Determining the systemic redshift of Lyman-α emitters with neural networks and improving the measured large-scale clustering.

Siddhartha Gurung-López,1,2⋆ Shun Saito,1,3 Carlton M. Baugh,4 Silvia Bonoli,5 Cedric G. Lacey,4 and Álvaro A. Orsi6,2.

1 Institute for Multi-messenger Astrophysics and Cosmology, Department of Physics Missouri University of Science and Technology, 1315 N. Pine St., Rolla MO 65409, USA
2 Centro de Estudios de Física del Cosmos de Aragón, Plaza San Juan 1, piso 2, Teruel, 44001, Spain.
3 Kavli Institute for the Physics and Mathematics of the Universe (WPI), Todai Institutes for Advanced Study, the University of Tokyo, Kashiwanoha, Kashima, Chiba 277-8583, Japan
4 Institute for Computational Cosmology, Durham University.
5 DIPC, Manuel Lardizabal Ibilbidea, 4, 20018 San Sebastian, Spain.
6 PlantTech Research Institute Limited. South British House, 4th Floor, 35 Grey Street, Tauranga 3110, New Zealand

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We explore how to mitigate the clustering distortions in Lyman-α emitters (LAEs) samples caused by the mis-identification of the Lyman-α (Lyα) wavelength in their Lyα line profiles. We use the Lyα line profiles from our previous LAE theoretical model that includes radiative transfer in the interstellar and intergalactic mediums. We introduce a novel approach to measure the systemic redshift of LAEs from their Lyα line using neural networks. In detail, we assume that, for a fraction of the whole LAE population their systemic redshift is determined precisely through other spectral features. We then use this subset to train a neural network that predicts the Lyα wavelength given a Lyα line profile. We test two different training sets: i) the LAEs are selected homogeneously and ii) only the brightest LAE are selected. In comparison with previous approaches in the literature, our methodology improves significantly both accuracy and precision in determining the Lyα wavelength. In fact, after applying our algorithm in ideal Lyα line profiles, we recover the clustering unperturbed down to \(1\) cMpc/\(h\). Then, we test the performance of our methodology in realistic Lyα line profiles by downgrading their quality. The machine learning techniques work well even if the Lyα line profile quality is decreased considerably. We conclude that LAE surveys such as HETDEX would benefit from determining with high accuracy the systemic redshift of a subpopulation and applying our methodology to estimate the systemic redshift of the rest of the galaxy sample.

Key words: Radiative transfer – Intergalactic medium – ISM – High-redshift – Emission lines

1 INTRODUCTION

Since the detection of the first galaxies emitting Lyman-α radiation more than twenty years ago (e.g. Steidel et al. 1996; Hu et al. 1998; Rhoads et al. 2000; Malhotra & Rhoads 2002), Lyα radiation (with wavelength \(\sim 1215.68\) Å in rest frame) has been used as a successful tracer of the high redshift universe (Orlitová et al. 2018; Henry et al. 2018; Ouchi et al. 2008; Oyarzun et al. 2017; Matthee et al. 2017; Caruana et al. 2018), detecting galaxies even at the epoch of reionization (Sobral et al. 2015; Ouchi et al. 2018; Shibuya et al. 2018). Ongoing cosmological galaxy surveys, such as the Hobby-Eberly Telescope Dark Energy Experiment (Adams et al. 2011; Hill et al. 2008, HETDEX) and the Javalambre Physics of the Accelerating Universe Astrophysical Survey (Benitez et al. 2014, J-PAS), aim at unveil-

⋆ E-mail: sidgurung@cefca.es

© 2020 The Authors
ing the nature of the Dark Energy using LAEs at the high redshift Universe. One of the most useful tools to extract cosmological information from galaxy surveys is the galaxy clustering (e.g., Shoji et al. 2009). Therefore, understanding the spatial distribution of LAEs has become more important than ever before.

The complexity in understanding the LAE clustering resides in the radiative transfer of Lyα photons inside neutral hydrogen (Harrington 1973; Neufeld 1990). In first place, Lyα photons are emitted in the HII regions around OB-type stars. Then they have to cross the interstellar medium (ISM), then the circumgalactic medium (CGM) and the intergalactic medium (IGM) until they finally reach our observatories. In all these three mediums there is neutral hydrogen, and therefore, they are optically thick to Lyα radiation. Inside galaxies, it is commonly thought that Lyα photons escape through outflow that modify the Lyα flux and line profile (e.g., Ahn et al. 2000; Zheng & Miralda-Escudé 2002; Ahn 2003; Verhamme et al. 2006; Orsi et al. 2012; Gronke et al. 2016; Gurung-Lopez et al. 2018a). Then, in the CGM the Lyα radiation is spread around the galaxy creating the so-called Lyα halos (e.g., Zheng et al. 2010; Leclercq et al. 2017; Behrens et al. 2017). Finally, the Lyα radiation enters inside the IGM, where, to a first approximation, the radiation bluewards the Lyα wavelength is absorbed (e.g., Zheng et al. 2011; Laursen et al. 2011; Byrohl et al. 2019; Gurung-López et al. 2020).

Furthermore, the clustering property of LAEs can be sensitive to the selection function which is typically determined by the flux threshold. For example, Lyα radiation is very sensitive to dust. Therefore, galaxies with low metallicities are preferentially observed as LAEs (Sobral et al. 2018). This translate into a lower clustering amplitude (Gurung Lópe et al. 2018b), as galaxies with lower metallicity exhibit are hosted in smaller dark matter halos (Lacey et al. 2016, e.g.). Moreover, it has been pointed out that the large scale properties of the IGM might play a role on the selection function of LAEs (Zheng et al. 2010, 2011). This could distort the clustering of LAEs and reduce the accuracy of LAE cosmological surveys (Wyithe & Dijkstra 2011). However, there is still debate in the community whether if there is a large scale IGM coupling with the observed Lyα luminosity (Gurung-López et al. 2020) or not (Behrens et al. 2017). These facts contribute to the complexity of understanding the radiative transfer of the Lyα radiation and its impact on the clustering statistics.

A key challenge lies in determining the systemic redshift of an LAE only from their Lyα line profile. Due to the radiative transfer, the peak of the Lyα line rarely matches the Lyα wavelength. In Verhamme et al. (2018), authors probed that taking into account the radiative transfer in the ISM was crucial to improve the determination of the systemic redshift of LAEs. In their work, they used a relation between the shift of the line peak from Lyα and the width of the line to correct for the redshift due to the radiative transfer into the ISM. However, this recipe assumed the commonly used Thin Shell toy model for the outflow geometry. In fact the relation between the width of the line and peak shift depends strongly on the outflow geometry used (Gurung-Lopez et al. 2018a). More recently, Muzahid et al. (2019), linked the LAEs star formation rate to the line peak offset, also, getting better results than when assuming that the peak of the Lyα line is the Lyα wavelength.

Along a similar line, Byrohl et al. (2019) studied how the incorrect systemic redshift determination distorts the clustering of LAEs along the line of sight. In detail, an inaccurate Lyα wavelength determination is translated into an imprecise redshift, thus into an uncertainty in its radial position in redshift space. This causes that LAEs look like they are more spread along the line of sight than what they actually are. In fact, this shuffling in the LAE radial position can be interpreted as a extra random radial velocity dispersion component, which translates into an additional ‘Finger-of-God’ suppressing the apparent clustering along the line of sight. Furthermore, Byrohl et al. (2019) found that the clustering distortion was mitigated after correcting the shift of the peak with different recipes, such as the those in Verhamme et al. (2018). However, the developed recipes in the literature Verhamme et al. (2018); Byrohl et al. (2019); Muzahid et al. (2019) to estimate the Lyα wavelength have limitations as the dispersion of the estimated Lyα wavelength was around 1Å (in rest frame). Such a large scatter introduces significant distortions to the apparent clustering of LAEs on scales ~ 5 cMpc/h in the monopole and up to k ≥ 0.1 h/cMpc in Fourier-space, as we will explicitly show in this paper.

We propose a novel approach to determine the systemic redshift from the Lyα line profiles of LAEs using neural networks. This is motivated by the fact that, there must be information on the Lyα wavelength in an entire spectral range of Lyα line profiles. We explore whether or not a given survey that observes LAEs only through Lyα emission (e.g., HETDEX (Hill et al. 2008)) could benefit from acquiring a subsample with a systemic redshift (without using Lyα), for example, by Hα observations. Then, this subset could be used to train a neural network to predict the Lyα wavelength in the rest of the main LAE population. In this work, we train different neural networks using subsets of the Lyα line profiles computed by our model to predict their Lyα wavelength.

