Wang et al. 2020b).

(iii) ‘Dark’ neutral ISM mass can further be constrained by combining dust, atomic ISM, and molecular ISM tracers. Similar comparisons have been made in both M31 and M33 (Leroy et al. 2011; Dalcanton et al. 2015; Gratier et al. 2017), as well as nearby galaxies (Sandstrom et al. 2013), though typically on  $>100 \text{ pc}$  scales. Extending similar comparisons to  $\sim 30 \text{ pc}$  scales will be helpful for resolving between  $\text{H}_2$ -dominated lines of sight towards GMCs and the surrounding atomic gas, which may include a cold, opaque component (e.g. see Williams et al. 2019, and Smith et al.,



in preparation for IR dust observations on these scales in M33 and M31).

One issue in using the dust-H I-CO comparison will be distinguishing whether the ‘dark’ ISM component is opaque H I or CO-dark H<sub>2</sub> (e.g. see Murray et al. 2018b, for constraints on this distinction in the Milky Way). However, this comparison on  $\sim 30$  pc scales will provide strong constraints on the *total* dark gas mass, which has otherwise been limited to statistical arguments over large spatial areas (e.g. Gratier et al. 2017). The H I line shape and possibility of individual opaque components, as described above, may be the key for making this distinction. Additional observations of C II and/or C I can help resolve this degeneracy in the phase of a ‘dark’ component (Bisbas, Schruha & van Dishoeck 2019; Madden et al. 2020), however, mapping the large angular areas covered by M31 and M33 at high resolution remains prohibitive for current sub-mm and IR telescopes.

While the approaches described above seem hopeful, there remains the clear advantage of a direct measurement of H I opacity from the absorption against bright background sources, despite the limited number of lines of sight where the absorption can be measured (Dickey & Brinks 1993). Though spatially limited, these H I absorption measurements are crucial to compare with opacity measurements from other methods (e.g. Fig. 12).

## 7 SUMMARY

In this paper, we explore whether extragalactic H I emission can be described by a single opaque velocity component model to constrain the H I line opacity and measure the ‘missing’ opaque atomic ISM mass. Using new sensitive 21-cm H I observations at high spectral resolution in M31 and M33, we compare two common spectral models: a multi-Gaussian model and a single-component, opaque H I model (e.g. Braun et al. 2009). With the improved spectral resolution of our data, we demonstrate a strong preference for the multi-Gaussian model for  $>80$  per cent of the spectra, while  $<2$  per cent prefer the opaque model and the remainder does not have a strong model preference. Adopting the opaque model, we further compare the inferred H I opacity corrections to dust emission and dust extinction properties, where we find that the H I opacity corrections require large variations in dust abundance that lack observational evidence. Combining our model selection and dust comparison results, we show that models for a single opaque H I component are not supported by the observations, which demonstrates there is a lack of lines of sight dominated by *only* cold and opaque H I on  $\sim 80$  pc scales. Finally, we show that the inferred optical depth from the opaque model fit to H I emission alone is subject to significant systematic uncertainty, which in turn leads to large uncertainties on the inferred opaque H I mass. Our main conclusions from this work are:

(i) Improvements in the spectral resolution of extragalactic 21-cm H I data reveal a wealth of spectral structure. The observations we present extend detailed H I studies that were previously restricted to the Milky Way and the Magellanic Clouds.

(ii) We identify a strong preference for an optically thin, multi-Gaussian model over a single opaque velocity component across both M31 and M33 by comparing the difference of the BICs between the two models (Fig. 5). This preference persists when considering only opaque models with a large inferred optical depth ( $\tau_p > 1$ ), and when the BIC is computed only over velocities where the opaque model measures  $\tau(v) > 0.5$ .

(iii) We verify our model selection test by applying the same fitting procedures to 20 000 spectra drawn from mock observations of the

opaque model, sampled at the same  $0.42 \text{ km s}^{-1}$  channel width and noise levels as the observations. We find that  $<1.5$  per cent prefer the multi-Gaussian model, which is the opposite trend the we find from the observations. This test demonstrates that our observations and analysis methods are sensitive enough to distinguish between the two models.

