The THESAN project: Lyman-α emitter luminosity function calibration

Clara Xu [ID],[1]⋆ Aaron Smith [ID],[2,1]† Josh Borrow [ID],[1] Enrico Garaldi [ID],[3] Rahul Kannan [ID],[2,4]

Mark Vogelsberger [ID],[1,5] Rüdiger Pakmor [ID],[3] Volker Springel [ID][3] and Lars Hernquist[2]

[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*
[2]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[3]*Max-Planck Institute for Astrophysics, Karl-Schwarzschild-Str. 1, D-85741 Garching, Germany*
[4]*Department of Physics and Astronomy, York University, 4700 Keele St., Toronto, Ontario, Canada, MJ3 1P3*
[5]*The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, Massachusetts Institute of Technology, Cambridge MA 02139, USA*

ABSTRACT

The observability of Lyman-alpha emitting galaxies (LAEs) during the Epoch of Reionization can provide a sensitive probe of the evolving neutral hydrogen gas distribution, thus setting valuable constraints to distinguish different reionization models. In this study, we utilize the new THESAN suite of large-volume ($L_{\rm box} = 95.5$ cMpc) cosmological radiation-hydrodynamic simulations to directly model the Lyα emission from individual galaxies and the subsequent transmission through the intergalactic medium. THESAN combines the AREPO-RT radiation-hydrodynamic solver with the IllustrisTNG galaxy formation model and includes high- and medium-resolution simulations designed to investigate the impacts of halo-mass-dependent escape fractions, alternative dark matter models, and numerical convergence. We find important differences in the Lyα transmission based on reionization history, bubble morphology, frequency offset from line centre, and galaxy brightness. For a given global neutral fraction, Lyα transmission reduces when low mass haloes dominate reionization over high mass haloes. Furthermore, the variation across sightlines for a single galaxy is greater than the variation across all galaxies. This collectively affects the visibility of LAEs, directly impacting observed Lyα luminosity functions (LFs). We employ Gaussian Process Regression using SWIFTEMULATOR to rapidly constrain an empirical model for dust escape fractions and emergent spectral line profiles to match observed LFs. We find that dust strongly impacts the Lyα transmission and covering fractions of $M_{\rm UV} \lesssim -19$ galaxies in $M_{\rm vir} \gtrsim 10^{11}$ M$_\odot$ haloes, such that the dominant mode of removing Lyα photons in non-LAEs changes from low IGM transmission to high dust absorption around $z \sim 7$.

Key words: galaxies: high-redshift – cosmology: dark ages, reionization, first stars – radiative transfer – methods: numerical

1 INTRODUCTION

The prominent Lyman-alpha (Lyα) emission line of atomic hydrogen can serve as a powerful probe of the high-redshift Universe (Partridge & Peebles 1967; Ouchi et al. 2020). The intrinsic Lyα luminosity closely traces the global star-formation rate (SFR) of galaxies on relatively short timescales ($\lesssim 10$ Myr) throughout cosmic history (Kennicutt & Evans 2012; Dayal & Ferrara 2018). These Lyα photons then undergo a complex resonant scattering process coupling the emission to the neutral hydrogen and dust column densities, kinematics, and geometry, such that the emergent spectra and escape fractions can be difficult to model and interpret (Dijkstra 2019). In addition, neutral hydrogen in the intergalactic medium (IGM) far from the source can remove photons in the vicinity of the Lyα line with a single scattering out of the line of sight (Gunn & Peterson 1965). These effects are particularly strong during the Epoch of Reionization (EoR), before radiation from stars and galaxies completely ionized the hydrogen

throughout the Universe (Barkana & Loeb 2001; Wise 2019). At these high redshifts Lyα photons are typically only transmitted through the IGM if they scatter in frequency to the red damping wing of the line profile (Miralda-Escudé 1998; Madau & Rees 2000; Mesinger et al. 2015). Together these phenomena allow us to leverage Lyα emitting galaxies (LAEs) as a probe of reionization (e.g. Malhotra & Rhoads 2004; McQuinn et al. 2007; Dijkstra 2014; Kakiichi et al. 2016).

Surveys of high-redshift LAEs have made significant progress over the past decades, overcoming challenging conditions for observing the Lyα Universe (for a review see Ouchi et al. 2020). Improved strategies and technologies have helped to reach greater sensitivities, cover larger areas, or mitigate atmospheric and foreground contamination (e.g. Rhoads et al. 2000; Taniguchi et al. 2005; Finkelstein et al. 2009). At intermediate redshifts ($z \sim 2$–6) these now include novel integral field units (IFUs) to increase the number of known LAEs and better understand their detailed spatial–spectral Lyα characteristics. For example, the MUSE Integral Field Spectrograph on the Very Large Telescope (VLT; Bacon et al. 2010, 2017) and Hobby–Eberly Telescope Dark Energy Experiment (HETDEX; Adams et al. 2011; Gebhardt et al. 2021) among others are providing complementary in-

sights about large-scale structures and circumgalactic environments. At higher redshifts we rely on narrow-band surveys such as SILVER-RUSH mapping over 2000 LAEs at $z = 5.7$ and 6.6 (corresponding to advantageous low-contamination windows), which reveal target halo masses of $M_{halo} \sim 10^{11}$ M$_\odot$ (Ouchi et al. 2018). The Lyα equivalent widths are $\gtrsim 20$ Å for a small subset of the galaxy population (per cent level) as a result of either low duty cycles of bursty star-formation or small covering fractions of advantageous viewing angles. Spectroscopic searches well into the EoR at $z \gtrsim 7$ are impeded by low number statistics and so far mainly indicate that Lyα transmission is aided by very large ionized bubbles around UV bright galaxies (Jung et al. 2020, 2022; Endsley et al. 2021; Endsley & Stark 2022). We anticipate that data from the recently launched *James Webb Space Telescope* (*JWST*) will help unlock additional progress in the study of reionization-era LAEs (Katz et al. 2019; Smith et al. 2019). Finally, in addition to observing individual galaxies, future line intensity mapping of integrated diffuse light including Lyα and 21 cm emission will provide further definitive answers to questions about the EoR (Furlanetto & Lidz 2007; Heneka et al. 2017; Hutter et al. 2017).

There is strong motivation to better understand Lyα emission and transmission at $z \gtrsim 5$ from the perspectives of both galaxy formation (astrophysics) and reionization (cosmology). Constraints on the neutral fraction of the IGM can be derived from the evolution of LAEs and the number density of Lyman break galaxies (Ota et al. 2008; Ono et al. 2012; Jensen et al. 2013; Pentericci et al. 2014; Tilvi et al. 2014; Choudhury et al. 2015; Mesinger et al. 2015; Sarkar & Samui 2019), from the angular correlation function of LAEs (Sobacchi & Mesinger 2015), from the detection of Lyα emission in Lyman break galaxies (Mason et al. 2018a, 2019; Hoag et al. 2019; Jung et al. 2020), from a combination of Lyα luminosity, clustering and line profile (Ouchi et al. 2010; Goto et al. 2021), from the void probability function (Perez et al. 2022), and from the Lyα visibility (Dijkstra et al. 2011). There are also significant efforts to robustly measure and characterize the observed Lyα luminosity function (LF) at $z \approx$ 5–7 (Ouchi et al. 2008, 2010; Santos et al. 2016; Konno et al. 2018; Khusanova et al. 2020; Wold et al. 2022) also focusing on the bright end (Taylor et al. 2020, 2021). Interpreting the LF requires understanding both Lyα sources and transmission, which are correlated through the local environment and emission properties of galaxies (Mason et al. 2018a). However, the number density of LAEs is also connected to the more accessible ultraviolet luminosity function (UVLF) through the Lyα luminosity probability distribution for Lyman break galaxies, which allows some understanding of the relative evolution of galaxy and IGM effects (Dijkstra & Wyithe 2012; Gronke et al. 2015; Whitler et al. 2020; Morales et al. 2021).

Hydrodynamical simulations have proven to be a powerful framework for modeling the strong impact of frequency-dependent IGM transmission on photons in the vicinity of the Lyα line. Previous work has focused on the statistical nature of transmission curves either for post-reionization ($z \lesssim 5$) spectral modulations (Laursen et al. 2011; Byrohl & Gronke 2020) or predictions from large-volume reionization boxes (Gronke et al. 2021; Park et al. 2021; Smith et al. 2022a). Previous works have also explored the integrated Lyα transmission in the context of idealized galaxy spectral profiles (Jensen et al. 2013; Weinberger et al. 2019; Gangolli et al. 2021). These parameterized models can in principle be tuned to yield Lyα luminosities and equivalent widths that match observations after transmission through the IGM. Such studies based on semi-empirical modelling provide important predictions about the EoR and reveal the effective behaviour of these systems. However, they also gloss over specific details of Lyα radiative transfer, including uncertainties related to the emission and escape of radiation through the interstellar medium (ISM) and cir-

cumgalactic medium (CGM) of galaxies. Recently, simulations with higher resolution have been used to directly model the effects of dust absorption and predict equivalent widths, thus enabling a more self-consistent analysis of LAE populations (e.g. Garel et al. 2021). Still, the relatively small box size necessary for the high resolution also leads to difficulty in modeling large-scale effects and the bright end of LFs. Therefore, the empirical and direct transport approaches remain complementary.

In this paper we present a follow-up study of Smith et al. (2022a) in which we compare Lyα emission and transmission statistics from all runs from the THESAN suite of reionization simulations (Kannan et al. 2022b; Garaldi et al. 2022; Smith et al. 2022a). This allows us to provide additional context for reionization constraints including dependence on frequency, reionization history, and UV brightness. More importantly, we also perform an empirical machine learning study to determine what dust escape fractions and spectrally-averaged IGM transmission properties agree with the observed Lyα LFs. This is intended as a viable alternative to performing sub-grid modelling dependent Lyα radiative transfer calculations, thus acting as a bridge between the trusted but disconnected Lyα intrinsic emission and IGM transmission catalogues. The resulting simulation-based observed Lyα galaxy catalogues will also be publicly available as part of the THESAN data release. In the future we plan to apply this framework to study other Lyα-centric probes of reionization.

The remainder of the paper is organized as follows. In Section 2, we briefly describe the THESAN simulations employed throughout this paper including the Lyα catalogues. In Section 3, we present transmission statistics exploring the non-trivial dependence on frequency, redshift, and UV magnitude. In Section 4, we constrain an empirical model to match observed Lyα LFs and present dust and spectra reprocessed observed transmission statistics. Finally, in Section 5, we provide a summary and discussion on Lyα science from THESAN.

2 METHODS

In this section, we describe the THESAN simulations used in this work and the calculation of Lyα emission and transmission curves.

2.1 THESAN simulations

The THESAN project is a suite of large-volume radiation-magneto-hydrodynamical simulations that self-consistently model the reionization process and the resolved properties of the sources responsible for it. The simulations were performed with AREPO-RT (Kannan et al. 2019), a radiation-hydrodynamic extension of the moving mesh code AREPO (Springel 2010; Weinberger et al. 2020), which solves the fluid equations on an unstructured Voronoi mesh that is allowed to move along with the fluid flow for an accurate quasi-Lagrangian treatment of cosmological gas flows. Gravitational forces are calculated using a hybrid Tree-PM approach, splitting the force into a short-range force that is computed using an oct-tree algorithm (Barnes & Hut 1986) and a long-range force, estimated using a particle mesh approach.

Radiation fields are modelled using a moment based approach that solves the zeroth and first moments of the radiative transfer equation (Rybicki & Lightman 1986), coupled with the M1 closure relation, that approximates the Eddington tensor based on the local properties of each cell (Levermore 1984). They are coupled to the gas via a non-equilibrium thermo-chemistry module, which self-consistently calculates the ionization states and cooling rates from hydrogen and helium, while also including equilibrium metal cooling and Compton cooling of the CMB (see Section 3.2.1 of Kannan et al. 2019, for
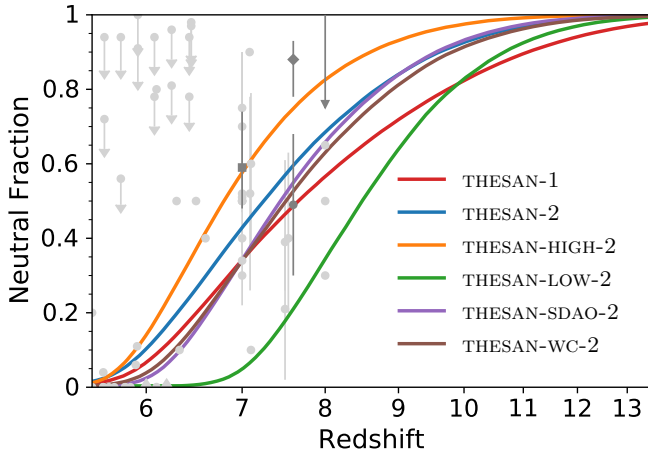
Figure 1. Reionization history of each simulation plotted as neutral fraction against redshift. THESAN-HIGH-2 and THESAN-LOW-2 bracket the other simulations, with THESAN-LOW-2 reionizing earlier than the others due to the greater contribution of low-mass galaxies to reionization, whereas in THESAN-HIGH-2 the rarer high-mass galaxies dominate. THESAN-2, THESAN-WC-2, and THESAN-SDAO-2 have similar reionization histories, though as expected THESAN-WC-2 is closest to THESAN-1. For reference, we also include observational constraints from the detection of Lyα emission in Lyman break selected galaxies (Mason et al. 2018b – square; Mason et al. 2019 – triangle; Hoag et al. 2019 – diamond; Jung et al. 2020 – circle). Constraints on the reionization history from other methods are included in light grey.

more details). Both stars and AGN act as sources of radiation, with the spectral energy distribution of stars taken from the Binary Population and Spectral Synthesis models (BPASS version 2.2.1; Eldridge et al. 2017), assuming a Chabrier IMF (Chabrier 2003). The AGN radiation output is scaled linearly with the mass accretion rate with a radiation conversion efficiency of 0.2 (Weinberger et al. 2018) and a Lusso et al. (2015) parametrization for the shape of its spectrum. For computational efficiency, we choose to only model the ionizing part of the radiation spectrum and discretize photons into energy bins defined by the following thresholds: $[13.6, 24.6, 54.4, \infty)$ eV. Each resolution element tracks the comoving photon number density and flux for each bin. Finally, we employ a reduced speed of light approximation with a value of $\tilde{c} = 0.2\,c$ (Gnedin 2016), which is large enough to accurately capture the propagation of ionization fronts and post-reionization gas properties.

The prescriptions for processes happening below the resolution limit of the simulations, such as star and black hole formation and feedback and metal production and enrichment, are taken from the IllustrisTNG model (Weinberger et al. 2017; Springel et al. 2018; Pillepich et al. 2018a,b; Naiman et al. 2018; Marinacci et al. 2018; Nelson et al. 2018, 2019; Pillepich et al. 2019). The model is augmented with a scalar dust model that tracks the production, growth, and destruction of dust using the formalism outlined in McKinnon et al. (2017). An additional birth cloud escape fraction parameter, $f_{\rm esc}^{\rm cloud} = 0.37$, is added to mimic the absorption of LyC photons happening below the resolution scale of the simulation. The parameter is tuned such that the simulation reproduces a realistic late-reionization history (Kannan et al. 2022b), which matches the observed neutral fraction evolution in the Universe (Greig et al. 2017).

All THESAN simulations follow the evolution of a cubic patch of the universe with linear comoving size $L_{\rm box} = 95.5$ cMpc, and utilize variance-suppressed initial conditions (Angulo & Pontzen 2016). We employ a Planck Collaboration et al.