This work is part of a series of papers studying the impact of the Lyα radiative transfer on the observed properties of LAEs. In our first work (Gurung-López et al. 2019a), we focused on the Lyα RT taking place in the ISM. There, we found that LAEs are a very peculiar population that exhibits a tight balance between star formation rate and metallicity: Then, in our second work (Gurung-López et al. 2020), we implemented the Lyα in the IGM and we focused on the different selection effects introduced by it. There, we studied the impact in the clustering on large scales due to the IGM-LAE coupling. In this third work we analyze the properties of the Lyα stacked line profiles and study the impact on small scales of the miss-identification of the Lyα wavelength. Nonetheless, we emphasize that, we adopt this simulation set for a proof of concept, and that our approach can be generally applied to any LAE spectroscopic observation, in principle.

We briefly describe our model in section §2, while for a more extensive description we refer the reader to Gurung-López et al. (2019a, 2020). Then, in §7.1 we study the properties of the Lyα stacked line profiles. In section §4 we describe the different methodologies used in this work to identify the systemic redshift of LAEs from their observed Lyα line profiles. Then, in §5 we describe the effects of the Lyα
miss-identification in ideal line profiles. Meanwhile, in §6 we artificially reduce the quality of our Lyα line profiles analyze the effects of the Lyα miss-identification in realistic Lyα line profiles. Then, we discuss our result in §7. Finally, we make our conclusions in §8.

Throughout this paper, all the properties related to Lyα line profiles are given in the rest frame of the LAEs.

2 LAE THEORETICAL MODEL

In this work, we adopt the LAEs simulated with a semi-analytic model in Gurung-López et al. (2019a, 2020). In this section, let us briefly describe the LAE model but emphasize on modeling the spectrum around the Lyα emission. Our LAE model is based on four main ingredients:

• 1) The dark matter N-Body simulation, P-Millennium (Baugh et al. 2019), that imprints the hierarchical growth of structures in the LCDM scenario. This state-of-the-art cosmological simulation consists in 5040^3 dark matter particles with mass of 1.061 × 10^8 M_☉ h^{-1} distributed in a volume of L^3_{box} = (542.16 cMpc h^{-1})^3. P-Millennium uses cosmological parameters: H_0 = 67.77 km s^{-1} Mpc^{-1}, Ω_M = 0.693, Ω_M = 0.307, σ_8 = 0.8288, consistent with Planck Collaboration et al. (2016).

• 2) The model of galaxy formation and evolution, GALFORM (Cole et al. 2000; Lacey et al. 2016; Baugh et al. 2019). In short, GALFORM populates galaxies and gases within the dark matter halos and tracks their evolution through the cosmic history. GALFORM follows recipes to estimate a whole bunch of galaxy properties such as metallicity or star formation rate (SFR). These recipes are calibrated to fit several observables, such as, the optical and near infrared luminosity functions at z = 0 and its evolution up to z = 3, the HI mass function at z = 0, the sub-mm galaxy number counts and their redshift distributions among others. Galaxies in GALFORM exhibit two components: the disks, where the quiescent star formation takes place and the bulges, where the strong star formation bursts take place. Each of these pieces exhibit different properties (such as metallicity, etc).

• 3) The Lyα radiative transfer in the ISM is implemented through the Python open source code, FLaREDN (Gurung-López et al. 2019b). FLaREDN is based on a pre-computed grid of outflow models using LyαRT (Orsi et al. 2012), spawning a wide range in neutral hydrogen column density (N_H), outflow expansion velocity (V_{exp}) and dust optical depth (τ_d). By using different machine learning and multidimensional interpolation algorithms, FLaREDN predicts the Lyα escape fraction f_{esc} and line profile φ(λ) with high accuracy for different outflow geometries. Here, we will focus on the ‘Thin Shell’ and ‘Galactic Wind’ outflow geometries.

Through out this work, we define the escape fraction of a given medium as the ratio between the flux injected into the medium and the that emerges from it. For example, for the ISM escape fraction, f_{esc,ISM} = L_{Lyα,ISM}/L_{Lyα,0} where L_{Lyα,0} and L_{Lyα,ISM} are the intrinsic Lyα luminosity and the luminosity passing through the ISM region, respectively. Also, in our convention the line profile, φ(λ), is normalized as ∫ φ(λ)dλ = 1.

In practice, our model links the galaxy properties predicted by GALFORM to outflow features through simple recipes (Gurung-López et al. 2019a). In this way, each component (disk and bulge) in each galaxy has a different parameter set of {N_H, V_{exp}, τ_d} through their SFR, metallicity, cold gas mass and stellar mass. Then we use FLaREDN to compute a Lyα line profile and an escape fraction for each galaxy and component.

• 4) The radiative transfer in the IGM is implemented by computing the optical depth of Lyα photons in the line of sight (fixed along Z axis) between the observer and each galaxy. Therefore, the IGM transmission depends on the particular properties of the environment of each galaxy, and in particular, on the IGM density ρ, its density gradient a long the line of sight ∂ρ/∂z, the IGM line of sight velocity V_Z and its gradient ∂_Z V_Z.

Our model provides the Lyα line profile and luminosity are computed by convolving the Lyα line profile emerging from the ISM with the IGM transmission curve for each component of a galaxy. In practice, the line profile and luminosity are given by

\[ φ(λ) = \frac{L_{Lyα,Disk} φ_{Disk} + L_{Lyα,Bulge} \phi_{Bulge}}{L_{Lyα}} \]  

and

\[ L_{Lyα} = L_{Lyα,Disk} + L_{Lyα,Bulge} \]  

where the luminosity for each component is evaluated as

\[ L_{Lyα} = L_{Lyα,0} f_{esc} X \]  

with X = (Disk, Bulge). f_{esc} is the escape fraction from the IGM.

Furthermore, LAEs are defined as galaxies with a Lyα emission line exhibiting a high contrast to the galaxy continuum. It is usually found in the literature that for a galaxy to be considered an LAE, it must exhibit a rest frame equivalent width EW_0 > 20Å (e.g. Gronwall et al. 2007; Konno et al. 2018). In this work we follow this criteria. In particular, we compute the EW_0 for each galaxy as a function of its continuum luminosity per unit of wavelength around Lyα wavelength L_c as

\[ EW_0 = L_{Lyα}/L_c \]  

where L_c is directly provided by GALFORM and it is based on the full evolution of the stellar population given a galaxy.

Finally, the free parameters in the model, which depend on the outflow geometry model, are adjusted so that the simulated LAEs reproduce the observed Lyα luminosity function at their corresponding redshift.

3 STACKED LYMAN-α LINE PROFILES

Since the main goal of this paper is to study the determination of the systemic redshift of LAEs from their Lyα line profile, we present here the detailed properties of the stacked Lyα line profile. First we focus on how Lyα radiative transfer impacts the observed Lyα stacked line profile in our model. Then we discuss the Lyα line profiles as a function of different galaxy and IGM properties to understand how they influence the stacked line profile.

MNRA 600, 1–24 (2020)
Figure 1. Stacked Lyα line profile at redshift 2.2, 3.0, 5.7 from left to right. The Thin Shell geometry is displayed in thick green lines and the Galactic Wind in blue. Our default models, including the IGM Lyα absorption, are shown in thick lines while, models without the IGM implementation are shown in thin lines. The stacked line profiles are normalized so that their global maximums match unity.

Figure 2. Break down of the Lyα stacked line profiles at $z = 3.0$ as a function of total star formation rate, Lyα rest frame equivalent width, Lyα luminosity and stellar mass from left to right. Top panels show the model with the Thin Shell outflow geometry, while bottom panels show the model using the Galactic Wind geometry. We rank our main LAE population by each of these galaxy properties and split it by the percentiles 33 ($Q(33)$) and 66 ($Q(66)$). The stacked line profiles are color coded according to the samples they are displaying. The lighter lines show the samples with the lowest values of the galaxy property $X$ ($Q(0) < X < Q(33)$). The darkest lines show the samples with the highest values of each galaxy property ($Q(66) < X < Q(100)$). Meanwhile, the intermediate values ($Q(33) < X < Q(66)$) are shown in intermediate colors. The Lyα stack profile of each subpopulation, $\langle \phi \rangle_Q$ is normalized to the maximum of the stacked line profile of the complete LAE population, $\langle \phi \rangle_{\text{max}}$. 
Throughout this work we compute the stacked Lyα line profiles in a consistent manner. First, we normalize all the Lyα line profiles so that they have an area of unity, i.e., compute $\phi(\lambda)$. Then we evaluate the stacked line profile, $\langle \phi(\lambda) \rangle$, as the median of the line profile collection.

### 3.1 The impact of radiative transfer

In Fig. 1 we compare the stacked Lyα profile $\langle \phi(\lambda) \rangle$ before (solid thin) and after (thick solid) being processed by the IGM. In general, at low redshift ($z = 2.2$, $3.0$), the IGM tends to absorb blue photons, i.e., $\lambda < L_{\text{Ly}\alpha}$, while it does not affect less for redder photons. Meanwhile, at $z = 5.7$, the IGM optical depth to Lyα is much greater and it affects to the stacked line profile up to wavelength 2Å redder than Lyα.