(iv) The inferred opaque H I correction factors produce highly uncertain estimates of the opaque H I mass. Using the opaque correction factor from all lines of sight, we find mass correction factors of  $118 \pm 47$  per cent for M31 and  $21 \pm 15$  per cent for M33, where the ranges are  $1\sigma$  uncertainties. However, these mass correction factors reduce to just  $\sim 1$  per cent in both galaxies when we restrict to only using lines of sight where the single opaque component model is preferred. This large range, coupled with the large uncertainty, demonstrates that the inferred opaque H I mass factors are highly sensitive to the goodness-of-fit threshold and the allowed parameter range for the fit.

(v) We show that the scatter in the relation between  $\Sigma_{\text{atomic}}$  and  $\Sigma_{\text{dust}}$  is significantly increased when the inferred H I opacity corrections are included (Figs 9 and 10). This is the opposite of what is expected if the H I surface density is missing a component from the optically thin assumption relative to the dust, tracing the total neutral gas surface density. This increased scatter from the inferred H I opacity correction can only be explained for a population of lines of sight with large gas-to-dust ratios ( $\text{GDR} \approx 1000$ ) and  $A_V < 2$  mag over a range of  $\Sigma_{\text{atomic}} \approx 20\text{--}80 \text{ M}_{\odot} \text{ pc}^{-2}$ , which lacks observational evidence. Since the scatter is reflected in  $A_V$  and the dust surface density from IR SED fitting, this rules out the possibility of cold dust not accounted for in the far-IR.

Our results highlight that estimates of the opaque H I mass from the opaque spectral line model are not reliable on 100 pc scales. The large difference in the inferred dark atomic H I mass based on fits to the opaque model points at the critical role of model assumptions on interpretations of the atomic ISM. We describe how future studies may improve on this modelling. However, given our current results, we suggest that H I opacity is best constrained by comparing H I absorption to background sources with nearby emission (e.g. Dickey et al. 2003), or through HISA (e.g. Gibson et al. 2005; Wang et al. 2020b), though current instrumentation limits the use of these methods in M31 and M33.

## ACKNOWLEDGEMENTS

We thank the reviewer for insightful comments that improved the clarity of this work. EWK and EWR are supported by a Discovery Grant from NSERC (RGPIN-2017-03987). EWK acknowledges support in completing this paper from the Smithsonian Institution as a Submillimeter Array (SMA) Fellow. TGW acknowledges funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. 694343). The work of JC, KS, IC, AKL, and DU is supported by NASA ADAP grants NNX16AF48G and NNX17AF39G and National Science Foundation grant No. 1615728. The National Radio Astronomy Observatory and the Green Bank Observatory are facilities of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. This research was enabled in part by support provided by WestGrid ([www.westgrid.ca](http://www.westgrid.ca)), Compute Canada ([www.computecanada.ca](http://www.computecanada.ca)), and CANFAR ([www.canfar.net](http://www.canfar.net)).

**Code Bibliography:** CASA (versions 4.4, 4.7, 5.4; McMullin et al. 2007); ASTROPY (Astropy Collaboration 2013, 2018); radio-astro-



tools (spectral-cube (Ginsburg et al. 2019), radio-beam, uvcombine; [radio-astro-tools.github.io](https://radio-astro-tools.github.io)); MATPLOTLIB (Hunter 2007); SEABORN (Waskom et al. 2017); LMFIT (Newville et al. 2020); NUMPY & SCIPY (Virtanen et al. 2020); UNCERTAINTIES (Lebigot 2020).

## DATA AVAILABILITY

The software and code used in this paper are publicly available. Scripts to reproduce the results in the paper are available at <https://github.com/e-koch/HI-LineShapes><sup>12</sup> The H I data products for M33 were released with Koch et al. (2018b). The H I data products for M31 will be publicly released upon acceptance of in preparation manuscripts.