(2016) cosmology (more precisely, the one obtained from their `TT,TE,EE+lowP+lensing+BAO+JLA+H`$_0$ dataset), i.e. $H_0 = 100\,h\,{\rm km\,s^{-1}Mpc^{-1}}$ with $h = 0.6774$, $\Omega_{\rm m} = 0.3089$, $\Omega_\Lambda = 0.6911$, $\Omega_{\rm b} = 0.0486$, $\sigma_8 = 0.8159$, and $n_s = 0.9667$, where all symbols have their usual meanings. The highest resolution THESAN-1 simulation has a total number of dark matter and (initial) gas particles of 2100^3 each with mass resolutions of $m_{\rm DM} = 3.12 \times 10^6\,{\rm M}_\odot$ and $m_{\rm gas} = 5.82 \times 10^5\,{\rm M}_\odot$, respectively. The gravitational forces are softened on scales of 2.2 ckpc with the smallest cell sizes reaching 10 pc. This allows us to model atomic cooling haloes throughout the entire simulation volume. (See Smith et al. (2022a) for additional discussion showing that THESAN has sufficient resolution to accurately model Lyα intrinsic emission and IGM transmission, but that there would be significant uncertainties in the emergent spectra resulting from Monte Carlo Lyα radiative transfer calculations.)

THESAN also includes a suite of medium resolution simulations with the same initial conditions aiming to investigate the impacts on reionization caused by different physics as described in Kannan et al. (2022b). The THESAN-2 simulation is the same as THESAN-1 but with two (eight) times lower spatial (mass) resolution. The THESAN-WC-2 weak convergence simulation slightly increases the birth cloud escape fraction to compensate for lower star formation in the medium resolution runs. The THESAN-SDAO-2 assumes an alternative dark matter model that includes couplings to relativistic particles giving rise to strong Dark Acoustic Oscillations (sDAOs) cutting off the linear matter power spectrum at small scales. The THESAN-HIGH-2 (THESAN-LOW-2) simulations use a halo-mass-dependent escape fraction, with only haloes above (below) $10^{10}\,{\rm M}_\odot$ contributing to reionization. We note that to match the observed neutral hydrogen fraction, the different THESAN runs adopt different birth cloud escape fractions $f_{\rm esc}^{\rm cloud}$: THESAN-1 and THESAN-2 have 0.37, THESAN-WC-2 has 0.43, THESAN-HIGH-2 has 0.8, THESAN-LOW-2 has 0.95, and THESAN-SDAO-2 has 0.55. The reionization histories of each simulation are plotted in Fig. 1, where THESAN-LOW-2 reionizes faster due to the greater contribution of low-mass galaxies, with the opposite being true of THESAN-HIGH-2. THESAN-WC-2 is also closer than THESAN-2 to the reionization history of THESAN-1. The THESAN simulations have been used to study 21 cm power spectra (Kannan et al. 2022b) including an effective bias expansion in redshift space (Qin et al. 2022), IGM–galaxy connections including ionizing mean free path and Lyα transmission statistics (Garaldi et al. 2022), multitracer line intensity mapping (Kannan et al. 2022c), and galaxy ionizing escape fractions (Yeh et al. 2023).

2.2 Lyα emission and transmission

We summarize the methodology described in Smith et al. (2022a) for Lyα emission and IGM transmission calculations, noting that we now include all THESAN simulations. For each snapshot, we produce a Lyα catalogue directly mirroring the friends-of-friends (FoF) halo and SUBFIND subhalo catalogues. This provides supplemental data for dark matter haloes and gravitationally bound galaxies, as well as high level information such as the global emissivity. For each galaxy we track the total Lyα luminosity L_α [erg s^{-1}] including contributions from resolved recombination $L_\alpha^{\rm rec}$, collisional excitation $L_\alpha^{\rm col}$, and unresolved H II regions $L_\alpha^{\rm stars}$, stellar continuum spectral luminosities $L_{\lambda,{\rm cont}}$ [erg s^{-1}Å^{-1}] at $\lambda_{\rm cont} = \{1216, 1500, 2500\}$ Å, ionizing luminosity from active galactic nuclei $L_{\rm ion}^{\rm AGN}$ [erg s^{-1}], and centre of Lyα luminosity position $\boldsymbol{r}_\alpha$ [kpc], peculiar velocity $\boldsymbol{v}_\alpha$ [km s^{-1}], and 1D velocity dispersion σ_α [km s^{-1}]. The details for calculating the total intrinsic emission due to recombinations, collisional excita-
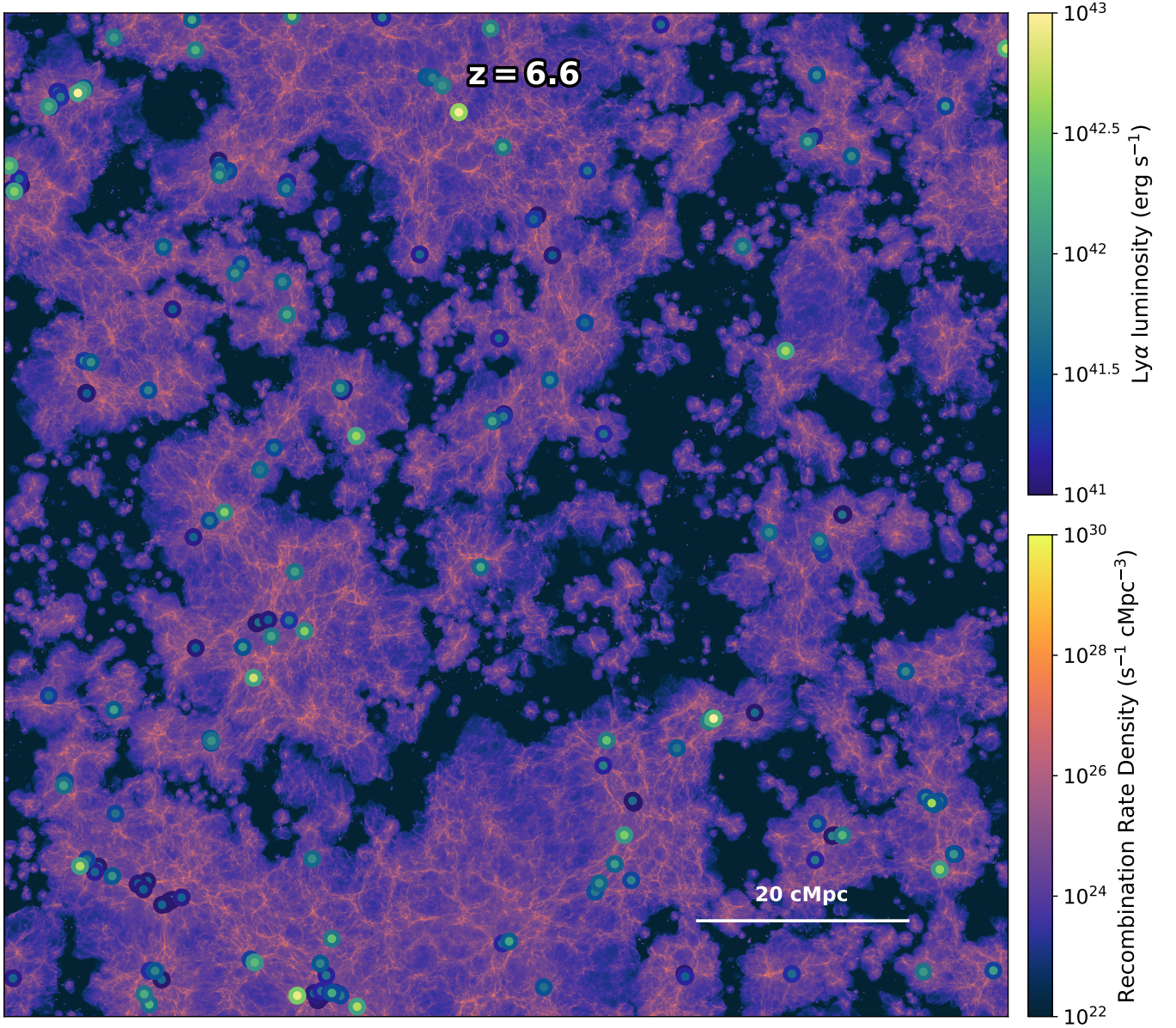
Figure 2. Recombination rate density in ᴛʜᴇꜱᴀɴ-1 at $z = 6.6$ covering the middle 2.5% of the box volume. The locations of galaxies with intrinsic Ly α luminosity greater than 10^{41} erg s^{-1} are plotted over the background, with the centre (edge) colour indicating the intrinsic (observed) luminosity as calculated in section 4. Galaxies with high Ly α emission are clearly correlated with the overall large-scale structure of the simulation.

tion, and local stars are described in Smith et al. (2022a) (see their equations 1–3), which we summarize in a single equation as:

$$\frac{L_\alpha}{h\nu_\alpha} = \int \left[P_{\rm B}\alpha_{\rm B}n_p + q_{\rm col}n_{\rm H\,I} \right] n_e\,{\rm d}V + 0.68(1 - f_{\rm esc}^{\rm cloud})\dot{N}_{\rm ion}\,. \quad (1)$$

Here $h\nu_\alpha = 10.2$ eV, $P_{\rm B}$ is the Ly α conversion probability per recombination event, $\alpha_{\rm B}$ is the case B recombination coefficient, $q_{\rm col}$ is the collisional rate coefficient, and n_p, $n_{\rm H\,I}$, and n_e are number densities for protons, neutral hydrogen, and electrons. Finally, 0.68 is the fiducial conversion probability, $f_{\rm esc}^{\rm cloud}$ denotes the escape fraction of ionizing photons calibrated for each simulation to match reionization history constraints, and $\dot{N}_{\rm ion}$ is the age and metallicity dependent emission rate of ionizing photons from stars taken from the BPASS models (v2.2.1). Combining resolved and sub-resolution emission mechanisms is necessary given that $f_{\rm esc}^{\rm cloud} < 1$. Moreover, this ap-

proach allows us to utilize the on-the-fly ionization field while also reproducing the expected ionizing photon budget.

We also produce catalogues for the frequency-dependent transmission of Ly α photons through the IGM for all simulations at $z = \{5.5, 6, 6.6, 7, 8, 9, 10, 11, 13\}$ (similar to Laursen et al. 2011; Byrohl & Gronke 2020; Gronke et al. 2021; Garel et al. 2021; Park et al. 2021). As in Smith et al. (2022a), for each central halo selected with at least 32 star particles (and less resolved ones if they are among the 90% most massive centrals) we extract 768 radially outward rays corresponding to equal area healpix directions of the unit sphere. We use the COsmic Ly α Transfer code (ᴄᴏʟᴛ; Smith et al. 2015, 2019, 2022b) to perform exact ray-tracing through the native Voronoi unstructured mesh data. We start the rays at initial distances of $R_{\rm vir}$, defined as the radius within which
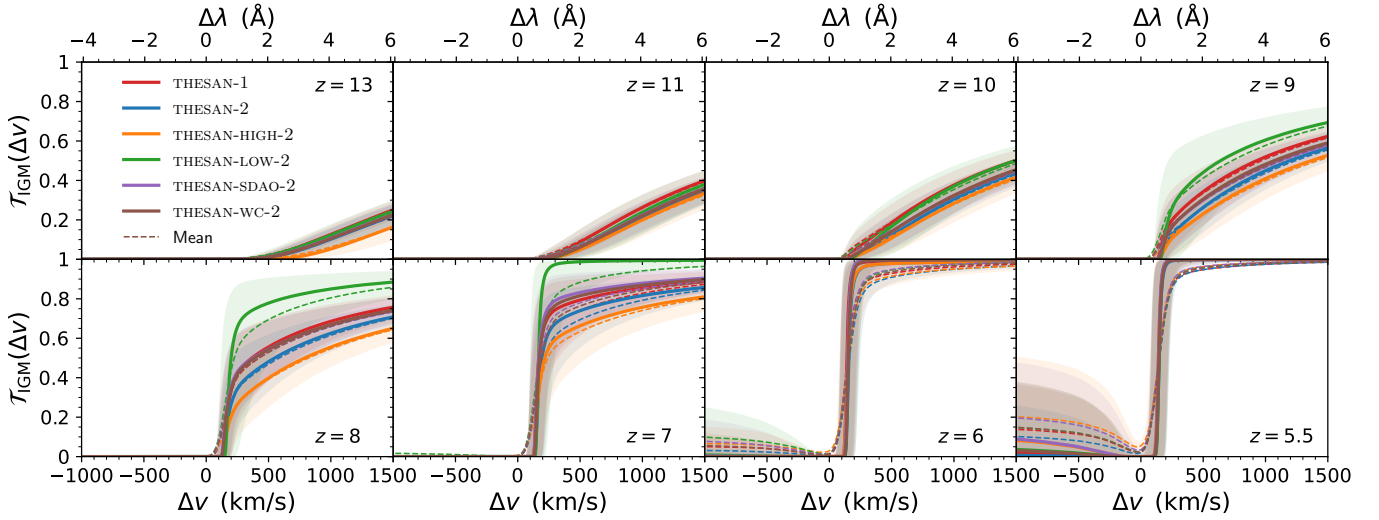
Figure 3. IGM transmission $\mathcal{T}_{\mathrm{IGM}}$ as a function of velocity offset Δv and rest-frame wavelength offset $\Delta\lambda$ around the Lyα line for each simulation. Each panel shows a different redshift over $z \in [5.5, 13]$. The solid (dashed) curves show the catalogue median (mean) statistics and shaded regions give the 1σ confidence levels. With equal weight to all galaxies and sightlines, this view is biased towards more numerous lower-mass haloes. From redshifts $z \approx 7$–9, the various simulations produce higher or lower transmission curves as a result of the different reionization histories and bubble morphologies. The shapes of coeval curves are also affected, e.g. with steeper or flatter slopes around $\Delta v \approx 500\,\mathrm{km\,s^{-1}}$. In order of decreasing transmissivity these are THESAN-LOW-2, THESAN-SDAO-2, THESAN-WC-2, THESAN-1, THESAN-2, and THESAN-HIGH-2.

the mean density becomes 200 times the cosmic value (R_{200}). We take the systemic location and rest-frame from the subhalo intrinsic Lyα luminosity averaged position $\boldsymbol{r}_\alpha$ and velocity $\boldsymbol{v}_\alpha$. We select a broad wavelength range of $\Delta v \in [-2000, 2000]\,\mathrm{km\,s^{-1}}$ sampled at a high spectral resolution of $5\,\mathrm{km\,s^{-1}}$ or a resolving power of $R \approx 60\,000$. We perform the integrations out to a distance of $4000\,\mathrm{km\,s^{-1}}/H(z) \approx 40\,\mathrm{cMpc}\,[(1+z)/7]^{-1/2}$. We calculate the traversed optical depth based on the continuous Doppler shifting scheme described in Smith et al. (2022a), which incorporates velocity gradients encountered during propagation. The total optical depth τ defines the frequency-dependent transmission function for each ray:

$$\mathcal{T}_{\mathrm{IGM}}(\Delta v) \equiv \exp\left[-\tau(\Delta v)\right], \qquad (2)$$

which describes the fraction of flux not attenuated by the IGM after escaping the halo. Similar to Park et al. (2021), we incorporate the distant IGM beyond the local rays in a statistical sense based on the global reionization history. This additional damping-wing absorption τ_{DW} has a very minor impact after the EoR but is increasingly important at higher redshifts (Miralda-Escudé 1998). Finally, we adopt the complete first-order quantum-mechanical correction to the Voigt profile presented by Lee (2013), which strengthens the red wing due to positive interference of scattering from all other levels. Detailed calculations for τ_{DW} and the Voigt profile correction are given in Smith et al. (2022a).

While THESAN is fully capable of capturing IGM scale effects, detailed LAE modelling is strongly affected by ISM scale sourcing and radiation transport through the CGM. The uncertainties mainly arise from the galaxy formation model (and limited resolution), which includes temporary decoupling of wind particles from the hydrodynamics and the use of an effective equation of state (EoS) for cold gas above the density threshold $n_{\mathrm{H}} \approx 0.13\,\mathrm{cm}^{-3}$ (Springel & Hernquist 2003). In this study, we choose to calibrate an empirical model for the spectral line profile emerging from galaxies and absorption of Lyα photons by dust (see Sec. 4.1). We emphasize that due to dust and IGM reprocessing the bright end of the observed Lyα LF differs

from the intrinsic one by up to two orders of magnitude. Thus, our approach informs us about what is required to match observations and provides observed LAE luminosity catalogues from the THESAN simulations for follow-up comparison studies. In Fig. 2, we show a preview visualization of the LAE catalogue along with the resolved recombination rate density (for details see Section 4.3).