Furthermore, the impact of the IGM in the stacked line profiles is different for each geometry. We find that at $z = 2.2$ and $z = 3.0$ the IGM affects more the Thin Shell geometry than the Galactic Wind. In particular, the blue peak which is present before the line profiles are processed by the IGM (especially at $z = 3.0$), is mostly vanished after the RT in the IGM. Meanwhile, at $z = 5.7$ the IGM affects similarly both cases. This is consistent with the values of the Lyα IGM escape fraction in our model (see Fig. 10 and 11 in Gurung-López et al. (2020)). These findings originate from the differences in the family of line profiles generated by each outflow geometry, as the IGM absorption depends on the wavelength of the photons, as discussed in detail in Gurung-López et al. (2020).

Our model predicts different shape of $\langle \phi(\lambda) \rangle$ for the Thin Shell and Galactic Wind models at all the redshift bins studied in this work. In general, we find that the stacked line profile for the Thin Shell model is bluer than that for Galactic Wind. Meanwhile, at each different epoch, the stacked line profiles for two outflow geometry models differ in unique fashions. For example, at $z = 2.2$, the peaks of $\langle \phi \rangle$ of both geometries match, but the Thin Shell exhibits a broader $\langle \phi \rangle$ ($\sim 3\AA$) than the Galactic Wind ($\sim 1.5\AA$). At $z = 3.0$, the width of the stacked profiles are comparable between both geometries, but the Galactic Wind profile is more redshifted than the Thin Shell one. Another difference between our two outflow models is that at redshift of 2.2 and 3.0 the Thin Shell exhibits a weak peak bluer than Lyα, while the Galactic Wind lacks this blue peak. Finally, at $z = 5.7$, the Thin Shell profile is broader than the Galactic Wind one.

The differences between the ‘Thin Shell’ and ‘Galactic Wind’ outflow geometries arise due to two facts. First, the distribution of the ISM parameters, $\{V_{\text{exp}}, T_{\alpha}, N_{H}\}$ of the LAEs are different for both geometries (see Fig. A1 and A2 of Gurung-López et al. 2020). Second, the radiative transfer in each geometry leads to a different line profile, even for the same parameter set of $\{V_{\text{exp}}, T_{\alpha}, N_{H}\}$. For example, the ‘Thin Shell’ geometry model is more prone to exhibit a blue peak than the ‘Galactic Wind’ model (Gurung-López et al. 2019b), as we see in the stacked profile at redshifts 2.2 and, in particular, at $z = 3.0$.

In summary, the radiative transfer impacts the shape of the stacked profile including the peak position and the width from the ISM to IGM scales in a non-trivial manner.

### 3.2 Dependence of the stacked profile on galaxy and environmental properties

So far we have considered the stacked profile for all LAEs in our simulation. In this section we split the Lyα stacked line profile according to different galaxy and IGM properties. To this end, for a given galaxy or IGM property X, we compute the percentiles 33.33 ($Q(33)$) and 66.66 ($Q(66)$). Then, we split the LAEs into three subsamples of the same size containing the galaxies with the lowest ($Q(0) < X < Q(33)$), intermediate ($Q(33) < X < Q(66)$) and the greatest ($Q(66) < X < Q(100)$) values of $X$.

#### 3.2.1 Imprints of the galaxy properties

In Fig. 2 we split the Lyα stacked line profile as a function of the star formation rate, rest frame Lyα equivalent width, Lyα luminosity and stellar mass (from left to right). The Lyα stack line profile of each subpopulation is normalized to the maximum of the stacked line profile of the complete LAE sample, $\langle \phi \rangle_{\text{max}}$. Here we show the snapshot at redshift 3.0 only, but have confirmed that the other two redshift bins ($z = 2.2$ and 5.7) exhibit similar trends.

Overall our model predicts that the stacked Lyα line profile properties depend on the galaxy properties. For instance, the peak of the Lyα line profile correlates positively with the SFR and $L_{\text{Ly}\alpha}$ for both outflow geometries. The peak of the line profile anti-correlates with EW$_0$ for both outflow geometries. Interestingly, the dependence of the stacked Lyα line profile on the stellar mass behaves differently for two outflow geometries; In Thin Shell the peak anti-correlates with $M_*$, while there is no apparent trend in Galactic Wind.

We also find that, not only the peak position, but also the shape of the stacked line profile changes through the dynamical range of these galaxy properties. In particular, we find that for both outflow geometries, the stacked line profile becomes broader at higher SFR and $L_{\text{Ly}\alpha}$ values, while it shrinks for high values of EW$_0$. Meanwhile, when the ‘Thin Shell’ is implemented, increasing the stellar mass leads to broader stacked profiles. In contrast, the width of the stacked line profile using the ‘Galactic Wind’ remains constant through the stellar mass dynamical range.

These non-trivial differences between Thin Shell and Galactic Wind are consequences of the complicated interplay between the properties of galaxies and outflows (see equations XX and XX in Gurung-López et al. (2020)). This highlights the importance of having different outflow models to model the RT in the ISM.

#### 3.2.2 Imprints of the IGM

Here we study how the different large scale IGM properties change the observed Lyα profile. In order to do so, we split the Lyα stacked line profile $\langle \phi \rangle$ by IGM properties for both outflow geometries. In Fig. 3 we show the difference between the stacked line profile of the full LAE population and the split samples. The IGM properties used for dividing the LAE population are density $\rho$, density gradient along the line of sight $\nabla_{\perp} \rho$, velocity along the line of sight $v_{\perp}$ and its gradient along the line of sight $\nabla_{\perp} v_{\perp}$. These properties were computed in a regular grid of cubic cells of 2 cMpc/h side, as in Gurung-López et al. (2020). Here
we focus on the snapshot at \( z = 5.7 \), where IGM is optically thicker to Ly\( \alpha \) photons than at \( z = 2.2 \) and \( z = 3.0 \). Note that we also find the same trends at \( z = 2.2 \) and 3.0, but with a lower amplitude.

We find that, although differences are tiny (\( \lesssim 10\% \) compared to \( \langle \phi \rangle_{\max} \)), both outflow geometries exhibit the same trends with a clear dependency of the IGM properties. This suggests that there is a smooth dependence between the IGM properties and the stacked line profile. For example, the higher the IGM density is, the more flux is absorbed at bluer wavelengths. This causes that the observed stacked line profile is slightly more redshifted in high IGM density regions. Also, our model predicts that LAEs located in regions with high \( \partial Z \rho \), \( V_Z \) and \( \partial Z V_Z \) exhibit a bluer stacked line profile, while the opposite is true for low values of these IGM properties.

The trends in the stacked line profile are in agreement with our previous work \citep{gurung2020}, where the IGM transmission generally anti-correlates with \( \rho \) and positively correlates with \( \partial Z \rho \), \( V_Z \) and \( \partial Z V_Z \) \citep{gurung2020}. We also showed that, the bluer the wavelength around Ly\( \alpha \), the more sensitive the line profile is to IGM absorption \citep[see Fig.5 in][]{gurung2020}. Combining these two facts, the LAEs lying in regions with lower IGM transmission will exhibit a redder line profile than LAEs lying in regions with higher transmission, which is consistent with Fig. 3.

4 DETERMINING THE REDSHIFT OF LAES

As we showed in the previous section, the wavelength of photons initially emitted at the Ly\( \alpha \) wavelength changes as they travel through the ISM and the IGM. In this way, the Ly\( \alpha \) line profiles are modified in a non-trivial way \citep{zheng2011}. In general, in each Ly\( \alpha \) line profile, the true Ly\( \alpha \) wavelength (\( \lambda_{\alpha} \)) and the wavelength set as Ly\( \alpha \) (\( \lambda_{\alpha}^{\text{Obs}} \)) can differ, as we show below. Thus, each Ly\( \alpha \) line profile is shifted by \( \Delta \lambda = \lambda_{\alpha}^{\text{Obs}} - \lambda_{\alpha} \) along wavelength. In other words, the Ly\( \alpha \) line profiles transforms as

\[
\phi(\lambda) \rightarrow \phi(\lambda + \Delta \lambda), \tag{5}
\]

where \( \phi(\lambda + \Delta \lambda) \) is the observed Ly\( \alpha \) line profile for given a \( \lambda_{\alpha} \).