## REFERENCES

- Astropy Collaboration, 2013, *A&A*, 558, A33
- Astropy Collaboration, 2018, *AJ*, 156, 123
- Bühr S. et al., 2015, *A&A*, 580, A112
- Bisbas T. G., Schruha A., van Dishoeck E. F., 2019, *MNRAS*, 485, 3097
- Blagrave K., Martin P. G., Joncas G., Kothers R., Stil J. M., Miville-Deschênes M. A., Lockman F. J., Taylor A. R., 2017, *ApJ*, 834, 126
- Braun R., 1990, *ApJS*, 72, 761
- Braun R., 1997, *ApJ*, 484, 637
- Braun R., 2012, *ApJ*, 749, 87
- Braun R., Thilker D. A., Walterbos R. A. M., Corbelli E., 2009, *ApJ*, 695, 937
- Caldú-Primo A., Schruha A., 2016, *AJ*, 151, 34
- Chiang I. D., Sandstrom K. M., Chastenot J., Johnson L. C., Leroy A. K., Utomo D., 2018, *ApJ*, 865, 117
- Corbelli E., Lorenzoni S., Walterbos R., Braun R., Thilker D., 2010, *A&A*, 511, A89
- Cornwell T. J., 2008, *IEEE J. Sel. Top. Signal Process.*, 2, 793
- Dalcanton J. J. et al., 2012, *ApJS*, 200, 18
- Dalcanton J. J. et al., 2015, *ApJ*, 814, 3
- Dickey J. M., Brinks E., 1993, *ApJ*, 405, 153
- Dickey J. M., McClure-Griffiths N. M., Gaensler B. M., Green A. J., 2003, *ApJ*, 585, 801
- Druard C. et al., 2014, *A&A*, 567, A118
- Ferrière K. M., 2001, *Rev. Mod. Phys.*, 73, 1031
- Field G. B., Goldsmith D. W., Habing H. J., 1969, *ApJ*, 155, L149
- Freedman W. L. et al., 2001, *ApJ*, 553, 47
- Gibson S. J., Taylor A. R., Higgs L. A., Brunt C. M., Dewdney P. E., 2005, *ApJ*, 626, 195
- Ginsburg A. et al., 2019, radio-astro-tools/spectral-cube: Release v0.4.5. doi:10.5281/zenodo.3558614
- Gordon K. D. et al., 2014, *ApJ*, 797, 85
- Gratier P. et al., 2010, *A&A*, 522, A3
- Gratier P. et al., 2012, *A&A*, 542, A108
- Gratier P. et al., 2017, *A&A*, 600, A27
- Haud U., 2000, *A&A*, 364, 83
- Heiles C., Troland T. H., 2003, *ApJ*, 586, 1067
- Henshaw J. D. et al., 2016, *MNRAS*, 457, 2675
- Hill A. S., Mac Low M.-M., Gatto A., Ibáñez-Mejía J. C., 2018, *ApJ*, 862, 55
- Hunter J. D., 2007, *Comput. Sci. Eng.*, 9, 90
- Jameson K. E. et al., 2019, *ApJS*, 244, 7
- Kam S. Z., Carignan C., Chemin L., Foster T., Elson E., Jarrett T. H., 2017, *AJ*, 154, 41
- Kass R. E., Raftery A. E., 1995, *J. Am. Stat. Assoc.*, 90, 773
- Keown J., Di Francesco J., Teimoorinia H., Rosolowsky E., Chen M. C.-Y., 2019, *ApJ*, 885, 32
- Kim S., Staveley-Smith L., Dopita M. A., Sault R. J., Freeman K. C., Lee Y., Chu Y.-H., 2003, *ApJS*, 148, 473
- Koch E., Rosolowsky E., Leroy A. K., 2018a, *Res. Notes Am. Astron. Soc.*, 2, 220
- Koch E. W. et al., 2018b, *MNRAS*, 479, 2505
- Koch E. W., Rosolowsky E. W., Schruha A., Leroy A., Kepley A., Braine J., Dalcanton J., Johnson M. C., 2019, *MNRAS*, 485, 2324
- Krumholz M. R., McKee C. F., Tumlinson J., 2009, *ApJ*, 699, 850
- Kurono Y., Morita K.-I., Kamazaki T., 2009, *PASJ*, 61, 873
- Lebigot E. O., 2020, Uncertainties: a Python package for calculations with uncertainties. Available at: <http://pythonhosted.org/uncertainties/>
- Lee M.-Y., Stanimirović S., Murray C. E., Heiles C., Miller J., 2015, *ApJ*, 809, 56
- Leroy A. K. et al., 2011, *ApJ*, 737, 12