3 TRANSMISSION STATISTICS

The strong absorption of photons near the Lyα line provides a powerful probe of the structure and evolution of reionization. In this section we explore the isolated impact of IGM transmission in THESAN. We note that we apply a UV continuum selection cut of $M_{1500} < -19$ when aggregating haloes for summary statistics to avoid resolution biases in comparisons of different simulations (see Figs. 3–6). Due to the low number of bright galaxies we relax this restriction to $M_{1500} < \{-18.5, -17.5, -16.5, -15.5\}$ for $z = \{9, 10, 11, 13\}$, respectively.

3.1 Comparing transmission curves

We first consider the frequency dependence and redshift evolution of global catalogue statistics. In Fig. 3 we show the IGM transmission fraction $\mathcal{T}_{\mathrm{IGM}}$, including the local and damping-wing contributions as a function of velocity offset Δv and rest-frame wavelength offset $\Delta\lambda$ around the Lyα line for each simulation. Each panel shows the transmission at a different redshift across the range $z \in [5.5, 13]$. The solid (dashed) curves show the full catalogue median (mean) statistics and shaded regions give the 1σ confidence levels. With equal weight to all galaxies and sightlines, this view is biased towards lower masses that dominate the catalogues by halo count. As expected, transmission redward of line centre increases as redshift decreases due to the decreasing neutral fraction, while the blue peak is highly suppressed until the end of the simulation. In Appendix A,
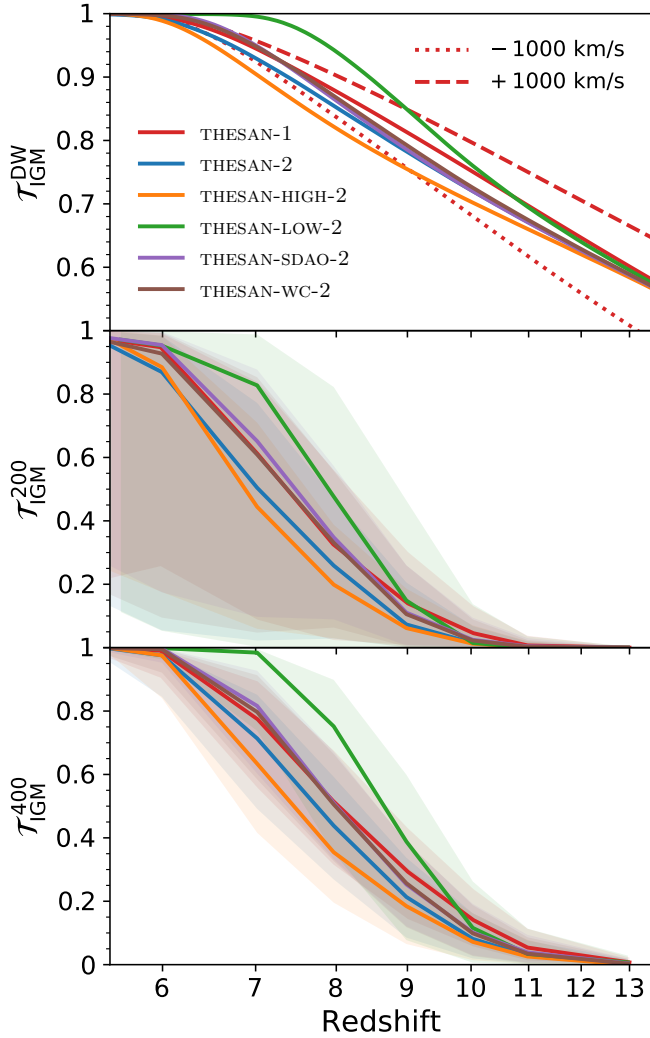
Figure 4. *Top:* Damping-wing IGM transmission $\mathcal{T}_{IGM}^{DW} = \exp(-\tau_{DW})$ at $\Delta v = 0$ for all simulations as a function of redshift. For THESAN-1 we also plot $\mathcal{T}_{IGM}^{DW}$ at $\Delta v = \pm 1000\,\mathrm{km\,s^{-1}}$ to indicate the range of frequency dependence. *Middle and bottom:* Overall transmission averaged over $50\,\mathrm{km\,s^{-1}}$ spectral windows at velocity offsets of $\Delta v = 200$ and $400\,\mathrm{km\,s^{-1}}$ plotted over redshift, which are qualitatively similar, but $\mathcal{T}_{IGM}^{400}$ has slightly higher transmission and less variation. The effects of different reionization histories can be seen again with THESAN-LOW-2 and THESAN-HIGH-2 bracketing the other simulations.

Figure 5. The same quantities as in Fig. 4, i.e. the IGM transmission from the damping wing $\mathcal{T}_{IGM}^{DW}$ and at velocity offsets of $\Delta v = 200$ and $400\,\mathrm{km\,s^{-1}}$ (from top to bottom), but as functions of the global neutral fraction. In this view the transmission order for THESAN-LOW-2 and THESAN-LOW-2 have switched places, illustrating the importance of bubble morphology in addition to the reionization history.

at all redshifts, as a consequence of having the latest reionization history and mainly providing a local transmissivity boost to rarer, higher-mass galaxies.

3.2 Redshift evolution

To highlight the rapid change in transmissivity throughout the EoR, we focus on the behaviour in wavelength bands relevant for Lyα related science. Specifically, we define the integrated or mean transmission as $\mathcal{T}_{IGM}^{int} \equiv \int \mathcal{T}_{IGM}\,\mathrm{d}\Delta v / \int \mathrm{d}\Delta v$, focusing on velocity offset windows at 200 and $400\,\mathrm{km\,s^{-1}}$ averaged over widths of $50\,\mathrm{km\,s^{-1}}$ to match typical spectroscopic instrument capabilities. These locations are chosen to roughly correspond to the left and right sides of observed red peaks at low and high redshifts (e.g. Ouchi et al. 2020). In Fig. 4 we plot the damping wing transmission at line center as a function of redshift in the top panel, and the transmission at $\Delta v = 200$ and $400\,\mathrm{km\,s^{-1}}$ with redshift in the bottom two panels. In the top panel, we also plot damping wing

we explore the impact of IGM resolution on the frequency-dependent transmission curves and conclude that while there are strong differences near line centre, the difference in the red peak transmission ($\Delta v \gtrsim 200\,\mathrm{km\,s^{-1}}$) is relatively minor.

From redshifts $z \approx 7$–9, the various simulations can produce significantly higher or lower transmission curves as a result of the different reionization histories and bubble morphologies with THESAN-LOW-2 and THESAN-HIGH-2 bracketing the range. The shapes of coeval transmission curves are also affected; for example, THESAN-LOW-2 looks more like a step function, and around a velocity offset of $\Delta v \approx 500\,\mathrm{km\,s^{-1}}$ the other simulations have steeper slopes. THESAN-LOW-2 has the earliest global reionization history but also the greatest advantage for local transmission around the abundant low-mass galaxies, as these dominate reionization in this scenario. In contrast, THESAN-HIGH-2 has the lowest degree of transmission
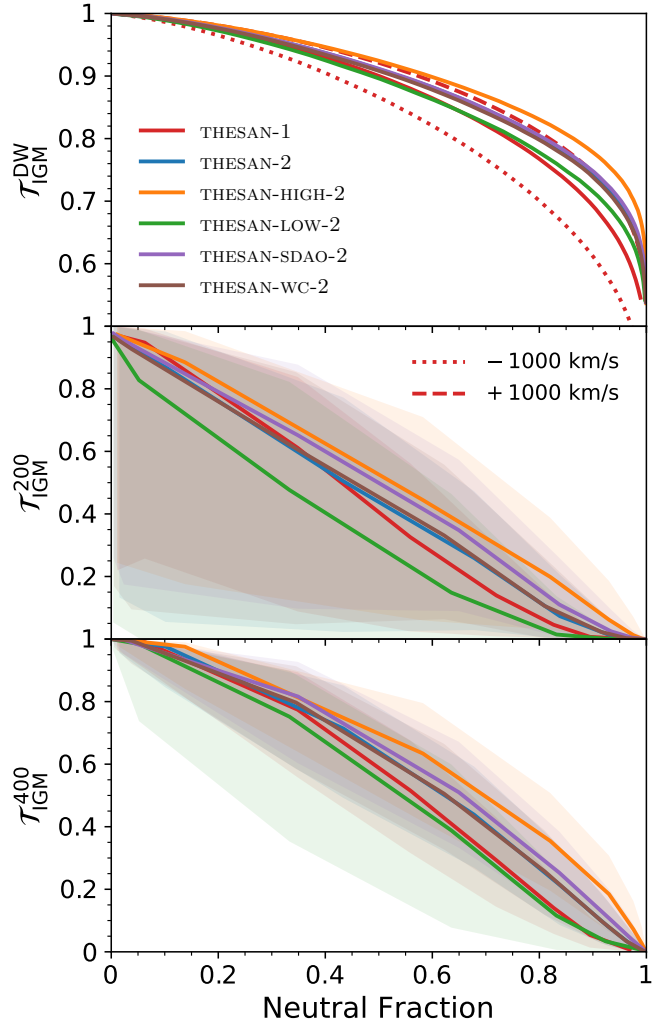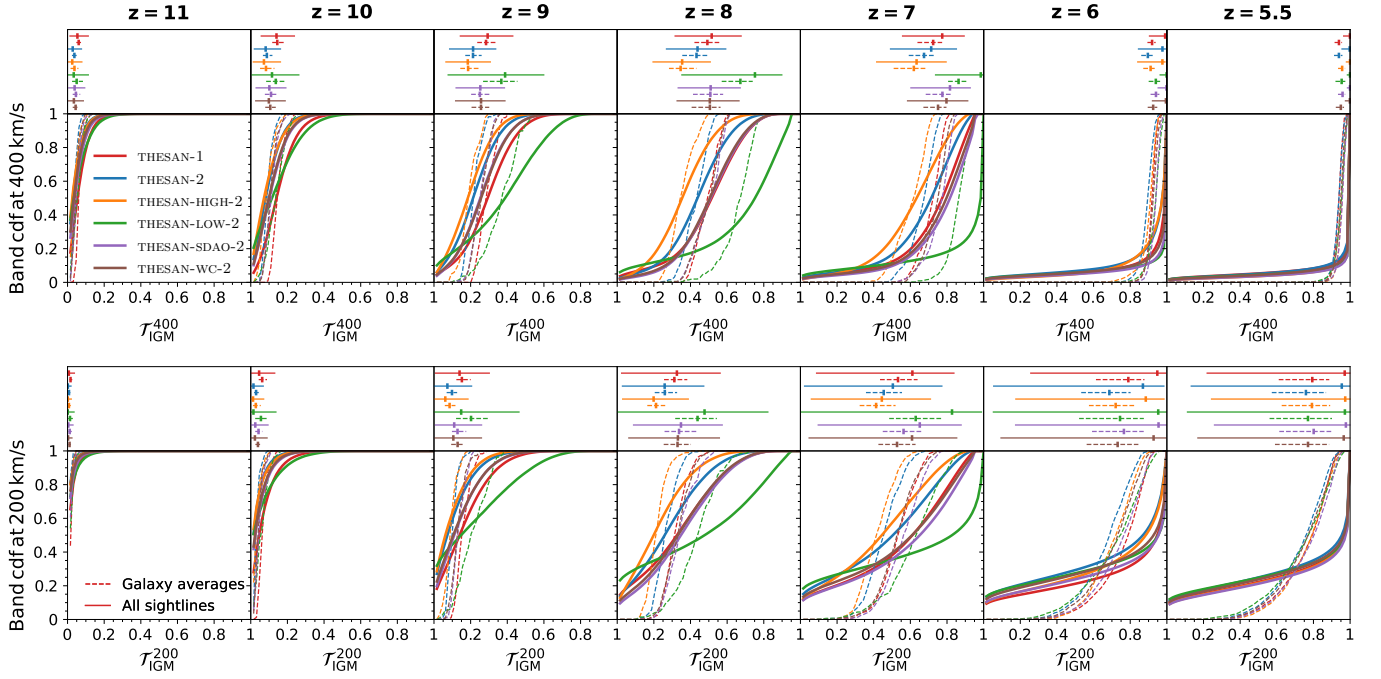
Figure 6. Cumulative probability distribution functions (CDF) for transmission at 200 and 400 km s^{-1} for each simulation at several redshifts. Median and 1σ summary statistics are plotted in the upper panels. We also include each CDF of galaxy averages plotted as dashed curves, which result in slightly lower medians and have narrower distributions. Thus, the variation across sightlines from a single LAE is larger than the variation between different galaxies.

transmission at $\Delta v = \pm 1000$ km s^{-1} for THESAN-1 to demonstrate the range of dependence on frequency; there is higher transmission for red velocity offsets, and vice versa. At $\Delta v = 400$ km s^{-1}, there is less variation within and among the simulations than at 200 km s^{-1}. We interpret this as a difference between probing the distant absorption across a more homogeneous large-scale IGM and the nearby imprint of the stochastic local environment. The damping wing transmission in THESAN-1 is generally higher than the other simulations, with the exception of THESAN-LOW-2, due to its higher resolution that allows more low-mass galaxies to contribute ionizing photons and influence the early reionization history. However, the overall transmission in THESAN-1 typically lies between the THESAN-2 and THESAN-WC-2 simulations. We note that the 8 times higher resolution of THESAN-1 is expected to induce systematically lower transmission because the covering fraction of Lyman-limit systems increases (van de Voort et al. 2019).

The effects of the different reionization histories can be seen again, with THESAN-LOW-2 bracketing the simulations from above and THESAN-HIGH-2 from below in all panels. To disentangle the effects of differing reionization histories, in Fig. 5 we also plot the same values against the global volume-weighted neutral fraction. In this perspective the transmission fractions essentially increase linearly with decreasing neutral fraction, while damping wing transmission increases faster at higher neutral fractions before leveling off. It is also interesting that THESAN-LOW-2 (THESAN-HIGH-2) now has lower (higher) transmission than the other simulations, switching places and highlighting the importance of bubble morphology in addition to the reionization history; THESAN-LOW-2 (THESAN-HIGH-2) likely allows less (more) transmission at the same ionization fraction due to its smaller (larger) bubble sizes. In addition, THESAN-SDAO-2 and THESAN-WC-2 clearly have higher transmission than THESAN-1 at the same neutral fraction, especially at larger neutral fractions. This is

also due to smaller bubble sizes in THESAN-1 at higher redshifts during first formation, enabled by its higher resolution (see section 3.4 of Kannan et al. 2022b, and Neyer et al. in prep.). Otherwise, THESAN-SDAO-2 and THESAN-WC-2 trend closely to THESAN-1 in other statistics. We note that the damping wing contribution has less variation between simulations due to the assumption of a homogeneous cosmological neutral hydrogen density, removing information about morphology but retaining an imprint of the timing of reionization. As before, both Fig. 4 and Fig. 5 are biased towards lower mass haloes.