In the following we introduce the different methods that we use through this work to find the Ly\( \alpha \) wavelength (\( \lambda_{\alpha}^{\text{Obs}} \)) directly from the Ly\( \alpha \) line profile. First, in §4.1, we describe two different methods to retrieve \( \lambda_{\alpha}^{\text{Obs}} \) that have been already used in the literature. These algorithms depend on line profile characteristics such as the width and the position of the global maximum of the Ly\( \alpha \) line. Then, in §4.2, we introduce a novel method that makes use of the full Ly\( \alpha \) line profile in order to predict \( \lambda_{\alpha}^{\text{Obs}} \) through neural networks.
4.1 Standard methodologies

- **GM** (Global Maximum): This is the simplest method to assign a Lyα wavelength. Basically, the position of the global maximum (\(\lambda_{\text{Ly},\text{Max}}\)) is set as the Lyα wavelength, i.e.,

\[
\lambda_{\text{Ly},\text{obs}} = \lambda_{\text{Ly},\text{Max}}.
\]

4.1.1 Theoretical framework

For all the LAE population while the true escape channel has significative differences with the assumed model.

- **GM-F** : This method takes into account that the Lyα photon tend to be redshifted as they escape the galaxies through outflows. As a result, the position of the red peak is shifted from the Lyα frequency. This shift depends on the outflow properties and can be related to the FWHM of the red peak Verhamme et al. (2018); Gurung-López et al. (2019b). As we have shown (see §7.1) the Lyα line profiles predicted by our model are clearly dominated by a prominent red peak and only a faint blue peak is found. Therefore, \(\lambda_{\text{Ly},\text{Max}}\) and the maximum of the read peak matches. Thus, in this method, we compute the Lyα wavelength as

\[
\lambda_{\text{Ly},\text{obs}} = \lambda_{\text{Ly},\text{Max}} - \text{FWHM}_\text{Red}.
\]

where FWHM\text{Red} is the FWHM of the red peak of the Lyα line profile. This relation is compatible with the observational results found in Verhamme et al. (2018), whom first suggested this kind of correction. However, this trend depends strongly in the outflow geometry that is assumed Gurung-López et al. (2019b). In particular, this relation works well for the Thin Shell geometry, while the Galactic Wind deviates slightly from it. In this work we use Eq.7 for both, the Thin Shell and the Galactic Wind in order to qualitatively study the impact in the clustering and stacking of using a relation that is slight off. This mimics the observational framework in which an outflow geometry is assume for all the LAE population while the true escape channel has some differences with the assumed model.

4.2 Neural networks

We propose a novel method to determine the systemic redshift of LAEs from their whole Lyα line profile through a neural network. In this section we explain all the ingredients of the neural networks implemented in this work. First we describe the architecture and the different training sample. Then we study how the full line profile helps us determine the Lyα wavelength.

4.2.1 Neural network architecture

The architecture of the neural network in this work consists in an input layer, a single hidden layer and an output layer. We remark that this work does not focus on finding the best architecture to solve the Lyα wavelength determination problem, as this would depend on the Lyα observation characteristics (e.g. spectral resolution, signal to noise ratio, etc). Instead, we adopt this simple architecture as a proof of concept.

In this work we seek for an algorithm that could be replicated in observational experiments. With this goal in mind, for a given Lyα line profile \(\phi(l)\), we set as input \(\phi(\Delta \lambda_{\text{obs}})\), where have mapped

\[
\lambda \rightarrow \Delta \lambda_{\text{obs}} = \lambda - \lambda_{\text{Ly},\text{Max}}.
\]

In this way, the global maximum of the Lyα line profile is always centered at \(\Delta \lambda_{\text{obs}} = 0\), which can be easily replicated in observational experiments. Additionally, we rescale each individual \(\phi(\Delta \lambda_{\text{obs}})\) in a way that the minimum of the line profile is 0 and the maximum is 1.

4.2.2 Training sets

Throughout this work we implement two neural networks. Both of them have the same architecture, but distinct training sets:

- **NN:Uniform** : The training sample is randomly selected from the whole LAE population. In other words, there is no dependence on any LAE property.
- **NN:Bright** : The training sample is constructed only by the brightest LAEs. In practice, we rank our LAE population by their Lyα luminosity. Then, we split the LAE population in two and use for training the brightest subset.

The motivation behind each training set is different. On one hand, NN:Uniform represents an ideal scenario where, in a given survey, a subset of the whole LAE population is homogeneously selected and re-observed at a wavelength range that allows the measurement of spectral features (other than the Lyα line) and the assignment of the true systemic rest frame. However, it is, in general, challenging to obtain systematic redshifts of such a homogeneous subsample, since the flux of other spectral features tend to be less prominent than Lyα (Trainor et al. 2015). On the other hand, NN:Bright
We remark that the mean and median of \( \lambda \) for each combination of redshift, outflow geometry and spectral quality (see NN:Bright) for each combination of redshift, outflow geometry and spectral quality (see NN:Bright) for each combination of redshift, outflow geometry and spectral quality (see NN:Bright).

Figure 5. Distribution of the difference between the assigned Ly\( \alpha \) wavelength and the intrinsic \( \lambda_{\text{Ly} \alpha} \) for each of the different Ly\( \alpha \) identification algorithms. The GM algorithm is displayed in blue, GM-F in green, NN:Bright in orange and NN:Uniform in yellow. Each column shows a different redshift bin (2.2, 3.0 and 5.7 from left to right). The models using the Thin Shell (Galactic Wind) geometry are shown in the top (bottom) panels.

is designed to study how well the LAE redshift determination works even only with the brightest LAEs re-observed at other wavelength, which is closer to a realistic situation than the NN:Uniform case.

We train both neural networks (NN:Uniform and NN:Bright) for each combination of redshift, outflow geometry and spectral quality (see §5).

The performance of the neural networks are linked to the size of the training sample. In general, the larger the training sample, the more accurate the neural network becomes. In Fig. 4 we show the accuracy of our neural networks (NN:Uniform and NN:Bright) for the different outflow geometries at redshift 3.0 as a function of the number density of LAEs for the training, \( n_{\text{LAE}}^{\text{Training}} \). Here, we use the standard deviation of \( \lambda_{\text{Ly} \alpha}^{\text{Obs}} - \lambda_{\text{Ly} \alpha} \) (noted as \( \sigma(\Delta \lambda) \)) to quantify the quality of the neural network, as the clustering of LAEs is sensitive to the distribution of \( \Delta \lambda \) (Byrohl et al. 2019).

We remark that the mean and median of \( \lambda_{\text{Ly} \alpha}^{\text{Obs}} - \lambda_{\text{Ly} \alpha} \) are, in general, one order of magnitude smaller than \( \sigma(\Delta \lambda) \), and they have little impact on the two-point statistics of the auto-correlation function.

As we show in Fig. 4, the neural networks are able to assign a \( \lambda_{\text{Ly} \alpha}^{\text{Obs}} \) close to \( \lambda_{\text{Ly} \alpha} \). Overall, the accuracy increases as the training sample size is increased. However, the NN:Uniform and NN:Bright algorithms behave slightly different. On one hand, in the NN:Uniform algorithms, \( \sigma(\Delta \lambda) \) decreases until \( n_{\text{LAE}}^{\text{Training}} \sim 4 \times 10^{-4} \text{ (cMpc/h)}^{-3} \) (~10% of the total LAE sample), where it reaches a plateau around 0.1Å. This means that at this value of \( n_{\text{LAE}}^{\text{Training}} \), the training sample is big enough to cover the full variety of Ly\( \alpha \) line profiles. Thus, adding more galaxies to the training sample beyond \( 10^{-4} \text{ (cMpc/h)}^{-3} \) add little information, leaving the accuracy constant. On the other hand, the performance of the NN:Bright algorithm is worse at low \( n_{\text{LAE}}^{\text{Training}} \). Meanwhile, the NN:Bright converge to the NN:Uniform accuracy when \( n_{\text{LAE}}^{\text{Training}} \sim 2 \times 10^{-3} \text{ (cMpc/h)}^{-3} \) (~50% of the total sample). This is because the Ly\( \alpha \) line profile depends on galaxy and IGM properties (see Figs. 2 and 3). In particular, the Ly\( \alpha \) line profile of the bright LAEs is more redshifted than the Ly\( \alpha \) line profiles of the faint LAEs. Hence, the training sample in the NN:Bright is biased towards redshifted Ly\( \alpha \) line profiles and it does not contain typical line profiles from faint LAEs. This reduces the accuracy at low \( n_{\text{LAE}}^{\text{Training}} \) in comparison to NN:Uniform, which makes an uniform selection on \( L_{\text{Ly} \alpha} \). Then, for larger values of \( n_{\text{LAE}}^{\text{Training}} \), the training samples of the NN:Uniform and NN:Bright become more similar, which makes them converge to the same accuracy. Finally, the line profiles generated using the Galactic Wind outflow geometry seem slightly more complex than the Thin Shell...
counterparts, which results, typically, in a better accuracy for the Thin Shell geometry.