- Leroy A. K. et al., 2020, *ApJS*, preprint (arXiv:2104.07739)
- Lindner R. R. et al., 2015, *AJ*, 149, 138
- Liu B., Li D., Staveley-Smith L., Qian L., Wong T., Goldsmith P., 2019, *ApJ*, 887, 242
- Lockman F. J., Free N. L., Shields J. C., 2012, *AJ*, 144, 52
- Madden S. C. et al., 2020, *A&A*, 643, A141
- Marchal A., Miville-Deschênes M.-A., Orieux F., Gac N., Soussen C., Lesot M.-J., d'Allonnes A. R., Salomé Q., 2019, *A&A*, 626, A101
- Marx-Zimmer M., Herbstmeier U., Dickey J. M., Zimmer F., Staveley-Smith L., Mebold U., 2000, *A&A*, 354, 787
- McMullin J. P., Waters B., Schiebel D., Young W., Golap K., 2007, in Shaw R. A., Hill F., Bell D. J., eds, ASP Conf. Ser. Vol. 376, Astronomical Data Analysis Software and Systems XVI. Astron. Soc. Pac., San Francisco, p. 127
- Murray C. E. et al., 2014, *ApJ*, 781, L41
- Murray C. E., Stanimirović S., Kim C.-G., Ostriker E. C., Lindner R. R., Heiles C., Dickey J. M., Babler B., 2017, *ApJ*, 837, 55
- Murray C. E., Stanimirović S., Goss W. M., Heiles C., Dickey J. M., Babler B., Kim C.-G., 2018a, *ApJS*, 238, 14
- Murray C. E., Peek J. E. G., Lee M.-Y., Stanimirović S., 2018b, *ApJ*, 862, 131
- Newville M. et al., 2020, lmfit/lmfit-py 1.0.1. Available at: <https://doi.org/10.5281/zenodo.3814709>
- Nieten C., Neininger N., Guélin M., Ungerechts H., Lucas R., Berkhuijsen E. M., Beck R., Wielebinski R., 2006, *A&A*, 453, 459
- Peek J. E. G., Heiles C., Peek K. M. G., Meyer D. M., Lauroesch J. T., 2011, *ApJ*, 735, 129
- Riener M., Kainulainen J., Beuther H., Henshaw J. D., Orkisz J. H., Wang Y., 2020, *A&A*, 633, A14
- Rohlfs K., Braunsfurth E., Mebold U., 1972, *AJ*, 77, 711
- Roman-Duval J. et al., 2014, *ApJ*, 797, 86
- Saintonge A. et al., 2017, *ApJS*, 233, 22
- Sandstrom K. M. et al., 2013, *ApJ*, 777, 5
- Schruha A., Bialy S., Sternberg A., 2018, *ApJ*, 862, 110
- Schwarz G., 1978, *Ann. Statist.*, 6, 461
- Sokolov V., Pineda J. E., Buchner J., Caselli P., 2020, *ApJ*, 892, L32
- Stanimirović S., 1999, PhD thesis, Univ. Western Sydney Nepean
- Stanimirović S., Murray C. E., Lee M.-Y., Heiles C., Miller J., 2014, *ApJ*, 793, 132
- Thilker D. A., Braun R., Walterbos R. A. M., 2002, in Taylor A. R., Landecker T. L., Willis A. G., eds, ASP Conf. Ser. Vol. 276, Seeing Through the Dust: The Detection of HI and the Exploration of the ISM in Galaxies. Astron. Soc. Pac., San Francisco, p. 370
- Utomo D., Chiang I. D., Leroy A. K., Sandstrom K. M., Chastenot J., 2019, *ApJ*, 874, 141
- Vilardell F., Ribas I., Jordi C., Fitzpatrick E. L., Guinan E. F., 2010, *A&A*, 509, A70
- Virtanen P. et al., 2020, *Nat. Methods*, 17, 261
- Walter F. et al., 2020, *ApJ*, 902, 111
- Wang Y. et al., 2020a, *A&A*, 634, A83
- Wang Y. et al., 2020b, *A&A*, 634, A139
- Waskom M. et al., 2017, mwaskom/seaborn: v0.8.1 (september 2017). Available at: <https://doi.org/10.5281/zenodo.883859>
- Whitworth A. P. et al., 2019, *MNRAS*, 489, 5436
- Williams T. G., Gear W. K., Smith M. W. L., 2019, *MNRAS*, 483, 5135