3.3 Transmission distributions

To further investigate the different distributions of transmission between simulations, in Fig. 6 we plot the cumulative probability distributions for transmission at $\Delta v = 200$ and 400 km s^{-1} for each simulation at several redshifts. For convenience in comparing simulations, the median values and 1σ summary statistics are plotted in the upper panels. This view is again biased towards lower masses. The solid curves indicate the distribution for all sightlines across all galaxies, while for the dashed lines, we first average the transmission for each galaxy over its 768 sightlines before plotting the distribution for all galaxies. Generally, the galaxy-averaged transmission is slightly lower and has a more narrow distribution, especially at lower redshifts. This implies that the variation across sightlines from a single LAE is larger than the variation between different galaxies. This arises from the exponential sensitivity on optical depth ($e^{-\tau}$) allowing the same galaxy to have both large and small values; i.e. the broad and bimodal sightline distributions are not due to different large-scale IGM environments. There will always be sightlines with very low transmission due to filaments and other self-shielding structures common at high-z (Park et al. 2021; healpix plots from Smith et al. 2022a).
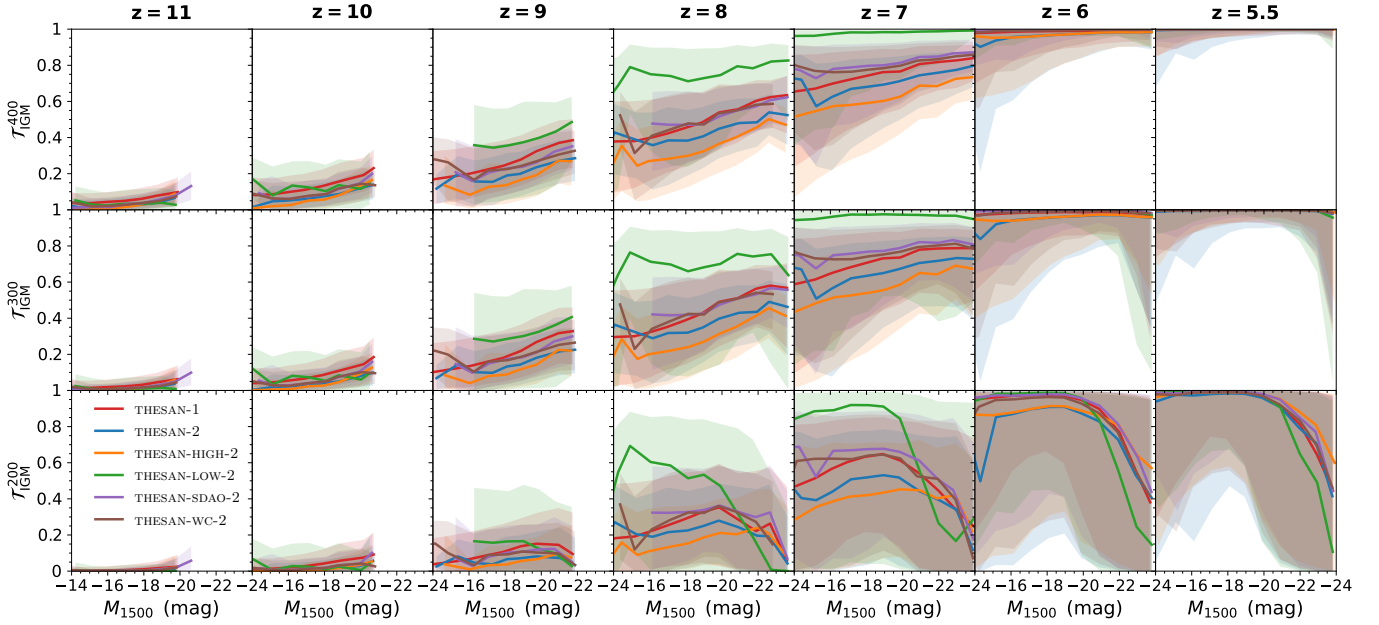
Figure 7. Band integrated IGM transmission median and 1σ statistics as a function of UV magnitude M_{1500} for each redshift. We find clear trends of higher transmission for brighter galaxies, especially for $\Delta v \gtrsim 300\,\mathrm{km\,s^{-1}}$ at $z \gtrsim 6$. Another striking feature is the strong absorption for bright galaxies ($M_{1500} \lesssim -20$) due to infalling gas at $\Delta v \lesssim 200\,\mathrm{km\,s^{-1}}$.

The distributions at $\Delta v = 200\,\mathrm{km\,s^{-1}}$ are much broader than those at $400\,\mathrm{km\,s^{-1}}$, again indicating that it has more variation in transmission. This is both in terms of values among galaxies in the same simulation and between different simulations (recall the shaded regions in Fig. 4). There are also lower median transmission values at $200\,\mathrm{km\,s^{-1}}$ at a given redshift because the closer photons are to line center, the more susceptible they are to resonant infalling motions within the local environments of galaxies, which vary significantly.

In addition, THESAN-1, with its higher resolution, is able to resolve lower-mass galaxies, and THESAN-LOW-2 has lower-mass galaxies dominate reionization; the two simulations therefore have earlier ionized bubble formation around lower mass galaxies. This leads to the higher transmission values that we initially see at $z \gtrsim 10$. Eventually, THESAN-LOW-2 diverges as it has a much earlier completion of reionization, while THESAN-HIGH-2 lags behind the others due to its later reionization history. Overall, the similar behaviour of these statistics despite changing resolution and physics demonstrates that Lyα transmission does indeed act as a powerful probe of the EoR.

3.4 Dependence on UV magnitude

We now explore the dependence of transmission on galaxy properties to quantify environmental effects and compare similar galaxies across different simulations. In Fig. 7, we plot the median transmission for $50\,\mathrm{km\,s^{-1}}$ wavebands centred at velocity offsets of $\Delta v = \{200, 300, 400\}\,\mathrm{km\,s^{-1}}$ as functions of UV magnitude M_{1500} (without any dust correction) at several redshifts with shaded regions indicating 1σ statistics. Transmission generally increases with magnitude, with the notable exception of a downturn for the brightest galaxies ($M_{1500} \lesssim -20$) at $\Delta v = 200\,\mathrm{km\,s^{-1}}$. This effect also bleeds into the transmission at $300\,\mathrm{km\,s^{-1}}$, but disappears by $400\,\mathrm{km\,s^{-1}}$. This suppression signature is mainly due to cosmic streams of infalling gas in the crowded environments around massive galaxies, as the frequency range likely to encounter a resonance point is extended

to the circular velocity $V_c \propto M_{\mathrm{halo}}^{1/3}$ (Santos 2004; Dijkstra et al. 2007).

We again see THESAN-LOW-2 and THESAN-HIGH-2 bracketing the other simulations. It is also interesting that although THESAN-LOW-2 generally has higher transmission, its typical transmission at $200\,\mathrm{km\,s^{-1}}$ dips below the other simulations in the downturn in the brightest galaxies, while THESAN-HIGH-2 goes above the others. This is due to the biased (dis)advantage artificially setting the birth cloud ionizing escape fractions to be (zero) higher in the (THESAN-LOW-2) THESAN-HIGH-2 runs. The remaining simulation orderings are consistent with differences in the reionization histories. In practice, due to the large variances it may be difficult to disentangle timing, morphology, and modelling effects from IGM transmission alone.

3.5 Covering fractions

Covering fractions are often used to understand stochastic visibility arising from anisotropic and bimodal sightline distributions. In our context we define the covering fraction of each individual galaxy as the fraction of sightlines with transmission below 20 per cent; i.e. $P(\mathcal{T}_{\mathrm{IGM}} < 0.2)$. In Fig. 8 we show the median and 1σ range of halo covering fractions as a function of UV magnitude M_{1500} for each simulation at different redshifts and wavebands. Overall, the covering fractions decrease with decreasing redshift, as the progression of reionization determines the amount of neutral hydrogen around these haloes. We also see that fainter galaxies have larger covering fractions, corresponding to more isotropic suppression due to the correlation with ionized bubble size. We also see the influence of infall velocity absorption and local environment with an upturn for bright galaxies ($M_{1500} \lesssim -20$) at $\Delta v = 200\,\mathrm{km\,s^{-1}}$, which is largely absent by $400\,\mathrm{km\,s^{-1}}$. Finally, lower (higher) transmission fractions translate to higher (lower) covering fractions, so the reverse ordering among simulations is also largely preserved.
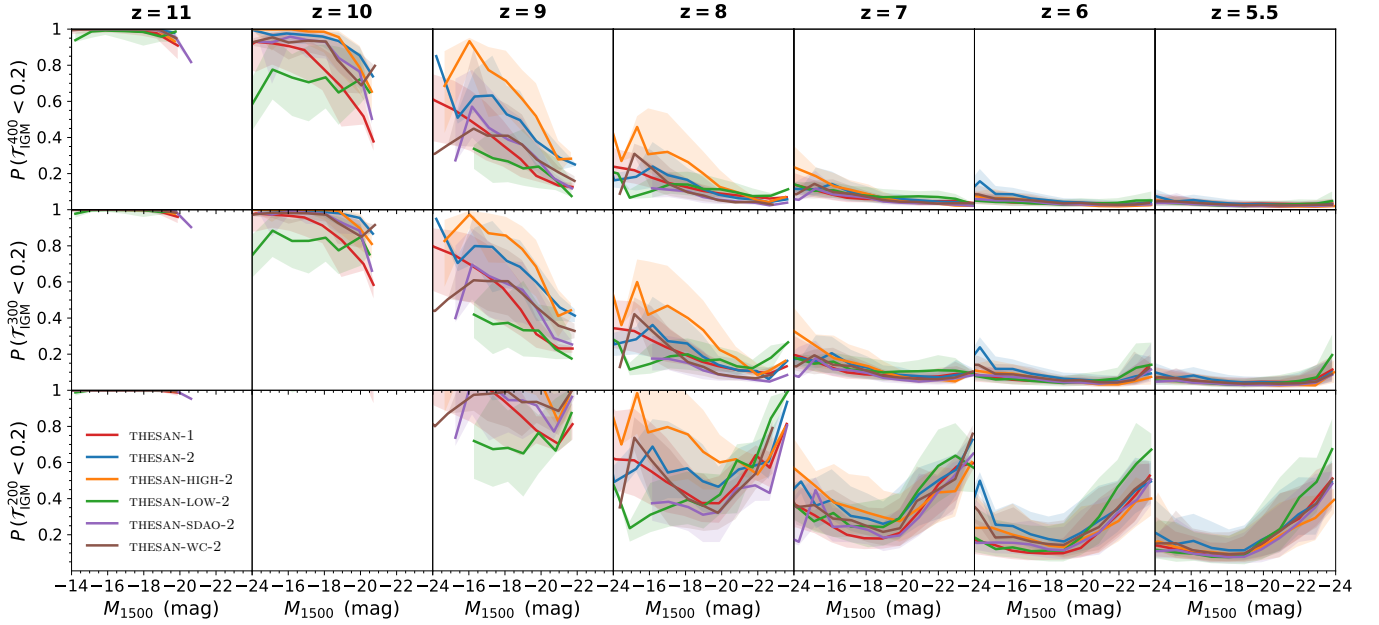
Figure 8. IGM transmission covering fractions defined as the fraction of sightlines around each galaxy with transmission below 20 per cent, i.e. $P(\mathcal{T}_{\rm IGM} < 0.2)$. The curves and shaded regions give the median and 1σ variation as a function of UV magnitude M_{1500} at velocity offsets of $\Delta v = \{200, 300, 400\}\,\mathrm{km\,s^{-1}}$ (bottom to top) at a range of redshifts (decreasing from left to right). See the text for additional discussion.

Table 1. Summary of various emergent spectra models used in previous studies. The columns from left to right are the acronym used in the text and figures, reference citation, model description, and relevant notes about the simulation or scientific application. These models are provided for comparison with our fully calibrated spectral model shown in Fig. 9, both as purely spectral models and also including our best-fit dust model. For notational compactness $G(\mu, \sigma)$ denotes a Gaussian with mean μ and standard deviation σ.

Model	Reference	Description
J13	Jensen et al. (2013)	Gaussian-minus-a-Gaussian[a]
W19	Weinberger et al. (2019)	$G(\{1.5, 1.8\}\,V_c, 88\,\mathrm{km\,s^{-1}})$[b]
G21	Gangolli et al. (2021)	$G(\{1.2, 1.4, 1.8\}\,V_c, V_c)$[c]
BG20G	Byrohl & Gronke (2020)	$G(0, 200\,\mathrm{km\,s^{-1}})$
BG20N	Byrohl & Gronke (2020)	Neufeld double-peaked profile[d]
BG20R	Byrohl & Gronke (2020)	Neufeld with red peak only

[a] The σ values depend on halo mass, calibrated to simulated lines
[b] The μ coefficient depends on their reionization model
[c] The μ coefficient depends on neutral fraction and redshift
[d] Temperature $T = 10^4$ K and H I column density $N_{\rm HI} = 10^{20}\,\mathrm{cm^{-2}}$

4 OBSERVED LUMINOSITY FUNCTIONS

Observed Lyα LFs are expected to differ significantly from intrinsic ones as a result of dust absorption and flux-averaged IGM transmission. In THESAN we can accurately model the Lyα intrinsic emission from galaxies and frequency-dependent IGM transmission, but ISM scale Lyα radiative transfer effects would be extremely challenging to model self-consistently. The empirical calibration approach presented in this section is intended as a reasonable alternative to sub-grid model dependent predictions, and provides a bridge framework to incorporate emission, dust, spectra, and transmission modelling components to match observed Lyα LFs.

4.1 Galaxy model calibration procedure

The choice of spectral model can significantly alter the predicted observed Lyα luminosities of galaxies. In Table 1, we provide a summary of previous simulation-based studies from across the literature that adopt empirical emergent Lyα spectra models. We also assign a name abbreviation to each reference and give a brief summary of relevant information. In Fig. 9 we apply these models to our THESAN-1 IGM transmission data at redshifts of 5.5 and 6.6 (we note that we do not show the model from Gangolli et al. (2021), as it overlaps with the curve from Weinberger et al. (2019)), alongside two additional curves representing our best-fit spectral and dust attenuation model at the specific redshift (dotted black) and when combining redshifts (thick dashed black) for which the calibration process is discussed in detail below. We see that none of the currently available models in the literature (coloured curves), when applied to the THESAN-1 simulation, provide an acceptable fit to the observed data (background points with error bars). This is especially problematic at the highest luminosities because none of the models accounts for dust absorption. This highlights the importance of including dust in addition to IGM transmission, which strongly suppresses more massive galaxies that are intrinsically bright but have low Lyα escape fractions. We emphasize that when we apply our dust model (detailed below) to the spectral models from the literature, the agreement with observations improves but not to a satisfactory level unless both models are simultaneously constrained.

Unfortunately, there are aspects of parameterised dust models that are degenerate with accompanying spectral models; for example, strong dust absorption will lower the observed Lyα luminosity, but so will a spectral model with very little flux in the red damping wing. To account for this, we choose to model both the dust and galaxy spectra for completeness. Fig. 10 contains a schematic representation of both our dust and spectral models, which we now describe.