From now on, we fix the number density of the training sample to \( n_\text{LAE}^{\text{Training}} = 4 \times 10^{-4} \text{(cMpc/h)}^3 \). We have chosen this value of \( n_\text{LAE}^{\text{Training}} \) for two main reasons: a) the information of the training sample of the NN:Uniform saturates and increasing \( n_\text{LAE}^{\text{Training}} \) does not add new information to it, and b) the NN:Bright has not yet converged to the NN:Uniform.

5 THE EFFECTS OF THE Ly\(\alpha\) WAVELENGTH DETERMINATION IN IDEAL LINE PROFILES.

Hereafter, we study how the miss-identification of the Ly\(\alpha\) wavelength modifies the clustering. In order to understand the physical consequences, in this section, we rather focus on an ideal case where we ignore binning artifacts and the instrumental noise in a LAE spectrum. We will consider more realistic situations in next section.

5.1 Algorithm performances in ideal Ly\(\alpha\) line profiles

In this section we compare the performance of the four methodologies to determine \( \Delta_{\text{Ly} \alpha} \) from a Ly\(\alpha\) line profile. In Fig. 5 we show the probability distribution function (PDF) of the deviation of \( \alpha_{\text{Obs}} \) from \( \alpha_{\text{Ly} \alpha} \) (\( \Delta \)) for the different redshifts, outflow geometries and algorithms to determine \( \Delta_{\text{Ly} \alpha} \). Also, we list the mean (\( \mu(\Delta) \)) and standard deviation (\( \sigma(\Delta) \)) of these distribution in Tab.1. Overall, the algorithms using neural networks outperform the standard algorithms (GM and GM-F). We find that the best methodology to retrieve \( \Delta_{\text{Ly} \alpha} \) is NN:Uniform, as the mean and standard deviation of \( \Delta \) are the smallest at all redshifts and outflow geometries. In detail, for all our models \( \sigma \) is below 0.1A and \( \mu(\Delta) \) is lower than 0.01A. The NN:Uniform is followed closely by the NN:Bright, which also exhibits a great performance, with \( \sigma(\Delta) \sim 0.15 \) and \( \mu(\Delta) < 0.01 \AA \).

Regarding the standard methodologies, GM-F performs better than GM. If we focus on GM (blue), we find that the performance is similar for both outflow geometries. The mean of the \( \Delta \) distribution is shifted \( \sim 2\Delta \) redwards \( \Delta_{\text{Ly} \alpha} \). This is a direct consequence of the Ly\(\alpha\) RT, as the photons get redshifted when they travel through neutral hydrogen. Then, as GM takes the global maximum of the Ly\(\alpha\) line profile as \( \alpha_{\text{Obs}} \), the whole \( \Delta \) is systematically redshifted. In fact, we find that for the GM algorithm, the peak of the \( \Delta \) distribution is located at the same position than the peak of their corresponding stacked line profiles (see Fig. 1). Moreover, for the Thin Shell at \( z = 5.7 \) the \( \Delta \) distribution of the GM algorithm exhibits a double peak shape, and the same position than the peaks present in the Ly\(\alpha\) stacked line profile of that model.

The performance of GM-F (green) depends strongly on the outflow geometry and redshift. On one hand, at low redshift (\( z = 2.2 \) and \( z = 3.0 \)), GM-F performs better than GM in both outflow geometries, as the GM-F distributions of \( \Delta \) are thinner and closer to \( \Delta_{\text{Ly} \alpha} \), which is a consequence of the different relation between \( \Delta_{\text{Ly} \alpha} \) and the true Ly\(\alpha\) line profile. In particular, \( \mu(\Delta) \) is close to zero and \( \sigma(\Delta) \) is smaller than those of GM. However, the GM-F performance in the Thin Shell is better than in the Galactic Wind. In particular, \( \mu(\Delta) \) is close to zero for the Thin Shell, while it is \( \sim 0.5\AA \) in the Galactic Wind. This is a consequence of the different relation between \( \Delta_{\text{Ly} \alpha, \text{Max}} \) and \( \text{FWHM}_{\text{Red}} \) in the Thin Shell and Galactic Wind outflow geometries (see Fig. 4 in Gurung-López et al. (2019b)). On the other hand, at \( z = 5.7 \), the \( \Delta \) distribution becomes bimodal, with a second less prominent peak centered around \( \Delta \sim -2\Delta \) (see Fig. 1). This is caused by the IGM modifying the Ly\(\alpha\) line profile in such a way that the \( \text{FWHM}_{\text{Red}} \) of the observed Ly\(\alpha\) line profile is larger than the initial. As a consequence, GM-F overcorrects the shift of the Ly\(\alpha\) peak, causing the peak in the \( \Delta \) distribution at \( \Delta < 0 \).

5.2 Impact on the redshift-space clustering of LAEs

The misidentification of the Ly\(\alpha\) wavelength has a non-negligible impact on the three-dimensional clustering of LAEs samples that rely only on their Ly\(\alpha\) line profile to determine their redshift, and hence, their radial position (Byrohl et al. 2019). On the other hand, the measured redshift of LAEs have three main contributions: a) the geometric redshift given by the

Table 1. Mean (\( \mu(\Delta) \)) and standard deviation (\( \sigma(\Delta) \)) of the difference between the Ly\(\alpha\) assigned as Ly\(\alpha\) and the true Ly\(\alpha\) frequency for the different Ly\(\alpha\) identification algorithms, redshifts and outflow geometries.

| Redshift | Geometry       | Algorithm      | \( \mu(\Delta) \) | \( \sigma(\Delta) \) |
|----------|----------------|----------------|-------------------|----------------------|
| 2.2      | Thin Shell     | GM             | 1.822             | 0.66                 |
|          |                | GM-F           | -0.3059           | 0.45                 |
|          |                | NN:Uniform     | 0.0032            | 0.07                 |
|          |                | NN:Bright      | -0.0087           | 0.08                 |
|          | Galactic Wind  | GM             | 1.8289            | 0.89                 |
|          |                | GM-F           | 0.4394            | 0.36                 |
|          |                | NN:Uniform     | 0.0022            | 0.11                 |
|          |                | NN:Bright      | -0.156            | 0.17                 |
| 3.0      | Thin Shell     | GM             | 1.4747            | 0.99                 |
|          |                | GM-F           | 0.0251            | 0.79                 |
|          |                | NN:Uniform     | 0.0303            | 0.08                 |
|          |                | NN:Bright      | 0.0292            | 0.1                  |
|          | Galactic Wind  | GM             | 1.9134            | 0.99                 |
|          |                | GM-F           | 0.4757            | 0.34                 |
|          |                | NN:Uniform     | 0.0007            | 0.11                 |
|          |                | NN:Bright      | 0.0112            | 0.16                 |
| 5.7      | Thin Shell     | GM             | 1.5134            | 1.22                 |
|          |                | GM-F           | -0.7141           | 1.44                 |
|          |                | NN:Uniform     | 0.0116            | 0.1                  |
|          |                | NN:Bright      | 0.0094            | 0.12                 |
|          | Galactic Wind  | GM             | 0.8751            | 0.97                 |
|          |                | GM-F           | -0.5247           | 1.09                 |
|          |                | NN:Uniform     | -0.0132           | 0.09                 |
|          |                | NN:Bright      | -0.0185           | 0.17                 |
Hubble flow, b) the redshift or blueshift given by the peculiar velocity of the galaxy along the line of sight (Kaiser 1987), normally dubbed RSD, and c) a redshift or blueshift rising from the Lyα wavelength misidentification as in Byrohl et al. (2019), i.e., $\Delta t \neq 0$.

In order to characterize the clustering of a galaxy population, it is useful to define the galaxy overdensity field

$$\delta_g = \frac{n_g(\vec{x})}{\langle n_g \rangle} - 1,$$

where $n_g(\vec{x})$ is the number density galaxies at the position $\vec{x}$, and $\langle n_g \rangle$ is its average value. Then the two-point correlation function (2PCF), $\xi$, is defined as

$$1 + \xi(\vec{r}) = \langle [1 + \delta_g(\vec{x})][1 + \delta_g(\vec{x} + \vec{r})] \rangle$$

where $\vec{r}$ is the pair vector between two points separated a distance $r$. We also consider the power spectrum $P_g(k)$, which is the Fourier transform of the 2PCF, i.e.,

$$P_g(k) = \frac{4\pi}{(2\pi)^3} \delta_D(\vec{k} + \vec{k}^*) = \delta_D(\vec{k}) \delta_D(\vec{k}^*)$$

where $k$ is the wavenumber, and $\delta_D(\vec{k})$ is the Dirac delta function.