<sup>12</sup>Code DOI: <https://doi.org/10.5281/zenodo.4665936>.

Winkel B., Kerp J., Flöer L., Kalberla P. M. W., Ben Bekhti N., Keller R., Lenz D., 2016, *A&A*, 585, A41  
 Wolfire M. G., McKee C. F., Hollenbach D., Tielens A. G. G. M., 2003, *ApJ*, 587, 278

## APPENDIX A: IMAGING APPROACH AND SINGLE-DISH COMBINATION

In this section, we describe the  $uv$ -data handling and imaging steps for the M31 VLA observations (Section 2.1) in further detail. The procedure is similar to that used in Koch et al. (2018b) for the M33 VLA observations, with some difference unique to M31 as a target and alterations to the imaging steps.

The M31 data from individual observing sessions were transformed into the LSRK frame and observations of matched VLA configurations were concatenated. We then subtracted continuum sources from the data in the  $uv$ -plane using UVCONTSUB with a zeroth-order spectral fit. The zeroth-order constant fit does not leave noticeable residuals in the continuum-subtracted data.

Next, we combined observations taken in different configurations. Because we imaged the data at the native spectral resolution, we paid special attention to matching the spectral channel edges between the data for the three configurations. By default (as of CASA v5.6 and before), regridding over shifts smaller than the channel size in MSTRANSFORM will use linear interpolation that creates a ‘beat’ pattern in the outputted data. While the ‘fftshift’ method for regridding should handle this, we found that the implementation in CASA v5.4.1 created large deviations near the edges of the spectral range in the outputted data. These issues are discussed at length in Leroy et al. (2020).

For our M31 observations, the difference in the spectral channel edges are very similar across the different configurations ( $\ll 1$  per cent the channel width). Because of this, we split the channels according to the velocity range that will be imaged to the nearest matching channel number and concatenated these channel-matched data sets together. A small frequency tolerance was allowed in the concatenation.

To handle the large computational time needed for imaging the H I data, we imaged and deconvolved each spectral channel separately, similar to Koch et al. (2018b). Though this includes the additional step of splitting the  $uv$ -data, this approach minimizes the size of the  $uv$ -data and the time for input/output operations. Furthermore, each channel can be imaged independently and can, for example, be parallelized by deploying the imaging for each channel as separate jobs on a cluster. This approach has two benefits: (i) different combinations of imaging parameters can be rapidly explored on a small number of channels; and (ii) divergent deconvolution solutions can be corrected without the need to re-image other spectral channels. We split the H I  $uv$ -data into individual channels at the target spectral resolution – in this case, we use the native spectral resolution of the data,  $0.42 \text{ km s}^{-1}$ .

We imaged each spectral channel in two steps. First, we deconvolved the data to the  $2\sigma_{\text{rms}}$  level within a pre-defined mask. An outer  $uv$ -taper of 5 arcsec was applied to produce a  $\sim 18$  arcsec beam resolution, roughly equivalent to VLA’s C-configuration. Since the H I emission is highly extended, we used multiscale CLEAN (Cornwell 2008) with scales of [0, 18, 36, 72, 180, 360, 720, 1440] arcsec, spanning the  $uv$ -coverage of the VLA observations,<sup>13</sup> and

major cycles were triggered every 500 iterations, which assisted the convergence of the deconvolution. We produced signal masks from the previously imaged D-configuration (58 arcsec) mosaic covering the entire star-forming disc<sup>14</sup> using the masking procedure described in Koch et al. (2018b) and also applied to the M33 H I data cube. Since the beam size differs by only a factor of  $\sim 3$  between the D-configuration data and the data product derived here, the D-configuration mosaic provides excellent prior information at higher resolution in M31. Using these signal masks was critical due to the highly elongated and extended H I emission for the blue-shifted channels in this portion of M31. Without the mask, there was a tendency for the deconvolution to diverge because of the large negative bowls between the blue-shifted ‘limbs,’ which were difficult to account for with the symmetric multiscale CLEAN elements.

After the first stage, we noticed low-lying emission or mild artefacts to remain in the residual and restored images. To handle this, we continued in the deconvolution in a second step by removing the signal mask from the first step. Since the vast majority of the bright emission is deconvolved in the first step, there is no longer an issue with negative bowling causing a divergence. We continue to deconvolve until the residuals in the entire map reach the  $2\sigma_{\text{rms}}$  level.

All imaging was run on Compute Canada’s Cedar cluster.<sup>15</sup> Each channel was imaged in parallel using eight cores and at least 4 GB of memory per core. We imaged four channels at once on a single 32 core node with 128 GB of memory, the standard size available at the time on the Cedar cluster. Each imaging stage required 20–30 h of wall time to complete per channel.

Once the imaging was completed, we recombined the channels to form the final spectral line data cube.