For simplicity, we choose to model dust as a piecewise function of galaxy mass. Below a minimum halo mass, $M_{\rm halo}^{\rm min}$, the Lyα escape
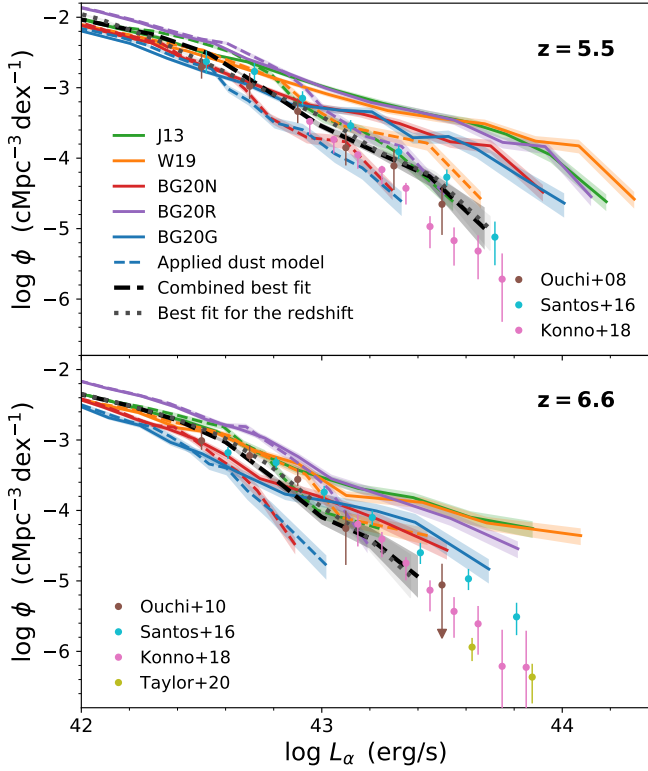
Figure 9. Comparison of observed Lyα luminosity functions at $z = 5.5$ (top panel) and $z = 6.6$ (bottom panel) for the spectral models described in Table 1, which do not include dust modelling, as applied to the THESAN-1 simulation (coloured solid curves; see the table for a mapping between identifier and model). For comparison, we show a collection of observational data in the background. The thick black dashed curve shows the best-fitting model with the same parameters used for both redshifts, with the shaded regions showing the variation in the LF across sightlines. The grey dotted curve shows the case where we only fit to the data for the specific redshift, with our calibrated models showing significantly better agreement with the observed data than the uncalibrated spectral models already available in the literature, even after also applying our dust model (coloured dashed curves). For the W19 model we chose 1.5 as the coefficient for μ.

fraction due to dust stays constant at unity, i.e. we assume negligible absorption. Above a maximum halo mass, $M_{\rm halo}^{\rm max}$, the escape fraction stays constant at a certain minimum escape fraction, $f_{\rm esc}^{\rm min}$. In the intermediate mass range, the escape fraction decreases log-linearly. The free parameters are thus the maximum and minimum halo masses, and the minimum escape fraction:

$$f_{\rm esc} = \begin{cases} 1 & M_{\rm halo} < M_{\rm halo}^{\rm min} \\ f_{\rm esc}^{\rm min} & M_{\rm halo} > M_{\rm halo}^{\rm max} \\ 1 + (f_{\rm esc}^{\rm min} - 1)\frac{(\log M_{\rm halo} - \log M_{\rm halo}^{\rm min})}{(\log M_{\rm halo}^{\rm max} - \log M_{\rm halo}^{\rm min})} & \text{otherwise} \end{cases} \quad (3)$$

Likewise, our spectral model is chosen to be a Gaussian line profile described by a constant width of $\sigma = 200\,{\rm km\,s^{-1}}$. We chose a constant line width as we found that varying the line width was strongly degenerate with changes in the modelled minimum escape fraction. Our spectral model is given by

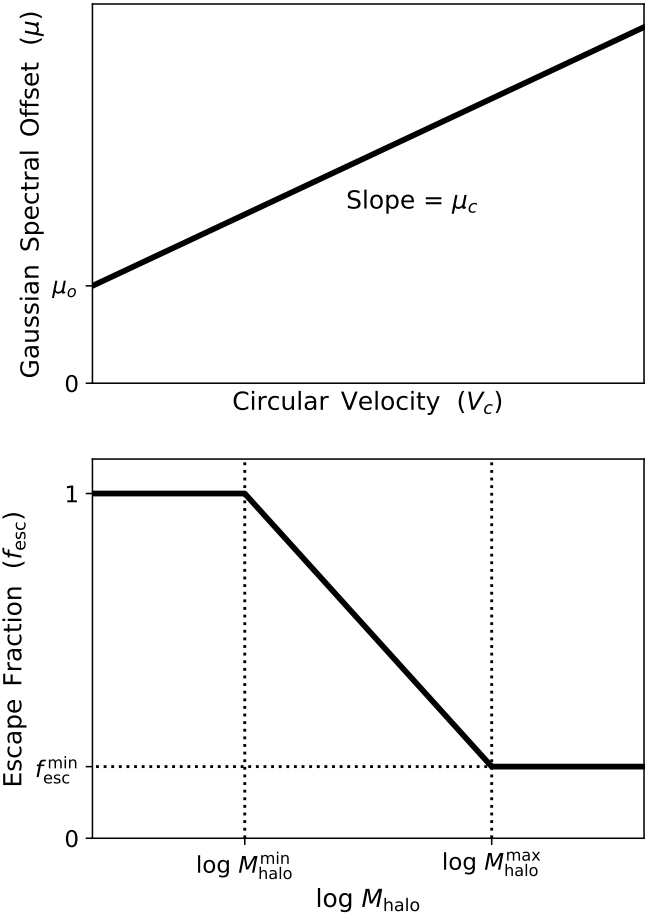$$J(\nu) \propto \exp\left(-\frac{1}{2}\frac{(\Delta\nu - \mu)^2}{\sigma^2}\right), \quad (4)$$

Figure 10. Schematic representations for the adopted model of the emergent Gaussian spectral offset μ as a function of halo circular velocity V_c (upper panel; see equation 5) and Lyα escape fraction $f_{\rm esc}$ due to dust absorption as a function of halo mass $M_{\rm halo}$ (lower panel; see equation 3).

with

$$\mu = \mu_c V_c + \mu_o, \quad (5)$$

where V_c is the circular velocity of the galaxy, while μ_c (coefficient) and μ_o (offset) are free parameters. This is consistent with observations and in line with theoretical expectations that more massive galaxies have larger velocity offsets (Yang et al. 2016; Verhamme et al. 2018). We ignore the normalization constant since it will cancel out later when calculating the fraction of transmitted luminosity.

Since we assume emergent spectra for the observed models which are uncertain, we take an empirical approach by assuming idealized Gaussian profiles, but of course there are unmodelled processes including RT effects that induce anisotropic dust absorption, Doppler shifting, and line broadening. Despite these uncertainties, it is a well-defined optimisation problem to calibrate our model to observed LFs, and the resulting ratios of observed-to-intrinsic Lyα luminosities are relevant for understanding and interpreting observations of LAEs.

To calculate this transmission ratio, the line profile from each galaxy is multiplied by the frequency-dependent IGM transmission fraction of the central galaxy in its group, if known (see 2.2 for details). Otherwise the line profiles from that group's galaxies are instead multiplied by the catalogue-median transmission curve. This is because we expect satellites to generally be fainter than the central but have similar cosmological scale IGM transmission with the

Table 2. Critical volume of parameter space for training the Gaussian Process emulator. We also require that $M_{\rm halo}^{\rm min} < M_{\rm halo}^{\rm max}$ for parameter viability. Note that the units for mass and μ_o are $M_\odot$ and $\rm km\,s^{-1}$, respectively.

Range	$\log M_{\rm halo}^{\rm min}$	$\log M_{\rm halo}^{\rm max}$	$f_{\rm esc}^{\rm min}$	$\mu_{\rm c}$	$\log \mu_{\rm o}$
[min, max]	[9.5, 11.5]	[10.5, 12]	[0.05, 0.4]	[0, 2]	[1, 2]

Table 3. Best-fit parameters for capturing smooth representations of observed Lyα luminosity functions with Schechter functions (see equation 8). Note that the units for ϕ^* and L^* are $\rm cMpc^{-3}\,dex^{-1}$ and $\rm erg\,s^{-1}$, respectively.

Redshift	$\log \phi^*$	$\log L^*$	α
5.5	−3.992	43.45	−2.534
6.6	−4.123	43.35	−2.501

exception of minor localized sightline and velocity offset effects. Therefore, we assume that satellites inherit transmission properties from the central halo and less-resolved galaxies inherit the median global transmission at each redshift. We integrate over frequency and then divide by the integral of the intrinsic profile:

$$\mathcal{T}_{\rm IGM} = \frac{\int {\rm d}\nu\, J(\nu) e^{-\tau(\nu)}}{\int {\rm d}\nu\, J(\nu)}\,. \tag{6}$$

We then multiply $\mathcal{T}_{\rm IGM}$ by the halo mass-dependent escape fraction from dust, $f_{\rm esc}(M_{\rm halo})$, which depends on the parameters of our dust model. This gives the final observed transmission fraction:

$$f_{\rm esc} \times \mathcal{T}_{\rm IGM} = L_\alpha^{\rm obs}/L_\alpha^{\rm int}\,. \tag{7}$$

From this, we calculate the observed Lyα luminosity, which gives us the observed LFs to calibrate to observations.

From our dust and spectral models, we thus have 5 parameters ($\log M_{\rm halo}^{\rm max}$, $\log M_{\rm halo}^{\rm max}$, $f_{\rm esc}^{\rm min}$, $\mu_{\rm c}$, $\log \mu_{\rm o}$) to optimize. Running a single model for all haloes is (relatively) computationally expensive, and thus standard minimisation techniques in such a high-dimensional parameter space can be costly. We employ a Gaussian Process Regression technique to build an emulator based on a suite of pre-performed simulations, allowing us to estimate the LF at various values of the free parameters in a rapid fashion. We use the SWIFTEMULATOR (Kugel & Borrow 2022), as this contains pre-written routines for emulating scaling relations like LFs.

We run the emulation at two redshifts, $z = 5.5$ and $z = 6.6$, generally following Morales et al. (2021). We note that for the $z = 5.5$ emulation, the observed data points we calibrated to were collected at $z = 5.7$, however the reionization history (global H I fraction, galaxy properties, etc.) between these two redshifts is very similar and the slight mismatch is unlikely to affect our results (see Kannan et al. 2022b). For the remainder of the paper, we refer to these observations as being at $z = 5.5$. We also tried an emulation at $z = 7$, but the available observations were not fully consistent with the LFs that THESAN predicts. Specifically, a Schechter function fit to the observed data points intersected the intrinsic LF at the faint-end.

We train the emulator with sufficient sampling within our five-dimensional parameter space, spread uniformly through the use of a Latin hypercube. The hypercube is generated with 300 realizations, from which we then reject parameter points that have $\log M_{\rm halo}^{\rm min} > \log M_{\rm halo}^{\rm max}$, an unphysical parameter combination, leaving 250 valid samplings. We list the ranges of parameters describing the critical volume of parameter space in Table 2, found after an initial exploration of the parameter space alongside physical considerations.

To train the emulator, we provide it with a vector specifying the location of the model in parameter coordinate space, i.e. $\vec{p} = (\log M_{\rm halo}^{\rm max}, \log M_{\rm halo}^{\rm max}, f_{\rm esc}^{\rm min}, \mu_{\rm c}, \log \mu_{\rm o})$, and the binned luminosity function ϕ constructed using an adaptive binning technique in $\log L_\alpha$ to control for Poisson error. We restrict the LF bin range to above $10^{41.5}\,\rm erg\,s^{-1}$ to avoid extrapolating beyond observational sensitivities. Once trained, the emulator can make highly accurate

predictions for $\phi(\vec{p}, L_\alpha)$ for any values of $\vec{p}$ and L_α covered by the hypercube and bins, respectively.

To find the best-fit parameters, we must compare our LFs to observed data. The Lyα LF shows the number density of LAEs in a given luminosity interval and is typically described using the Schechter (1976) function:

$$\phi(L){\rm d}(\log L) = \phi^* \left(\frac{L}{L^*}\right)^{\alpha+1} \exp\left(-\frac{L}{L^*}\right){\rm d}(\log L)\,, \tag{8}$$

where the redshift-dependent parameters are the normalization constant ϕ^*, power-law slope for the faint end α, and exponential cutoff scale for the bright end L^*. The rest-frame UVLF has been measured in detail out to $z \sim 10$ (Lyα LF out to $z \sim 7$) and is a powerful tool for studying the evolution of galaxy populations (e.g. Bouwens et al. 2015; Finkelstein et al. 2015; Oesch et al. 2015).

Our observational data is collated from Ouchi et al. (2008), Santos et al. (2016), and Konno et al. (2018) for $z = 5.5$, and from Ouchi et al. (2010), Santos et al. (2016), Konno et al. (2018), and Taylor et al. (2020) for $z = 6.6$. We fit Schechter functions to these data points using the curve_fit functionality available in the SCIPY library. The best-fit parameters for our Schechter functions are summarized in Table 3. Note that as we are only using these Schechter fits to 'smooth' out the data to simplify the interpretation of the fits, we do not fix any parameters in the Schechter fits. We then use the 'penalty' functionality available in SWIFTEMULATOR to compute the average absolute offset between the Schechter fit and the predicted scaling relation. Here, we calculate

$$I = {\rm mean}_{L_\alpha}\left(\frac{|\phi_{\rm fit}(L_\alpha) - \phi_{\rm em}(\vec{p}, L_\alpha)|}{\epsilon}\right)\,, \tag{9}$$

with a fixed offset (also known as model discrepancy) of $\epsilon = 0.4$, following the procedure in Rodrigues et al. (2017), for all L_α values across the range. We calculate this penalty for both redshifts, and take the combined penalty to be their mean. Our best-fit model is the model with the lowest combined penalty, which is calculated using the minimisation routine present in SCIPY along with the emulator.

4.2 Calibrated luminosity function results

We now present the results of the calibration along with our interpretation of the model parameter space. In Fig. 11 we illustrate the penalty structure within the critical volume of parameter space, combining both $z = 5.5$ and $z = 6.6$ maximum absolute offsets between the fit and emulated values as calculated by equation (9). For all parameter pairs, we show the combined penalty marginalized over the other model parameters, i.e. projected along the line of sight (LOS; upper right panels), and the volume fraction along the LOS spanned by implausibility defined by the penalty threshold $I < 0.4$ (lower left panels). Crucially, there is a small volume of parameter space that has an acceptable implausibility when matching both redshifts simultaneously, implying a tight constraint on many of our parameters. The best-fit model is shown as the white point within each panel, and we provide the specific parameter values in Table 4, including
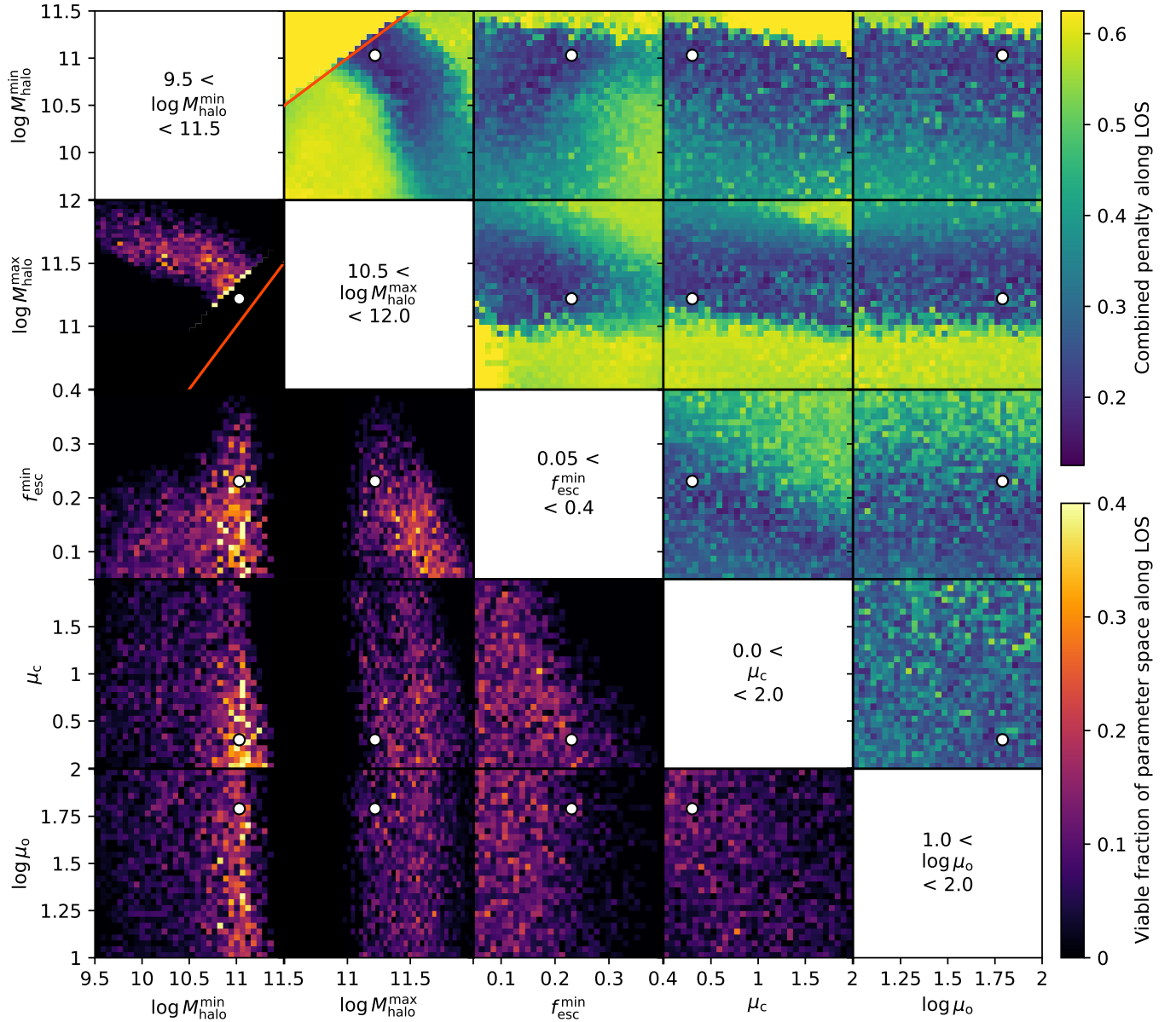
Figure 11. *Upper right:* Combined penalty for $z = 5.5$ and $z = 6.6$ data, as a function of all marginalized parameter pairs. *Lower left:* Volume fraction of parameter space along the line of sight spanned by implausibility $I < 0.6$. The red lines indicate the division where $M_{halo}^{min} > M_{halo}^{max}$ results in unphysical models. These views highlight the existence and prevalence of viable models, and place the best-fit model in the context of neighbouring penalty structures.

for the individual redshifts and our fiducial combined model. We note that μ_c, with a value of 0.3, is much lower than in other models, where it is generally around 1.5 (see Table 1). However, this may be compensated by having a dust model with a lower f_{esc}^{min}, or possibly a combination of adding a constant offset μ_o and different line widths. There is a clear indication in the panel of f_{esc}^{min} vs. μ_c that in good models these parameters are inversely correlated, or at least in projection there is a wedge of viable combinations.