Throughout this work we will focus on the clustering in redshift space. To incorporate the three redshift contributions to our clustering analysis, we recompute the position of our LAEs in redshift space as

$$\vec{x} = \vec{r} + \frac{V_{\text{LoS}} + \Delta V_{\text{Lyα}}}{a(z)H(z)} \hat{Z},$$

where we take $Z$ as the direction of the line of sight, assuming the global plain-parallel approximation (Beutler et al. 2014), $V_{\text{LoS}}$ is the velocity along the line of sight of the galaxy, $a(z)$ and $H(z)$ are the scale factor and the Hubble parameter at redshift $z$, respectively. Additionally,

$$\Delta V_{\text{Lyα}} = c \left(1 - \frac{\lambda_{\text{Lyα}}}{\lambda_{\text{Lyα}}^{\text{Obs}}}\right),$$

where $c$ is the speed of light. Finally, there is a fraction of the LAE population that are shifted outside of the box after transforming their line position to redshift space, i.e., including the contributions of $V_{\text{LoS}}$ and $\Delta V_{\text{Lyα}}$. For these galaxies we assume that our simulation box is periodic along the line of sight. Therefore, galaxies with $\pi < 0$, they are assigned $\pi = L_{\text{Box}} + \pi$ and galaxies with $\pi > L_{\text{Box}}$, they are assigned $\pi = \pi - L_{\text{Box}}$.

### 5.2.1 Clustering damping in Fourier space

First, we focus on the impact of the Lyα misidentification in the power spectrum. In Byrohl et al. (2019), the authors analytically showed that, due to the Lyα misidentification, the amplitude of the power spectrum is reduced along the line of sight at large wavenumber ($k$) values, i.e., at small scales. We define the damping of the power spectrum due to the Lyα miss-identification as in Byrohl et al. (2019), i.e.,

$$D_{\text{RT}}(k) = \frac{P_{\text{RT}}(k)}{P(k)} = |FT(\text{PDF}(\Delta t))|^2.$$
where $P(k)$ is the intrinsic redshift-space power spectrum of the LAE population, i.e., setting $\Delta V_{\lambda_{\alpha}} = 0$, in Eq. (12). Also, $DRT(k)$ is the power spectrum of the LAE sample after including the displacement along the line of sight due to the Ly$\alpha$ misidentification. Finally, the Fourier transformation is indicated as $FT$. We estimate the power spectrum from the simulated LAEs by making use of the Fast Fourier Transform (FFT). We set the number of grids as 512$^3$ with which the Nyquist wavenumber is $k_{\text{Nyq}} \sim 3h\text{Mpc}^{-1}$. The last equality holds only if i) the moments of $PDF(\Delta l)$ are scale independent and ii) $\Delta l$ is uncorrelated with the large-scale density and velocity fields (Byroh et al. 2019).

In Fig. 6 we show the damping in the power spectrum for our different Ly$\alpha$ wavelength recovering algorithms and multiple models. In general, we find that the amplitude of the power spectrum including the Ly$\alpha$ misidentification is lower than the intrinsic power spectrum at the scales relevant to the BAO and RSD measurements, as $DRT < 1$ at $0.1 < k_{\parallel} [h\text{Mpc}^{-1}] < 1$. In particular, the impact is greater on smaller scales (larger $k_{\parallel}$), while at large enough scales it disappears. This damping can be interpreted as the Finger-of-God effect and matches the results from Byroh et al. (2019), although the detailed suppressions behave differently as the PDFs are different.

$DRT$ has been computed in two different ways: i) computing the power spectra directly from the LAE positions in our simulation box (solid lines) and ii) by computing the Fourier transform of the one point PDF of $\Delta l$ (open circle points). We compare the two methods only for the GM in Fig. 6, and confirm that they are in a good agreement. This ensures that the misidentification is uncorrelated with the large-scale density or velocity field, while there is a small hint of the IGM interaction at $k_{\parallel} \sim 1$ for the Galactic Wind at $z = 5.7$. We have checked similar results for the other algorithms, and hence omitted them in the figure.

We find that, as naively expected, the algorithms with a higher accuracy for recovering the Ly$\alpha$ wavelength from the line profile show a shallower damping of the power spectrum. In particular, the NN:Uniform is the algorithm that is the least affected by the Ly$\alpha$ wavelength misidentification. In fact, the recovered power spectrum agrees at the 1% level up to $k_{\parallel} \sim 1$ for both outflow geometries and at all redshifts. The second best performance is achieved by the NN:Bright, which exhibits up to ~ 0.2 decreases in the power spectrum amplitude in the Galactic Wind, while in the Thin Shell, the damping is slight stronger than in the NN:Uniform. The GM and GM-F algorithms are heavily affected by the Ly$\alpha$ misidentification, as the amplitude of the power spectrum decreases dramatically on small scales at all redshifts and outflow geometries. In particular, GM-F is less affected than GM at redshift 2.2 and 3.0, while at $z = 5.7$, the performance of GM-F is comparable (in the Thin Shell) or slight worst (in the Galactic Wind) than GM.

5.2.2 Impact on the 2D clustering in configuration space

In this section we explore the 2PCF to illustrate qualitatively the clustering distortion produced by the misidentification of the Ly$\alpha$ wavelength. Later, we will further quantify the anisotropic distortions using the Legendre multipole moments.

We estimate 2PCF with the standard Landy-Szalay estimator (Landy & Szalay 1993). Fig. 7 shows the clustering divided into parallel and perpendicular to the line of sight components for our different algorithms. In particular, we display the model at redshift 2.2 using the Thin Shell geometry, but similar results are found for the other models too. To compare the performance of the algorithms, we show the contours with the clustering amplitude of $\xi(r_{\perp}, r_{\parallel}) \approx 10^{-1.0}$ and $10^{0.0}$ for no Ly$\alpha$ wavelength misidentification (solid) and for each algorithm (dashed). Overall, the misidentification of the Ly$\alpha$ wavelength causes an elongation of the LAE clustering along the line of sight. This elongation is more prominent for the algorithms with worst performance recovering the Ly$\alpha$ wavelength. In concordance with our previous findings, the GM and GM-F algorithm fail to recover the intrinsic redshift-space clustering of LAEs. Meanwhile, NN:Uniform and NN:Bright achieve almost a perfect recovery of the 2PCF.

5.2.3 Impact on the monopole

The multipole 2PCF is given by

$$\xi\ell(s) = \frac{2\ell + 1}{2} \int_{-1}^{1} d\mu \xi(s, \mu) L_\ell(\mu),$$

where $\ell$ is the multipole degree and $L_\ell$ is the Legendre’s polynomial of degree $\ell$ and $\mu$ is the cosine between the line of sight and the separation vector of a galaxy pair.

In Fig. 8 we show the ratio of $\xi_0$, the monopole ($\ell = 0$) of our LAE populations by using our different algorithms to determine the Ly$\alpha$ wavelength, and $\xi_0$, the monopole when the Ly$\alpha$ wavelength is identified perfectly, $\Delta l = 0$. Overall we find that the method used to determine the Ly$\alpha$ wavelength change the retrieved monopole of LAEs at scales below $10 \text{Mpc}/h$ for both outflow geometries at all redshifts. Meanwhile, the large scale clustering ($> 10 \text{Mpc}/h$) remain unchanged. In particular, at large scales, the monopole of the different algorithms converged to the intrinsic one with nearly no difference among them.

The clustering for the traditional algorithms (GM and GM-F) is suppressed at small scales. In contrast, the measured monopole using both our neural networks (NN:Uniform and NN:Bright) match extraordinary well the intrinsic clustering of LAEs at all scales. These differences in the monopole between the standard approaches and the neural networks are driven by the much better performance of NN:Uniform and NN:Bright when determining the Ly$\alpha$ frequency (see Fig. 5). In particular, the large dispersion of $\Delta l$ given by GM and GM-F translates in to a large scatter of $\Delta V_{\lambda_{\alpha}}$. Which means that, the position of the galaxies in redshift space are quite spread along the line of sight with respect to their original position in redshift space. This dilutes the clustering on small scales along the line of sight, which is causes a decrease of power in the monopole on these scales, $\gtrsim 1\text{Mpc}/h$.

Moreover, we find that, on the scales studied here, both neural networks produce very similar monopoles at all redshifts and for both outflow geometries. Meanwhile, we see that the power suppression in the GM-F monopole is generally smaller than the suppression in the GM monopole. Again, this comes from the different performance among
Figure 7. Redshift space clustering divided the parallel (\(\pi\)) and perpendicular (\(r_\perp\)) directions to the line of sight at \(z = 2.2\) for the Thin Shell outflow geometry. From left to right, in each panel, we display the GM, GM-F, NN:Uniform and NN:Bright algorithms, respectively. For each algorithm, we show the clustering levels \(\xi(r_\perp, \pi) = 10^{-1.0}\) and \(10^{0.0}\) in dashed black lines in their corresponding panels. The solid black lines is the same in every panel and corresponds with the same clustering levels in the case in which there no Ly\(\alpha\) wavelength miss-identification.