Finally, we included short-spacing information by feathering the VLA data cube with Effelsburg H I observations from EBHIS (Winkel et al. 2016) using the UVCOMBINE package.<sup>16</sup> The EBHIS data have a  $10''.8$  beam size that moderately overlaps with the shortest VLA baselines covering scales up to 16 arcmin. The overlapping range is less than the factor of 1.7 recommended by Kurono, Morita & Kamazaki (2009) for deriving a consistent scale factor in the overlap of the two data sets, however, we combine overlapping  $uv$  scales over many spectral channels to increase the number of samples to improve statistical estimates.

To combine the two datasets, we spectrally oversample the EBHIS data cube from the original  $1.3 \text{ km s}^{-1}$  to match the  $0.42 \text{ km s}^{-1}$  resolution of the VLA data. Because of the coarse resolution ( $10''.8 \approx 2.6 \text{ kpc}$ ), the H I spectra with bright emission are smooth and upsampling by  $\sim 3\times$  does not affect the single-dish spectral shape. We note, however, that this assumption about the upsampling holds only at very high S/N.

To ensure the data are optimally combined, we performed two  $uv$ -overlap tests recommended by Stanimirovic (1999) and also used for the M33 data (Koch et al. 2018b). Both of these are implemented in the UVCOMBINE package. First, we test whether there is a non-zero slope between the ratio of  $I_{\text{VLA}}/I_{\text{EBHIS}}$  with the  $uv$ -distance when weighted by the single-dish beam, as is used in feathering. If there is a non-zero slope, it suggests that the weighting from the single-dish data differs from the data, and the beam size moderately differs from what was assumed. We find no trend with  $k$  that would be consistent with the reported  $10''.8$  EBHIS beam size from Winkel et al. (2016)

<sup>13</sup>Different choices of scales for the multiscale CLEAN did not clearly change the final restored images.

<sup>14</sup>To be presented in Koch et al. (in preparation).

<sup>15</sup>[docs.compute.canada.ca/wiki/Cedar](https://docs.compute.canada.ca/wiki/Cedar)

<sup>16</sup>[github.com/radio-astro-tools/uvcombine](https://github.com/radio-astro-tools/uvcombine)

Second, we calculated the typical ratio of  $I_{\text{VLA}}/I_{\text{EBHIS}}$  to determine whether the flux calibration agrees between the two data sets. This test is best performed when the interferometric data cover most of the observed source, particularly for the large extended structure in M31 and the large EBHIS beam size. Because of this, we did this test using the full VLA D-configuration mosaic rather than the smaller 18 arcsec mosaic used throughout this paper.<sup>17</sup> Using the full mosaic improves the ratio test uncertainty as it includes more signal from the source, as well.

We take the Fourier transform of 500 channels in both the VLA and EBHIS data and sample the angular scales from 10.816 arcsec that both the EBHIS and VLA observations are sensitive to. We then combine these samples across channels assuming the flux scaling does not change across a narrow spectral range. We then model the histogram of  $I_{\text{VLA}}/I_{\text{EBHIS}}$  as Cauchy (or Lorentzian) distribution, following Koch et al. (2018b), that accounts for high and low tails in the distribution from noise in both the numerator and denominator. From the fitted Cauchy distribution, we find a relative scaling factor of  $1.07 \pm 0.01$ . Since the ratio is  $> 1$ , this scaling factor applied to the EBHIS data increased the flux by 10 per cent. In an independent test, Blagrove et al. (2017) find a similar scaling factor of  $1.10 \pm 0.01$  was needed for combining their observations from the Dominion Radio Astrophysical Observatory Synthesis Telescope with short-spacing from EBHIS.

The final VLA+EBHIS M31 H I data cube is the feathered product after applying a scaling factor of 1.07 to the EBHIS data. The data have a beam size of  $18 \text{ arcsec} \times 16 \text{ arcsec}$  and a sensitivity of  $2.8 \text{ K per } 0.42 \text{ km s}^{-1}$  channel. Within each channel, this corresponds to a  $5\sigma_{\text{rms}}$  optically thin H I column density of  $9.8 \times 10^{18} \text{ cm}^{-2}$ . This sensitivity is roughly equivalent to the sensitivity of the high-resolution WSRT H I observations from Braun et al. (2009) in  $2.3 \text{ km s}^{-1}$  spectral channels.