In Table 4, we also include parameter values for the best-fit model calculated for each redshift separately. The parameter space constraints for both redshifts are similar to Fig. 11 (not shown), and all values are within this similar range of good models. The individual best-fit parameters are also relatively similar, with the notable differences mainly indicating a slight redshift dependence on the dust

Table 4. Best-fit parameters as determined by the Gaussian Process Regression technique. Values are given for each redshift separately and for our fiducial model achieving the lowest combined mean penalty. Note that the units for mass and μ_o are $M_\odot$ and $km\,s^{-1}$, respectively.

Redshift	$\log M_{halo}^{min}$	$\log M_{halo}^{max}$	f_{esc}^{min}	μ_c	$\log \mu_o$
5.5	10.36	11.49	0.1758	0.6657	1.949
6.6	11.02	11.34	0.2110	0.2750	1.851
Combined	11.03	11.22	0.2307	0.3015	1.790

or spectral model. Specifically, as expected the role of dust is more important at lower redshifts as M_{halo}^{min} decreases; while the dust escape fraction stays more or less the same, the mass region affected
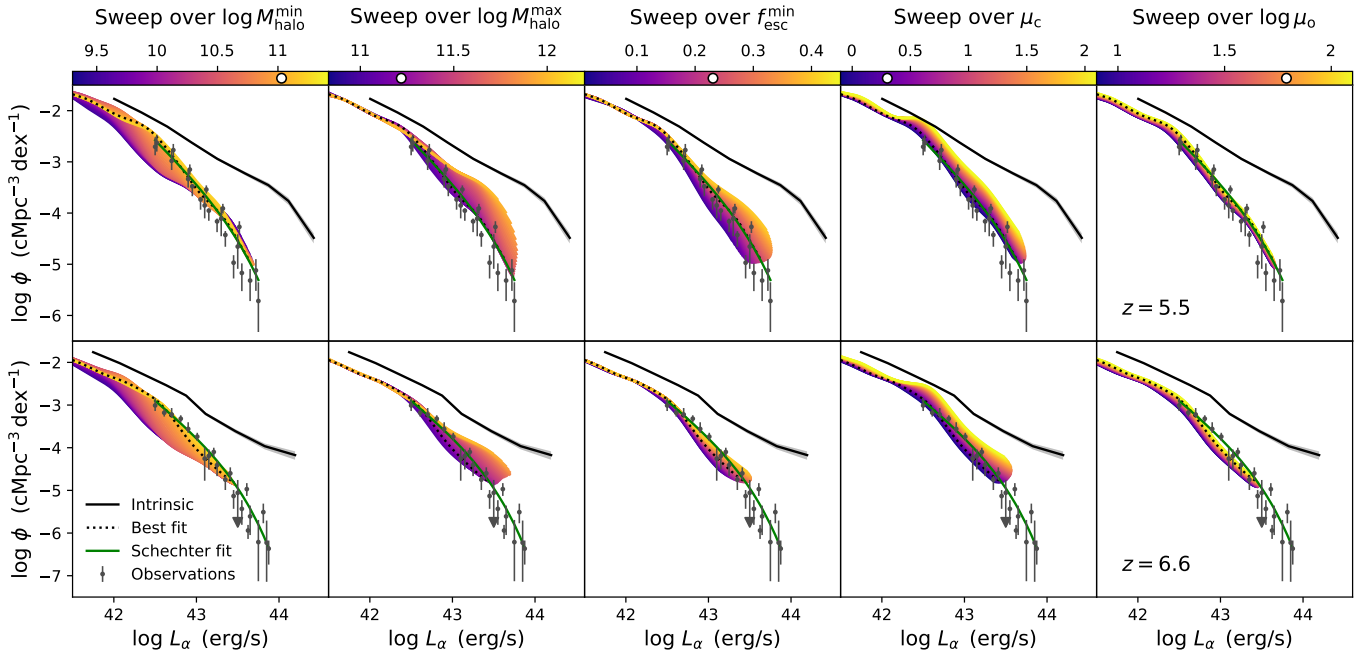
Figure 12. Sweep plots over all parameters, centred around the combined best fit for $z = 5.5$ and $z = 6.6$, with the best-fit model plotted as a dotted curve and also shown by white points in the colourbar axes. Observations are plotted in gray and the fitted Schechter function is in green. The intrinsic luminosity function is plotted in black. These trends confirm our physical intuition, and we also see that $\log \mu_o$ has less of an effect on the observed LF than the other parameters. We note that for sweeps over $\log M_{halo}^{min}$ and $\log M_{halo}^{max}$, we remove unphysical models with $M_{halo}^{min} > M_{halo}^{max}$.

by dust increases. For higher redshifts, the values of $\log M_{halo}^{min}$ and $\log M_{halo}^{max}$ are close together, making the escape fraction look close to a step function. In fact, Fig. 11 implies that this scenario for the escape fraction is the most reasonable, since the good models cluster near the red line where $\log M_{halo}^{min} = \log M_{halo}^{max}$. We note however that caveats to our dust model (detailed in section 4.3) mean that the shape of the f_{esc} line should be taken with a grain of salt. Finally, we also see that μ_c increases from $z = 6.6$ to $z = 5.5$. However, the constraints on μ_c from Fig. 11 are fairly weak so this difference is likely due to the parameter degeneracies in fitting the LF.

In Fig. 12, we plot sweeps over the range of each of the five parameters for both redshifts. Our best-fit model is plotted as a dotted line. The Schechter fits to the observations are shown in green, observations in gray, and the intrinsic LFs in black. The sweep for a particular parameter is performed by holding all other parameters constant at their best-fit values and predicting the LF for 50 values equally spaced over the specified parameter's range. This allows us to see how the specific parameter affects the LF. Specifically, we see that $\log \mu_o$ doesn't have much of an effect on the LF, while the other parameters provide much stronger constraints. This is supported by Fig. 11, where the penalty plots for $\log \mu_o$ have a larger region of 'good' models than the others.

Additionally, Fig. 12 verifies that the parameters make physical sense and affect the LF in reasonable ways. For all the parameters, having a lower value corresponds to the LF moving to the left. This is expected, as both a lower M_{halo}^{min} and M_{halo}^{max} lead to dust absorption affecting lower mass, dimmer haloes, leading to more absorption overall, and a lower LF. A lower f_{esc}^{min} also leads to more dust absorption. Finally, a smaller μ_c or μ_o means a smaller emergent spectral line offset, so the haloes are less luminous to begin with.

In Fig. 13, we apply our best-fit model to the entire suite of THESAN simulations and compare against observations. We find that all of the

simulated LFs fit the observations reasonably well. Still, the simulations do not align with each other so the model calibration would need to be redone in each case to account for the differences between them. However, individual calibrations would effectively give a dust and spectral model that counteracts the physical differences distinguishing each simulation. Furthermore, these are mainly carried over from the differences in the intrinsic LFs (plotted as dotted lines) and IGM transmission, caused by factors explored in the previous section such as reionization history and bubble morphology. For example, we once again see THESAN-LOW-2 and THESAN-HIGH-2 bracketing the other simulations. To test the sensitivity of the calibration to the different physics of the simulations, however, we performed a calibration for THESAN-WC-2 data and found very similar constraints; more detailed results are given in Appendix B.

4.3 Implications for observed statistics

Our spectral and dust model allows us to make broad predictions for observables that can be measured by current and future instruments, which we explore in this section. In Fig. 2, we show a visualization of the resolved recombination rate density along image sightlines calculated as $\int \alpha_B n_p n_e \, dV / \int dV$ similar to Equation (1) for a $95.5 \times 95.5 \times 2.3875 \, \text{cMpc}^3$ subvolume of the box at $z = 6.6$, along with points showing the intrinsic and observed Lyα luminosity of bright galaxies. We see that the background Lyα emission structure is highly correlated with both the intrinsic and observed luminosities, as expected, with galaxies of similar luminosities clustering together.

In Fig. 14, we plot the ratio of observed-to-intrinsic Lyα luminosity including dust and IGM transmission effects ($L_\alpha^{obs}/L_\alpha^{int} = f_{esc} \times \mathcal{T}_{IGM}$) for all THESAN simulations as functions of the global neutral fraction. We applied the same best-fit dust and spectral model to all simulations and redshifts. As before, we cut the sample at intrinsic
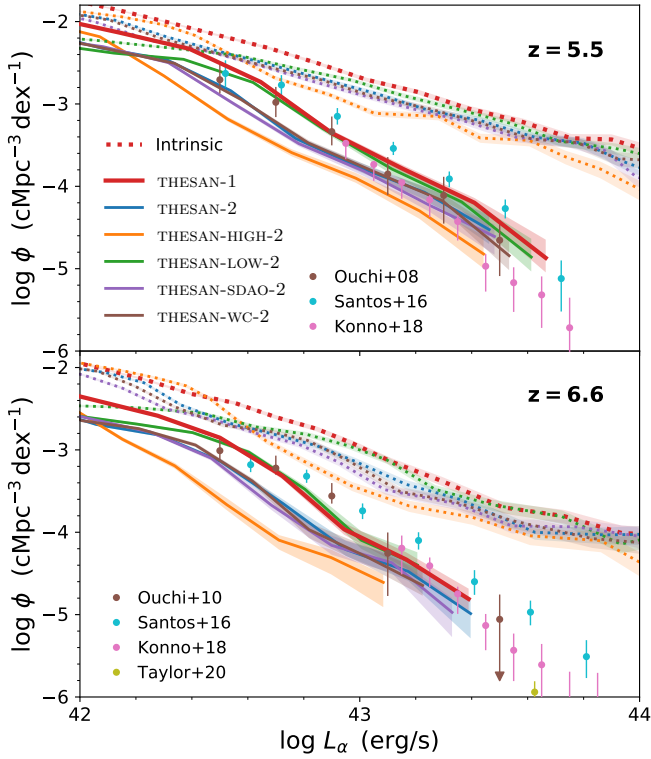
Figure 13. Comparison between the observed luminosity functions for all THESAN simulations after applying our best-fit model calibrated to THESAN-1 (thicker red line). Intrinsic luminosity functions are plotted as dotted lines. The ordering is preserved, indicating that physical differences affecting the intrinsic LFs are important in addition to differences in IGM transmission.

Table 5. Ratios of observed-to-intrinsic Lyα luminosity ($f_{\rm esc} \times \mathcal{T}_{\rm IGM}$) and covering fractions defined as the fraction of sightlines for each galaxy with ratios below 10 per cent $P(< 0.1)$ for the THESAN-1 simulation tabulated as functions of redshift (as percentages). The summary statistics are calculated including all central galaxies with UV brightness $M_{1500} < -19$ and are given in per cent units with median and asymmetric 1σ confidence regions.

Quantity	$z = 6$	$z = 7$	$z = 8$	$z = 9$	$z = 10$
$f_{\rm esc} \times \mathcal{T}_{\rm IGM}$	$32.9^{+14.4}_{-23.8}$	$25.6^{+12.9}_{-17.9}$	$16.6^{+10.5}_{-11.0}$	$9.8^{+7.4}_{-6.5}$	$4.4^{+4.9}_{-3.1}$
$P(< 0.1)$	$6.8^{+49.9}_{-2.5}$	$10.4^{+7.2}_{-3.0}$	$20.1^{+9.5}_{-5.2}$	$51.0^{+13.7}_{-15.6}$	$89.1^{+6.8}_{-10.7}$

magnitude, which is a bit counter intuitive because the IGM transmission bands show an increase with M_{1500} and strong infall velocity signatures. However, when considering the emergent spectral profile this seems to be erased, likely because brighter galaxies have larger velocity offsets.

While there is a correspondence between the observed-to-intrinsic luminosity and the fraction of galaxies that are LAEs, after completing our analysis we discovered the idealized dust model is incompatible with corrections typically used for $M_{\rm UV}$ (Gnedin 2014; Kannan et al. 2022a), so we do not predict realistic rest-frame equivalent widths (EWs) calculated as EW $= L_\alpha^{\rm obs}/L_{\lambda, \rm cont}^{\rm obs}$. Therefore, the predicted observed-to-intrinsic luminosity ratios should be viewed as statistically correct; i.e. they reproduce observed LFs, but the manner in which this is achieved yields too high of EWs for the most massive galaxies. This is because the Lyα escape fraction cannot be lower than $f_{\rm esc}^{\rm min}$ while the UV escape fraction can be arbitrarily low calculated as $\exp(-\tau)$. The degeneracy between dust absorption and IGM transmission effectively means that our massive galaxies have too high $f_{\rm esc}$ and too low $\mathcal{T}_{\rm IGM}$. However, we emphasize that the general $f_{\rm esc} \times \mathcal{T}_{\rm IGM}$ predictions remain trustworthy, e.g. UV bright galaxies have significant dust absorption while fainter objects ($M_{1500} \gtrsim -19$) are more direct probes of the IGM evolution. In any case, this motivates a future analysis with self-consistent Lyα and continuum dust corrections.

To qualitatively understand the evolution of the LAE fraction (Stark et al. 2011; Schenker et al. 2014; Mason et al. 2019), in the bottom panel of Fig. 15, we also plot a possible prediction of the rest-frame EW as a function of observed M_{1500}. We calculate the observed continuum escape fractions following the Gnedin (2014) correction and our best-fit model for the observed Lyα luminosities. The line continuum value at 1216 Å is extrapolated from the dust-corrected values at 1500 and 2500 Å, i.e. assuming power-law UV slopes (see section 3.2 of Smith et al. 2022a). However, realizing that the dust models are incompatible we also limit the UV dust correction escape fraction for each galaxy so it does not dip below our Lyα escape fraction, as otherwise it would not be physical. For reference, we plot a horizontal dotted grey line at 25 Å to mark the minimum threshold to be classified as an LAE. We see that starting at $z \lesssim 7$ fainter galaxies at observed UV magnitudes $M_{1500} \gtrsim -20$ largely fall above this line. This is exactly the LAE fraction evolution that we expect. However, the lack of Lyman break galaxies, i.e. with low-EWs and therefore not LAEs, is not realistic. This is again due to the escape fraction for the continuum correction around the line going down to zero, while our dust escape fraction for Lyα plateaus at $f_{\rm esc}^{\rm min}$ after $M_{\rm halo}^{\rm max}$, so our predicted EWs with escape fraction limiting are now well-behaved but are still generally too high at the bright end.

This problem would likely be solved with another dust model where the dust escape fractions for Lyα photons are required to be less than or equal to those of continuum photons. This would naturally
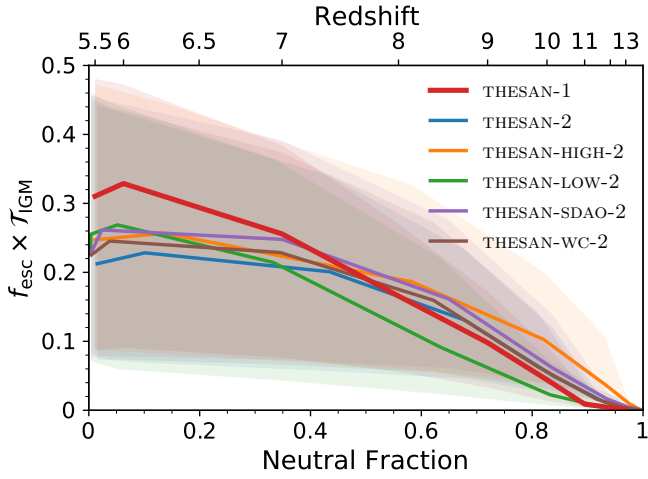
UV magnitudes of $M_{1500} < -19$ to avoid resolution biases across the runs, though similar to Figs. 3–6 the equal weighting to all galaxies results in a bias towards lower masses. We also include redshift markings along the top axis corresponding to the THESAN-1 neutral fraction history. We do not apply an equivalent width lower limit to select for LAEs, but note that this ratio would be higher if we did, as this removes dust-obscured and transmission-suppressed galaxies.