Figure 8. Ratio between the monopole in redshift space of the LAEs samples using different Ly\(\alpha\) frequency identification algorithms (thin colored lines) and the one assuming a perfect accuracy in the Ly\(\alpha\) wavelength identification (thick black line). Each column displays a redshift bin (2.2, 3.0 and 5.7 from left to right). Top (bottom) panels panels display the Thin Shell (Galactic Wind) outflow geometry models. The scale in the grey shaded region is logarithmic, while in the white region is linear.

these algorithms. GM-F performs better than GM since it provides a tighter distribution of \(\Delta \lambda\) in the PDF (Fig. 6).

5.2.4 Impact on the quadrupole

The quadrupole (\(\ell = 2\)) is more sensitive to the anisotropic clustering than the monopole. In Fig. 9 we show the quadrupole at different redshifts for both outflow geometries. The solid lines indicate the quadrupole when the peculiar motion of galaxies and the Ly\(\alpha\) wavelength misidentification are implemented (using Eq. 12). In contrast, in order to isolate the contribution of the Ly\(\alpha\) shift, we show with dashed lines the quadrupole using Eq. 12 but assuming that galaxies have no peculiar motion along the line of sight (\(V_{\text{los}} = 0\)).
Additionally, all solid curves converge on large scales, implying that the intrinsic and the suppression evolves with redshift, being lighter at low redshift. In fact, the typical scale at which the intrinsic and the quadrupole amplitude is enhanced, and its sign is flipped, specially at redshift 5.7. In detail, the GM-F algorithm works clearly better than the GM algorithm when the Thin Shell geometry is assumed, while they perform similarly when the Galactic Wind is implemented.

On the other hand, our neural networks work pretty well recovering the intrinsic quadrupole at all redshift and for both outflow geometries. In detail, we find very small differences between NN:Uniform and NN:Bright, as NN:Uniform performs slightly better. In other words, the contribution to the quadrupole given by the Lyα wavelength misidentification becomes negligible when the Lyα wavelength is computed using NN:Uniform and NN:Bright.

6 THE EFFECTS OF THE Lyα WAVELENGTH DETERMINATION IN REALISTIC LINE PROFILES.

In the previous sections we have studied the properties the Lyα line profiles directly predicted by our model. These Lyα line profiles are ideal in terms of i) signal to noise, which is effectively infinite and constant across all the Lyα luminosity range of our models (down to $10^{41.5}$erg s$^{-1}$), and ii) the size of the independent wavelength bins (0.1Å) . However, in observational data sets, reaching these conditions is challenging and/or impossible nowadays. In this section we study how the quality of the Lyα line profiles affects the clustering measurements in Lyα focused spectroscopic galaxy surveys.

---

**Figure 9.** Quadrupole of the LAE samples using Lyα identification algorithms (GM in blue, GM-F in green, NN:Bright in orange and NN:Uniform in yellow), at different redshifts (2.2, 3.0 and 5.7 from left to right) and implementing different outflow geometries (Thin Shell in the top and Galactic Wind in the bottom). The colored solid lines are computed with both, the contribution of the Lyα mis-identification and the peculiar motion of the galaxies. Meanwhile, the colored dashed line only include the shift due to the Lyα line profile. The scale in the grey shaded region is logarithmic, while in the white region is linear.

---

Overall, we find that the uncertainty in the Lyα wavelength determination changes the ratio between the clustering parallel and perpendicular to the line of sight. In this way, the quadrupole amplitude is enhanced at scales $\lesssim 10\,\text{cMpc}/h$. This result is consistent with the Finger-of-God effect as we have already confirmed in Fourier space. The Lyα misidentification can be regarded as an additional random shuffling along the line of sight. This interpretation is further assured by the fact that the dashed lines have negligible quadrupole amplitudes on large scales, $\gtrsim 10\,\text{cMpc}/h$. Additionally, all solid curves converge on large scales, $\gtrsim 10\,\text{cMpc}/h$, suggesting that the quadrupole 2PCF on such scales can be safely used to infer the peculiar velocity contribution in LAE surveys. Finally, The negative amplitude of the quadrupole 2PCF at large scales is qualitatively consistent with the Kaiser effect.

Focusing on the Lyα wavelength misidentification contribution, we find the same trend as the results in Fourier space. The different algorithms produce different quadrupole predictions, reflecting its algorithms’ performance.

On one hand, the quadrupole recovered by the standard algorithms (GM and GM-F) is heavily distorted with respect to the intrinsic one (black). We find that the amplitude of the suppression evolves with redshift, being lighter at low redshift. In fact, the typical scale at which the intrinsic and the observed LAE quadrupole converge is $\sim 5\,\text{Mpc}/h$, $\sim 8\,\text{Mpc}/h$ and $\sim 15\,\text{Mpc}/h$ at redshifts 2.2, 3.0 and 5.7 respectively. Additionally, at small scales, the quadrupole amplitude is enhanced, and its sign is flipped, specially at redshift 5.7. In detail, the GM-F algorithm works clearly better than the GM algorithm when the Thin Shell geometry is assumed, while they perform similarly when the Galactic Wind is implemented.

---

2 This value comes from the bin size used in FLaRED to store the Lyα line profiles.
where $L_{\text{Ly}\alpha}$ is the Ly$\alpha$ luminosity of each galaxy and $L_{\text{Ly}\alpha,F}$ is the Ly$\alpha$ luminosity of the faintest galaxy. Note that our models, by construction, provides good estimates of the LAE luminosity function (Gurung-López et al. 2020). Therefore, the $S/N$ distribution should also mimic observations. Next, to each pixel we add Gaussian noise with an amplitude that corresponds to the $S/N$ of that LAE. In this way, LAEs with lower $S/N$ have a noisier Ly$\alpha$ line profile and vice versa. Moreover, the larger $S/N_F$ is, the better signal to noise ratio has the faintest LAE and the whole LAE population. On the other side, for low $S/N_F$ values, some information from the Ly$\alpha$ line profile (e.g. light blue line) is vanished. In practice, the larger are $W_g$ and $\Delta \lambda_{\text{pix}}$ the more information is destroyed for a fixed value of $S/N_F$. We use four values for $S/N_F$: 6.0, 7.0, 10.0 and 15.0. Also, for the LAE sample studied here, $L_{\text{Ly}\alpha,F} \sim 1.63 \times 10^{42}$ erg s$^{-1}$.

For each combination of $(W_g, \Delta \lambda_{\text{pix}}, S/N_F)$, we produce a catalog of Ly$\alpha$ line profiles. In practice, we could determine directly the Ly$\alpha$ central wavelength from these line profiles, as we did in the ideal case in Sec. 5. However, as the line profiles are progressively downgraded it becomes more difficult to measure properly the FWHM$_{\text{red}}$ (necessary for the GM-F). Also, the global maximum gains more and more discretized as $\Delta \lambda_{\text{pix}}$ increases (needed for both, GM, GM-F). These two problems can be solved by fitting a Gaussian curve to the most prominent peak of the Ly$\alpha$ line profile and then, measure FWHM$_{\text{red}}$ and $\lambda_{\text{Ly}\alpha,\text{Max}}$ from the Gaussian. Another motivation behind this proceeding is that it can also be applied to observational data. Note that, we only use the Gaussian fitting to compute FWHM$_{\text{red}}$ and $\lambda_{\text{Ly}\alpha,\text{Max}}$. In this way, the neural networks use directly the modeled line profiles and not the Gaussian resulting from the fitting.

### 6.2 Ly$\alpha$ wavelength displacement

In this section we compare how the performance of the different algorithms to determine the Ly$\alpha$ wavelength from a Ly$\alpha$ line profile vary with the quality of the spectrum. In Figs. 11, 12, 13 and 14 we show the distributions of the displacement between the assigned Ly$\alpha$ wavelength and the true one for the GM, GM-F, NN:Uniform and NN:Bright algorithms respectively. Overall, for all our four algorithms decreasing the quality of the Ly$\alpha$ line profiles, i.e., decreasing $S/N_F$ and increasing $\Delta \lambda_{\text{pix}}$ and $W_g$, causes a larger uncertainty in the identification of the Ly$\alpha$ wavelength.

Focusing on the GM (Fig. 11), we find that this algorithms is insensitive to lowering the quality of the Ly$\alpha$ line profile. In fact, the mean of the distribution is always entered around 1.5$\AA$. This is due to the fact that Ly$\alpha$ photos are red-shifted as they escape through the galaxy outflows. Moreover, the width of the $\Delta \lambda$ remains also constant ($\sim 1$ $\AA$) until $\Delta \lambda_{\text{pix}}$ = 0.5$\AA$ and $W_g$ = 1.0$\AA$. From that point, it steadily grows. Additionally, the shape of the distribution changes with the quality of the Ly$\alpha$ line profile. On one hand, low values of $\Delta \lambda_{\text{pix}}$ and $W_g$ the distribution is skewed, exhibiting a tail towards large $\Delta \lambda$ values. On the other hand, as $\Delta \lambda_{\text{pix}}$ and $W_g$ increase, the distribution becomes more symmetric.