## APPENDIX B: MULTI-GAUSSIAN MODEL TESTING

In Section 3.2, we describe our approach for modelling the H I spectra with multiple Gaussian components, building on previous work by Lindner et al. (2015) and Riener et al. (2020). Here, we test this method on synthetic spectra drawn from randomly selected multi-Gaussian models sampled at the  $0.42 \text{ km s}^{-1}$  spectral resolution of our observations with similar noise levels added.

We produce multi-Gaussian synthetic spectra in three steps:

(i) We draw the number of Gaussian components randomly between 1 and 8, where the values are weighted based on the fraction of M31 spectra with a given number of Gaussian components (e.g.  $\sim 40$  per cent with 1 component,  $\sim 30$  per cent with 2 components,  $\sim 20$  per cent with 3 components, and  $\sim 10$  per cent with  $> 3$  components).

(ii) Next, we randomly draw the amplitude, centroid, and line width from uniform distributions again matching the range of parameters recovered from the observations. Specifically, we sample parameters from uniform distributions:

$$A \sim \mathcal{U}(2.7, 120) \text{ K} \quad (\text{B1})$$

$$v_0 \sim \mathcal{U}(-50, 50) \text{ km s}^{-1} \quad (\text{B2})$$

<sup>17</sup>When using a smaller mosaic, the single-dish observations are tapered by the interferometric primary beam response, and this test depends more strongly on the low level behaviour of the response function.

$$\sigma \sim \mathcal{U}(0.1, 60) \text{ km s}^{-1}. \quad (\text{B3})$$

The lower limit for  $A$  is the rms noise in the M31 data. We allow for narrow features below the  $0.42 \text{ km s}^{-1}$  channel width to test the recovery of unresolved features where the averaging within channels biases the line shape (e.g. Koch et al. 2018a).

While the observed parameter distributions show correlations between the amplitude and line width, we do not include these correlations for the synthetic spectra. This choice to draw from uncorrelated uniform distributions provides a simpler way to determine completeness limits for the observed parameter distributions.

(iii) The synthetic spectrum is evaluated at  $4\times$  finer spectral resolution compared to the observations. We then average this model spectrum to reach  $0.42 \text{ km s}^{-1}$  channels, matching the observations. Finally, we add Gaussian noise sampling from  $\mathcal{N}(0, 2.7)$  to match the M31 rms noise.

We draw 10 000 synthetic spectra to test our Gaussian fitting procedure. To examine the affect of the noise alone, 10 different noise draws are added to each synthetic spectrum and fit. Though not fully independent, this yields a total 100 000 fits to examine the fitting method. From fitting the synthetic spectra, we present two results: (1) the recovery fraction between the actual and fit number of Gaussian components, and (2) the parameter distributions of components correctly recovered in the fits.

Due to the issue of covariance between Gaussian components, we introduce a ‘distinctness’ criterion to distinguish between Gaussian components that are blended (large covariance) versus distinct (small covariance). Distinct components in a multi-Gaussian spectrum are characterized by a clear peak, corresponding to a local minimum in the second derivative, which allows for that shape to be well constrained when fitting. Because the component shape is distinguished in the multi-Gaussian spectrum, we can identify distinct components using a similar approach to the ‘derivative spectroscopy’ we use in our fitting method (Section 3.2; Lindner et al. 2015).

### B1 Classifying distinct and blended Gaussian components

We define a component as ‘distinct’ when the ratio between the second derivative of the total model versus the single component exceeds  $f_{\text{Distinct}} = 0.75$ :

$$f_{\text{Distinct}} = \frac{\frac{d^2 T_b(v_{0,i})}{dv^2}}{\frac{d^2 T_{b,i}(v_{0,i})}{dv^2}}, \quad (\text{B4})$$

following the notation of equation (4) for the total spectrum  $T_b$  and the  $i$ th component  $T_{b,i}$ . When the component is strongly blended,  $f_{\text{Distinct}}$  is small since the shape is not apparent in the total spectrum and its second derivative will not be strongly peaked. On the other hand, a distinct peak in the total spectrum will have a large  $f_{\text{Distinct}}$  as both derivatives will contain a similar local minimum. In the trivial case of a single Gaussian component,  $f_{\text{Distinct}} = 1$ . We classify each component as distinct or blended using this criterion.

### B2 Number of components recovered

Our key results in this paper rely on the  $\Delta\text{BIC}$  comparison, where the BIC value for each fit depends on the number of free parameters. Since the multi-Gaussian model is flexible, as the number

of components can vary, it is important for the fitting method to broadly recover the correct number of components. These issues are discussed in depth in several works (e.g. Lindner et al. 2015; Riener et al. 2020).