The ordering of the simulations is slightly different from the plot of $\mathcal{T}_{\rm IGM}$ in Fig. 5. Here, THESAN-1 transmission is at the higher end of the simulations, while before it was lower. This is a reflection of differences in the intrinsic LFs and the stronger impact our model has on the other simulations for which it was not calibrated. As in Fig. 5, THESAN-LOW-2 and THESAN-HIGH-2 once again bracket the medium resolution simulations early on due to reionization history and morphology, with THESAN-LOW-2 below the others and THESAN-HIGH-2 above, but all simulations are roughly equivalent below $z \approx 7$. Another similarity is the more or less linear progression upwards as neutral fraction decreases, though of course the value does not increase to unity. Instead, THESAN-1 peaks around 30^{+10}_{-20} % showing significant sightline-to-sightline and galaxy-to-galaxy variations, as summarized in Table 5.

In the top panel of Fig. 15, we show the observed-to-intrinsic luminosity ($f_{\rm esc} \times \mathcal{T}_{\rm IGM}$) and transmission without dust ($\mathcal{T}_{\rm IGM}$; dotted curves) as functions of the observed UV magnitude for THESAN-1 at several redshifts. Including dust absorption causes the value to begin decreasing around a magnitude of -18 to -19, as expected due to the absorption fraction increasing with magnitude. Without dust absorption our model exhibits a relatively flat dependence on UV

Figure 14. Ratio of observed-to-intrinsic Lyα luminosity ($f_{esc} \times \mathcal{T}_{IGM}$) for all THESAN simulations as functions of the global neutral fraction after applying our best-fit model. As in Fig. 5, the sample is cut at $M_{1500} < -19$ (or brighter magnitudes depending on redshift) to avoid biases. The top redshift axis corresponds to the THESAN-1 reionization history.

reduce the EWs of the most massive galaxies $M_{halo} \gtrsim 10^{11.5}\,M_\odot$ to below LAE thresholds 25 Å at all redshifts. We leave this to a future study and include our current model as a first exploration of the emulated calibration technique that succeeds in matching observed Lyα LFs but requires improvements to predict realistic EWs.

5 CONCLUSIONS

In this paper, we have constructed catalogues for Lyα emission and transmission for the entire suite of THESAN simulations. The THESAN simulations have been shown to provide a unique framework for statistically studying Lyα properties throughout the EoR (Smith et al. 2022a), and we have extended the analysis to include comparisons between the medium resolution simulations, each with different reionization histories and bubble morphologies. This allowed us to explore the effects of spatial resolution, alternative dark matter models, halo mass-dependent escape fractions, and numerical convergence on the frequency-dependent IGM transmission throughout the EoR ($5.5 \leq z \leq 13$). It is unclear how to interpret Monte Carlo Lyα radiative transfer calculations in the context of the galaxy formation model used in THESAN. Therefore, we chose to employ an empirical model for the dust escape fractions and emergent spectral profiles to act as a bridge between the much more robust intrinsic emission and IGM transmission predictions. We then used a Gaussian Process Regression technique to calibrate our model to optimally reproduce the observed Lyα luminosity function. The dust and spectral models have degenerate absorption properties, which complicates our ability to uniquely determine the precise physical origin of the large difference between intrinsic and observed LFs. However, we were still able to find a small volume of parameter space with good models that fit observations, which demonstrates the power of this type of empirical modelling. We summarize our main conclusions as follows:

(i) In all simulation variations, transmission redward of line centre increases with decreasing redshift while transmission blueward is highly suppressed until around $z \approx 6$, and remains low throughout the tail-end of reionization, as expected (see Fig. 3).
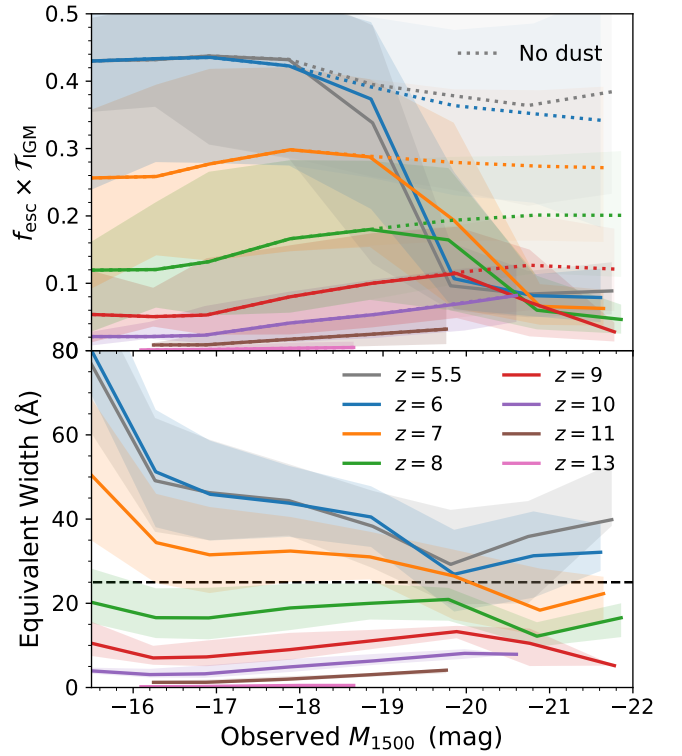

Figure 15. *Upper panel:* Observed-to-intrinsic Lyα luminosity ($f_{esc} \times \mathcal{T}_{IGM}$) and transmission without dust ($\mathcal{T}_{IGM}$; dotted curves) as functions of observed M_{1500} for THESAN-1 at several redshifts. *Lower panel:* Observed rest-frame EW as a function of observed M_{1500}. As described in the text, we limit the continuum escape fractions by our best-fit model Lyα escape fractions to avoid unrealistic escape scenarios. Despite the lack of non-LAE massive galaxies, the redshift evolution around $M_{1500} \sim -19$ roughly agrees with our expectation for the declining LAE fraction in the pre-reionization epoch.

(ii) Transmission around the red peak is also highly sensitive to frequency. For a waveband centred at $\Delta v = 200\,\mathrm{km\,s^{-1}}$, transmission is more suppressed and has more variation than at $\Delta v = 400\,\mathrm{km\,s^{-1}}$, since it is more affected by the local environment and infalling gas (see Fig. 4).

(iii) THESAN-LOW-2 has the earliest reionization history due to the dominant role of low-mass galaxies contributing to reionization. On the other hand, THESAN-HIGH-2 is delayed, as high-mass galaxies are much rarer at early times. This leads to differences in the transmission curves: THESAN-LOW-2 transmission is boosted while THESAN-HIGH-2 transmission is more suppressed for all redshifts.

(iv) When correcting for reionization history by looking at the global neutral fraction instead, the red peak transmission increases nearly linearly with decreasing neutral fraction. The difference between simulations also highlights the importance of bubble morphology on transmission, with THESAN-HIGH-2 having an advantage compared to THESAN-LOW-2 at the same neutral fraction (see Fig. 5).

(v) The variation across sightlines for a single galaxy is greater than the variation globally across all galaxies, mainly because a galaxy's environment can look significantly different in different directions (see Fig. 6).

(vi) Red peak transmission increases with UV magnitude M_{1500} at $z \gtrsim 6$, except near line centre, when there is a downturn for bright galaxies because of infalling gas. This behaviour is mirrored in the corresponding covering fractions defined as the fraction of

sightlines around each galaxy with low transmission $P(\mathcal{T}_{\text{IGM}} < 0.2)$ (see Figs. 7 and 8).

(vii) Previous empirical models do not account for dust absorption, however we find doing so is essential in realistic comparisons to observed luminosity functions, as this becomes the dominant mechanism for removing Lyα photons in massive galaxies at $z \lesssim 7$ (see Figs. 9 and 15).

(viii) In our idealized dust and spectral model, the minimum dust escape fraction ($f_{\text{esc}}^{\text{min}}$) and spectral profile offset from line centre (μ_c) are inversely correlated in the space for good models, as both cause a lower observed LF. Our framework also allows us to understand model degeneracies more generally (see Figs. 11 and 12).

(ix) Intrinsic luminosity functions and IGM transmission curves vary across different THESAN simulations, so parameter recalibration of the ratio of observed-to-intrinsic Lyα luminosity ($f_{\text{esc}} \times \mathcal{T}_{\text{IGM}}$) is recommended for different simulations. We found our best-fit model works well for THESAN-1 and is reasonable for the others, except in the case of THESAN-HIGH-2 (see Figs. 13 and 14).

(x) The dust escape fractions for Lyα and stellar continuum photons must be modelled self-consistently. Otherwise, it is possible to have $f_{\text{esc}}^{\text{Ly}\alpha} > f_{\text{esc}}^{\text{cont}}$ and the model overpredicts the equivalent widths of high mass galaxies. However, due to dust–spectral degeneracies, realistic predictions for observable LAE statistics beyond LFs are still possible (see Fig. 15).

In the future we plan to design and calibrate a dust model in which Lyα escape fractions are limited by continuum ones. Allowing f_{esc} to approach zero asymptotically for high mass galaxies will likely lead to more reasonable EW predictions. Such an improvement will expand the utility of our catalogues for further applications, including understanding the clustering of LAEs, 21 cm cross-correlations, connections to quasar absorption troughs, and other Lyα-centric observational probes of reionization. Overall, constrained empirical modelling represents a promising avenue to elucidate LAE science, especially when combined with state-of-the-art large-volume galaxy formation and reionization simulations with realistic Lyα emission and transmission.

DATA AVAILABILITY

All THESAN simulation data will be made publicly available in the near future, including Lyα catalogues. Data will be distributed via www.thesan-project.com. Before the public data release, data underlying this article will be shared on reasonable request to the corresponding author(s).

REFERENCES

Adams J. J., et al., 2011, ApJS, 192, 5
Angulo R. E., Pontzen A., 2016, MNRAS, 462, L1
Bacon R., et al., 2010, in McLean I. S., Ramsay S. K., Takami H., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 7735, Ground-based and Airborne Instrumentation for Astronomy III. p. 773508, doi:10.1117/12.856027
Bacon R., et al., 2017, A&A, 608, A1
Barkana R., Loeb A., 2001, Phys. Rep., 349, 125
Barnes J., Hut P., 1986, Nature, 324, 446
Bouwens R. J., et al., 2015, ApJ, 803, 34
Byrohl C., Gronke M., 2020, A&A, 642, L16
Chabrier G., 2003, ApJ, 586, L133
Choudhury T. R., Puchwein E., Haehnelt M. G., Bolton J. S., 2015, MNRAS, 452, 261
Dayal P., Ferrara A., 2018, Phys. Rep., 780, 1
Dijkstra M., 2014, Publ. Astron. Soc. Australia, 31, e040
Dijkstra M., 2019, Saas-Fee Advanced Course, 46, 1
Dijkstra M., Wyithe J. S. B., 2012, MNRAS, 419, 3181
Dijkstra M., Lidz A., Wyithe J. S. B., 2007, MNRAS, 377, 1175
Dijkstra M., Mesinger A., Wyithe J. S. B., 2011, MNRAS, 414, 2139
Eldridge J. J., Stanway E. R., Xiao L., McClelland L. A. S., Taylor G., Ng M., Greis S. M. L., Bray J. C., 2017, Publ. Astron. Soc. Australia, 34, e058
Endsley R., Stark D. P., 2022, MNRAS, 511, 6042
Endsley R., Stark D. P., Charlot S., Chevallard J., Robertson B., Bouwens R. J., Stefanon M., 2021, MNRAS, 502, 6044
Finkelstein S. L., Rhoads J. E., Malhotra S., Grogin N., 2009, ApJ, 691, 465
Finkelstein S. L., et al., 2015, ApJ, 810, 71
Furlanetto S. R., Lidz A., 2007, ApJ, 660, 1030
Gangolli N., D'Aloisio A., Nasir F., Zheng Z., 2021, MNRAS, 501, 5294
Garaldi E., Kannan R., Smith A., Springel V., Pakmor R., Vogelsberger M., Hernquist L., 2022, MNRAS, 512, 4909
Garel T., Blaizot J., Rosdahl J., Michel-Dansac L., Haehnelt M. G., Katz H., Kimm T., Verhamme A., 2021, MNRAS, 504, 1902
Gebhardt K., et al., 2021, ApJ, 923, 217
Gnedin N. Y., 2014, ApJ, 793, 29
Gnedin N. Y., 2016, ApJ, 833, 66
Goto H., et al., 2021, ApJ, 923, 229
Greig B., Mesinger A., Haiman Z., Simcoe R. A., 2017, MNRAS, 466, 4239
Gronke M., Dijkstra M., Trenti M., Wyithe S., 2015, MNRAS, 449, 1284
Gronke M., et al., 2021, MNRAS,
Gunn J. E., Peterson B. A., 1965, ApJ, 142, 1633
Heneka C., Cooray A., Feng C., 2017, ApJ, 848, 52
Hoag A., et al., 2019, ApJ, 878, 12
Hunter J. D., 2007, Computing In Science & Engineering, 9, 90
Hutter A., Dayal P., Müller V., Trott C. M., 2017, ApJ, 836, 176
Jensen H., Laursen P., Mellema G., Iliev I. T., Sommer-Larsen J., Shapiro P. R., 2013, MNRAS, 428, 1366
Jones E., Oliphant T., Peterson P., et al., 2001, SciPy: Open source scientific tools for Python, http://www.scipy.org/
Jung I., et al., 2020, ApJ, 904, 144
Jung I., et al., 2022, ApJ, 933, 87
Kakiichi K., Dijkstra M., Ciardi B., Graziani L., 2016, MNRAS, 463, 4019
Kannan R., Vogelsberger M., Marinacci F., McKinnon R., Pakmor R., Springel V., 2019, MNRAS, 485, 117
Kannan R., et al., 2022a, arXiv e-prints, p. arXiv:2210.10066
Kannan R., Garaldi E., Smith A., Pakmor R., Springel V., Vogelsberger M., Hernquist L., 2022b, MNRAS, 511, 4005

Kannan R., Smith A., Garaldi E., Shen X., Vogelsberger M., Pakmor R., Springel V., Hernquist L., 2022c, MNRAS, 514, 3857

Katz H., et al., 2019, MNRAS, 487, 5902

Kennicutt R. C., Evans N. J., 2012, ARA&A, 50, 531

Khusanova Y., et al., 2020, A&A, 634, A97

Konno A., et al., 2018, PASJ, 70, S16

Kugel R., Borrow J., 2022, The Journal of Open Source Software, 7, 4240

Laursen P., Sommer-Larsen J., Razoumov A. O., 2011, ApJ, 728, 52

Lee H.-W., 2013, ApJ, 772, 123

Levermore C. D., 1984, J. Quant. Spectrosc. Radiative Transfer, 31, 149

Lusso E., Worseck G., Hennawi J. F., Prochaska J. X., Vignali C., Stern J., O'Meara J. M., 2015, MNRAS, 449, 4204

Madau P., Rees M. J., 2000, ApJ, 542, L69

Malhotra S., Rhoads J. E., 2004, ApJ, 617, L5

Marinacci F., et al., 2018, MNRAS, 480, 5113

Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2018a, ApJ, 856, 2

Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2018b, ApJ, 856, 2

Mason C. A., et al., 2019, MNRAS, 485, 3947

McKinnon R., Torrey P., Vogelsberger M., Hayward C. C., Marinacci F., 2017, MNRAS, 468, 1505