Fig. B1 shows the actual versus recovered number of Gaussian components for all, distinct, and blended components. We find a broad range in the number of components identified by the fitting routine, though for all components, there is a bias to underestimate the true number of components for complex spectra ( $N_{\text{comp}} > 4$ ). By splitting the components into distinct and blended sets, we find that this bias for complex spectra is due to blended components, which are difficult to fully recover from a single spectrum with no additional information. On the other hand, our fitting method performs well for recovering distinct components.

These results demonstrate that the number of components identified by our multicomponent Gaussian fitting method is not systematically biased, consistent with previous works using this method (Lindner et al. 2015; Murray et al. 2017; Riener et al. 2020), and supports our conclusions for the  $\Delta\text{BIC}$  model selection in Section 4.3.

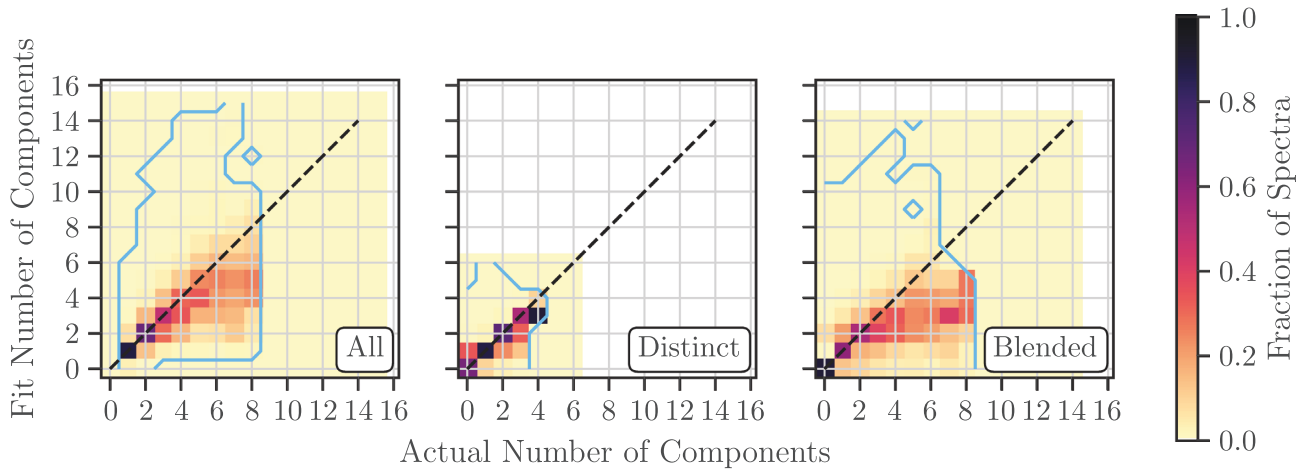
### B3 Recovery of Gaussian components

The investigation of recovering Gaussian component parameters, both in these tests and applied to the observations, is beyond the scope of this work and will be presented in Koch et al. (in preparation). Here, we note two conclusions regarding the recovery of the correct component parameters:

(i) By comparing to the uniform distribution of drawn parameters (equation B1), we find broad agreement over most of the chosen parameter space. The recovery fraction decreases for (i) extremely wide ( $>40 \text{ km s}^{-1}$ ; usually highly blended) components and (ii) for narrow ( $<1.2 \text{ km s}^{-1}$ ) components with  $S/N < 5$ .

(ii) Narrower components have higher recovery fractions as they are more likely to be classified as distinct (equation B4) and will less overlap with other components. This introduces a bias that may affect the average component line width and requires further exploration.

While the above points affect the component distributions of the multi-Gaussian fits, they do not strongly influence the overall fit and so do not affect our model selection results.



**Figure B1.** Confusion matrices comparing the actual versus fitted number of Gaussian components from the synthetic spectra for all (left), distinct (centre), and blended (right) components. The values shown are fractions normalized by column, and the cyan contours indicate the location of all non-zero values. We find a tendency to underestimate the true number of all components for complex spectra ( $N_{\text{comp}} > 4$ ), however, this is primarily due to blended components which are poorly distinguished in the total model. In contrast, the number of distinct components is well recovered in the fits. We conclude that our multi-Gaussian fitting is not systematically biased, thereby strengthening the comparison with the opaque model (equation 1).

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