McQuinn M., Hernquist L., Zaldarriaga M., Dutta S., 2007, MNRAS, 381, 75

Mesinger A., Aykutalp A., Vanzella E., Pentericci L., Ferrara A., Dijkstra M., 2015, MNRAS, 446, 566

Miralda-Escudé J., 1998, ApJ, 501, 15

Morales A. M., Mason C. A., Bruton S., Gronke M., Haardt F., Scarlata C., 2021, ApJ, 919, 120

Naiman J. P., et al., 2018, MNRAS, 477, 1206

Nelson D., et al., 2018, MNRAS, 475, 624

Nelson D., et al., 2019, MNRAS, 490, 3234

Oesch P. A., Bouwens R. J., Illingworth G. D., Franx M., Ammons S. M., van Dokkum P. G., Trenti M., Labbé I., 2015, ApJ, 808, 104

Ono Y., et al., 2012, ApJ, 744, 83

Ota K., et al., 2008, ApJ, 677, 12

Ouchi M., et al., 2008, ApJS, 176, 301

Ouchi M., et al., 2010, ApJ, 723, 869

Ouchi M., et al., 2018, PASJ, 70, S13

Ouchi M., Ono Y., Shibuya T., 2020, ARA&A, 58, 617

Park H., et al., 2021, ApJ, 922, 263

Partridge R. B., Peebles P. J. E., 1967, ApJ, 147, 868

Pentericci L., et al., 2014, ApJ, 793, 113

Perez L. A., Malhotra S., Rhoads J. E., Laursen P., Wold I. G. B., 2022, ApJ, 940, 102

Pillepich A., et al., 2018a, MNRAS, 473, 4077

Pillepich A., et al., 2018b, MNRAS, 475, 648

Pillepich A., et al., 2019, MNRAS, 490, 3196

Planck Collaboration et al., 2016, A&A, 594, A13

Qin W., Schutz K., Smith A., Garaldi E., Kannan R., Slatyer T. R., Vogelsberger M., 2022, Phys. Rev. D, 106, 123506

Rhoads J. E., Malhotra S., Dey A., Stern D., Spinrad H., Jannuzi B. T., 2000, ApJ, 545, L85

Rodrigues L. F. S., Vernon I., Bower R. G., 2017, MNRAS, 466, 2418

Rybicki G. B., Lightman A. P., 1986, Radiative Processes in Astrophysics

Santos M. R., 2004, MNRAS, 349, 1137

Santos S., Sobral D., Matthee J., 2016, MNRAS, 463, 1678

Sarkar A., Samui S., 2019, PASP, 131, 074101

Schechter P., 1976, ApJ, 203, 297

Schenker M. A., Ellis R. S., Konidaris N. P., Stark D. P., 2014, ApJ, 795, 20

Smith A., Safranek-Shrader C., Bromm V., Milosavljević M., 2015, MNRAS, 449, 4336

Smith A., Ma X., Bromm V., Finkelstein S. L., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2019, MNRAS, 484, 39

Smith A., Kannan R., Garaldi E., Vogelsberger M., Pakmor R., Springel V., Hernquist L., 2022a, MNRAS, 512, 3243

Smith A., et al., 2022b, MNRAS, 517, 1
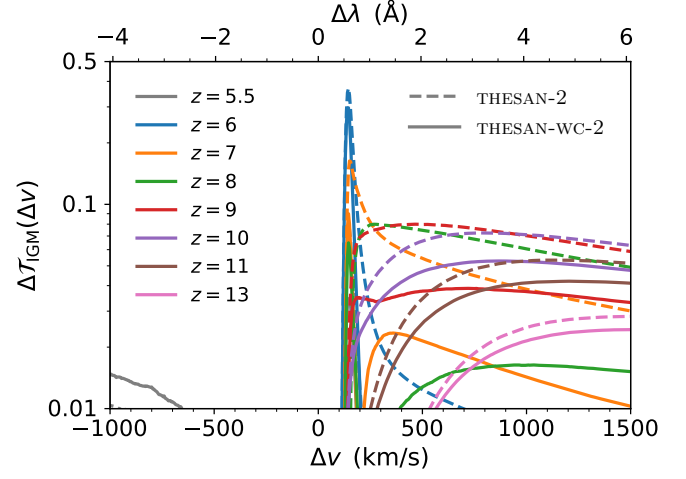
Sobacchi E., Mesinger A., 2015, MNRAS, 453, 1843

Figure A1. The absolute difference $\Delta\mathcal{T}_{\mathrm{IGM}} \equiv |\mathcal{T}_{\mathrm{IGM,1}} - \mathcal{T}_{\mathrm{IGM,2}}|$ between THESAN-1 and THESAN-2 (dashed lines) and THESAN-WC-2 (solid lines) simulations, shown for median statistics. This spatial resolution convergence comparison reveals a strong feature corresponding to the location of the step function transition and overall agreement at the level of $\Delta\mathcal{T}_{\mathrm{IGM}} \lesssim 0.05$ for the weak convergence run elsewhere.

Springel V., 2010, MNRAS, 401, 791

Springel V., Hernquist L., 2003, MNRAS, 339, 289

Springel V., et al., 2018, MNRAS, 475, 676

Stark D. P., Ellis R. S., Ouchi M., 2011, ApJ, 728, L2

Taniguchi Y., et al., 2005, PASJ, 57, 165

Taylor A. J., Barger A. J., Cowie L. L., Hu E. M., Songaila A., 2020, ApJ, 895, 132

Taylor A. J., Cowie L. L., Barger A. J., Hu E. M., Songaila A., 2021, ApJ, 914, 79

Tilvi V., et al., 2014, ApJ, 794, 5

Verhamme A., et al., 2018, MNRAS, 478, L60

Walt S. v. d., Colbert S. C., Varoquaux G., 2011, Computing in Science & Engineering, 13, 22

Weinberger R., et al., 2017, MNRAS, 465, 3291

Weinberger R., et al., 2018, MNRAS, 479, 4056

Weinberger L. H., Haehnelt M. G., Kulkarni G., 2019, MNRAS, 485, 1350

Weinberger R., Springel V., Pakmor R., 2020, ApJS, 248, 32

Whitler L. R., Mason C. A., Ren K., Dijkstra M., Mesinger A., Pentericci L., Trenti M., Treu T., 2020, MNRAS, 495, 3602

Wise J. H., 2019, Contemporary Physics, 60, 145

Wold I. G. B., et al., 2022, ApJ, 927, 36

Yang H., Malhotra S., Gronke M., Rhoads J. E., Dijkstra M., Jaskot A., Zheng Z., Wang J., 2016, ApJ, 820, 130

Yeh J. Y. C., et al., 2023, MNRAS, 520, 2757

van de Voort F., Springel V., Mandelker N., van den Bosch F. C., Pakmor R., 2019, MNRAS, 482, L85

APPENDIX A: RESOLUTION COMPARISON

Throughout this paper we have compared various results from THESAN-1 and the medium resolution simulations. In this section we directly examine where the main differences are in the frequency-dependent IGM transmission curves. Due to chaotic variations in galaxy formation modelling, we do not expect identical results at the level of individual haloes even for simulations run with the same random number generator seeds. However, it is reasonable to compare the median (or mean) transmission curves between THESAN-1 and THESAN-2, which has a factor of eight lower mass resolution but

Table B1. Best-fit parameters for THESAN-WC-2. Note that the units for mass and μ_o are $M_\odot$ and $km\,s^{-1}$, respectively.

Redshift	$\log M_{\rm halo}^{\rm min}$	$\log M_{\rm halo}^{\rm max}$	$f_{\rm esc}^{\rm min}$	$\mu_{\rm c}$	$\log \mu_{\rm o}$
5.5	11.07	11.24	0.1509	1.6986	1.663
6.6	11.18	11.33	0.0762	1.5284	1.316
Combined	10.95	11.28	0.1555	1.5648	1.744

is otherwise the same. In Fig. A1 we show the absolute difference $\Delta\mathcal{T}_{\rm IGM} \equiv |\mathcal{T}_{\rm IGM,1} - \mathcal{T}_{\rm IGM,2}|$ for the THESAN-2 and THESAN-WC-2 simulations. These both show a strong feature at $\Delta v \approx 100$–$200\,km\,s^{-1}$ resulting from the slightly different locations of the transition from blue suppression to red transmission, which occurs closer to line center for THESAN-1 as a result of resolving additional dense substructure for larger covering fractions of inflowing H I gas. There are also differences in the red damping-wing transmission at $\Delta v \gtrsim 300\,km\,s^{-1}$, which can be explained by differences in the reionization history and bubble morphology. The weak convergence run partially accounts for these differences by increasing the birth cloud escape fraction from 0.37 to 0.43, and this is reflected in the comparison in this region with $\Delta\mathcal{T}_{\rm IGM} < 0.08$ for THESAN-2 and much lower for THESAN-WC-2 in some redshift cases. Overall, the impact of resolution in the IGM (outside $R_{\rm vir}$) is minor.

APPENDIX B: THESAN-WC-2 CALIBRATION

Our dust and spectral model calibration in the main body of the paper was only done for THESAN-1. Here we present the results of the same calibration routine applied to THESAN-WC-2. The best-fit model, given in Table B1, has a higher $\mu_{\rm c}$ and lower $f_{\rm esc}^{\rm min}$ than was found for THESAN-1. However, the penalty structure in Fig. B1 is extremely similar to Fig. 11, with the same negative correlation between $\mu_{\rm c}$ and $f_{\rm esc}^{\rm min}$. It is therefore more likely that differences in the best-fit model are mainly due to parameter degeneracies rather than statistically significant physical differences. We also applied this best-fit model to all simulations at $z = 5.5$ and 6.6, the results of which are shown in Fig. B2. Again, all simulated LFs match observations, with differences carrying over from the intrinsic LFs. Overall, the simulation resolution did not have a significant effect on calibration results other than correcting slightly for physical differences.

This paper has been typeset from a TeX/LaTeX file prepared by the author.
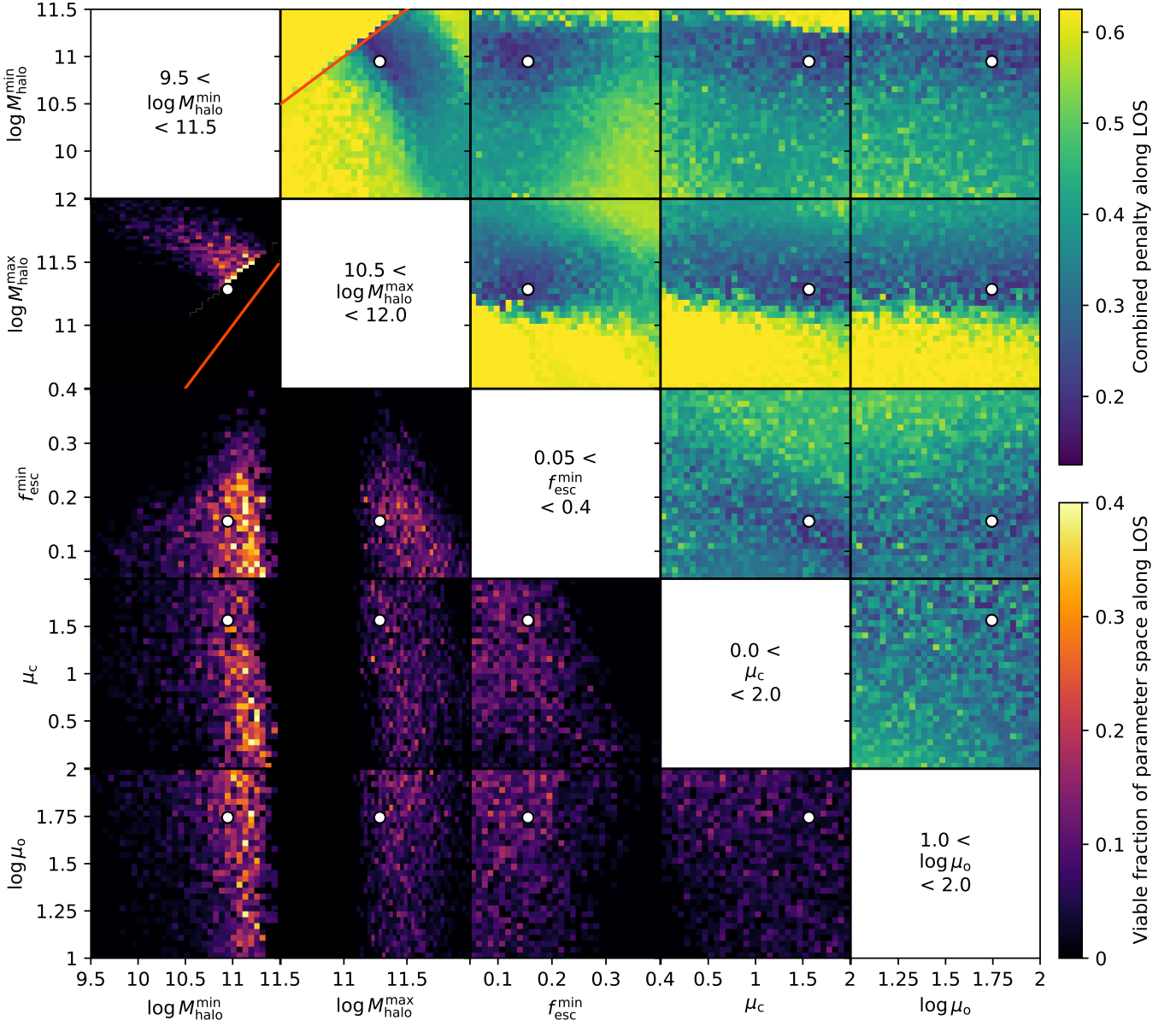
Figure B1. Counterpart to Fig. 11 for the ᴛʜᴇꜱᴀɴ-ᴡᴄ-2 calibration. *Upper right:* Combined penalty for $z = 5.5$ and $z = 6.6$ data. *Lower left:* Volume fraction of parameter space along the line of sight spanned by implausibility $I < 0.6$. The red lines indicate the division where $M_{\mathrm{halo}}^{\mathrm{min}} > M_{\mathrm{halo}}^{\mathrm{max}}$ results in unphysical models. The penalty structure for ᴛʜᴇꜱᴀɴ-ᴡᴄ-2 shows the same shape as the one for ᴛʜᴇꜱᴀɴ-1, implying that the resolution did not significantly affect the calibration.
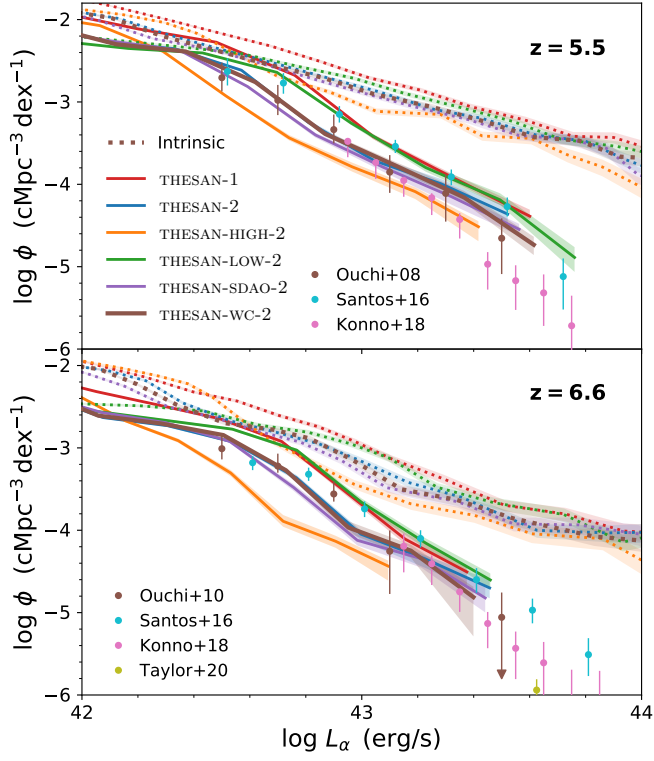
Figure B2. Comparison between the observed luminosity functions for all THESAN simulations after applying our best-fit model calibrated to THESAN-WC-2. Intrinsic luminosity functions are plotted as dotted lines. Much like Fig. 13, the ordering is preserved and all LFs match observations reasonably well.