Next, we find that the GM-F (Fig. 12) algorithms performance is heavily affected by the quality of the Ly$\alpha$ line profile. In the first place, when $\Delta \lambda_{\text{pix}}$ and $W_g$ are low, the $\Delta \lambda$ distribution is centered around 0. However, as the FWHM
of the line profile is increased, the $\Delta \lambda$ distribution moves progressively towards negative $\Delta \lambda$ values (down to -1Å). This is a consequence of the way in which GM-F derived the Lyα wavelength. GM-F corrects the displacement in the Lyα frequency found in GM by using a relation between the wavelength shift and the width of the red peak of the line. However, as $W_g$ increases, the width of the red peak also increases, resulting in Lyα wavelength that are over corrected by the shift due to the RT. In the second place, the spread of the $\Delta \lambda$ distribution increases as the quality of the Lyα line profiles becomes lower. In fact, we find that the standard deviation of the distribution reaches 2Å when $\Delta \lambda_{pix} = 1.0$Å. $W_g = 2.0$Å and $S/N_F = 6.0$.

Moreover, among the two neural network algorithms studied in this work, NN-Bright (Fig.14) is the one which is the most affected by the decrease of quality in the line profile. In general the shape of the $\Delta \lambda$ distribution is composed by a prominent peak located at $\Delta \lambda = 0$ and extended wings bluewards and redwards of Lyα. Increasing $\Delta \lambda_{pix}$ and $W_g$ causes that the wings become more elongated, and hence the accuracy of the algorithm is reduced. Meanwhile, for a set of fixed $\Delta \lambda_{pix}$ and $W_g$, decreasing $S/N_F$ lowers the peak contribution while the wings remain constant, which increases significantly the width of the distributions. The large dependence on $S/N_F$ comes from the fact that NN-Bright uses as a training set the 10% brightest LAEs. As a consequence, this algorithm is trained to reproduce lines with a much higher signal to noise ratio than average in the LAE population. This has little effect when the faintest LAE in the sample has $S/N_F = 15$, since all the galaxy population is going to have a very good S/N and quality. In detail, when $S/N_F = 15$, the S/N of the brightest LAEs is ~ 1000, but the difference in quality between the faint and the bright ends are very small, as the noise for the faintest galaxy is already very tiny. In other words, the quality of the Lyα line profile used for the training set is very similar to the one of the whole LAE sample, even though they exhibit a quite different S/N. However, when $S/N_F$ decreases, the differences in quality in the Lyα line profile become larger, as the faintest LAEs become more and more noisier, while the quality of bright LAEs remains almost unchanged.

Finally, we find that NN-Uniform (Fig.13) is the algorithm with the best performance in most of the range of the $\{\Delta \lambda_{pix}, W_g, S/N_F\}$ volume studied here. Increasing $\Delta \lambda_{pix}$ and $W_g$ and decreasing $S/N_F = 15$ reduce the performance of NN-Uniform. However, it is remarkable how little the spread of the $\Delta \lambda$ distribution is increased through the varied $\Delta \lambda_{pix}$ and $W_g$ range when $S/N_F = 15$ is kept fixed. In fact $\sigma(\Delta \lambda)$ varies from 0.2Å in the best case to only 0.35Å in the worst scenario. This highlights that this level of pixelization and line diluting (due to a FWHM ≠ 0), the Lyα line profiles still contain the necessary information to identify the true Lyα wavelength. However, when $S/N_F$ is reduced, the noise level increases and destroys part of this information. In particular, the higher $\Delta \lambda_{pix}$ and $W_g$ values, the more likely is to lose this information due to the noise.
Figure 12. Same as Fig.11 but using the GM-F algorithm.

Figure 13. Same as Fig.11 but using the NN:Uniform algorithm.
6.3 Monopole artifacts

In general, the misidentification of the Lyα wavelength translates into an incorrect redshift determination and, hence, into a shift in the position of the LAE along the line of sight. This has a direct impact in the measured clustering of LAEs on small scales, as we showed in §5.2. Also, we have just shown in §6.2 how the quality of a given set of observed Lyα line profiles is mirrored into the identification of the Lyα wavelength. In fact, the lower the quality, the more spread the Δλ distribution becomes. Here we study how the quality of the Lyα line profiles imprints the clustering on small scales, focusing on the 2PCF. In the following, we will show the results only for two extreme cases, i.e., the worst (GM) and the best (NN:Uniform) algorithms.

In Fig. 15 we show the monopole 2PCF for all the different combinations of {Δλpix, Wg, S/NP} including the contribution of peculiar velocities and Lyα misidentification (through Eq. 12) given by the NN:Uniform algorithm. Overall, we find a clustering suppression on small scales (ΔPpix, colored lines). At larger scales, the observed LAE clustering converges to the clustering that would be observed if the Lyα frequency was known at infinite precision (ΔPpix, dashed black line). For the GM algorithm we find that there is a strong clustering suppression across the quality range studied here (\(\geq 20\%\) at \(s \sim 1\) cMpc \(h^{-1}\)). Additionally, the suppression depends slightly on the quality of the Lyα line profiles. In fact, the suppression remain quite constant through all the \(\{\Delta \lambda_{\text{pix}}, W_g, S/N_P\}\) range, expect at \(\Delta \lambda_{\text{pix}} \approx 1\)Å and \(W_g \approx 2\)Å, where the lower \(S/N_P\), the stronger the suppression. In fact, we also show the monopole 2PCF for ideal line profiles (solid grey curve) to illustrate how little the clustering recovered by the GM algorithms is perturbed by the line quality. This shows that the Gaussian fitting that we are applying after downgrading the quality to recover the global maximum of the lines profiles works well.

For the NN:Uniform (Fig. 16), the convergence scale depends strongly on the quality of the observed Lyα line profiles. In general, the lower the quality, (i.e., the greater \(\Delta \lambda_{\text{pix}}\) and \(W_g\), and the lower \(S/N_P\)) the larger is the clustering suppression and hence the larger is the convergence scale. It is remarkable how well the NN:Uniform algorithm performs when the signal to noise ratio of the Lyα spectrum is good. In fact, the monopole is affected only on scales lower than 2 cMpc/\(h\) when \(S/N_P = 15\). This highlights that, although diluted, the Lyα line profiles still contain the information about the Lyα wavelength. However, the addition of noise easily destroys progressively this information, causing a greater suppression. In general, the higher \(\{\Delta \lambda_{\text{pix}}, W_g\}\), the most sensitive to noise becomes the clustering of LAEs.

6.4 Quadrupole artifacts

In Figs. 17 and 18 we display the quadrupole for all the multiple quality configuration for the GM and NN:Uniform algorithms, respectively. At the same time we show the samples including only the Lyα misidentification shift along the line of sight (dashed lines) and the samples including also the shift due to peculiar galaxy velocities (solid lines). In general, we find similar trends to the monopole. In particular, the lower the quality of a given set of Lyα line profiles, the larger is the clustering suppression along the line of sight.
GM

| $\Delta \lambda_{\text{pix}}$ | $\xi_{\text{LAE}}/\xi_{\text{R}}$ |
|-----------------------------|----------------------------------|
| 0.25 Å                      | Logarithmic                      |
| 0.5 Å                       | Linear                           |
| 1.0 Å                       |                                 |

$W_g = 0.5$ Å, $W_g = 1.0$ Å, $W_g = 2.0$ Å

Figure 15. Ratio between the observed monopole of LAE samples with ($\xi_{\text{LAE}}$) and without ($\xi_{\text{R}}$) the mis-identification of the Ly$\alpha$ frequency using the GM algorithm. The columns show $\Delta \lambda_{\text{pix}} = 0.25$ Å, 0.5 Å and 1.0 Å from left to right. Meanwhile, the rows show $W_g = 0.5$ Å, 1.0 Å and 2.0 Å from top to bottom. Additionally, in each panel, $S/N_F = 6.0$ is displayed in purple, 7.0 in blue, 10.0 in orange and 15.0 in red. The solid grey line is the monopole computed when the ideal profiles are used (same as Fig. 8). The dashed black line signalizes unity.

(18) S. Gurung-López. et al.

7 DISCUSSION

7.1 Relation between Ly$\alpha$ stack line profile and galaxy properties

As studied in , our model reproduces the trends observed between $\Delta \lambda$ and different galaxy properties. Observational studies have split the Ly$\alpha$ stacked line profile as function of several galaxy properties. For example, Guaita et al. (2017) studied the stacked spectrum of LAEs with spectroscopic observations from VIMOS ULTRA-DEEP SURVEY (VUDS, Le Fèvre et al. 2015; Tasca et al. 2017). Their sample consisted of 76 galaxies between redshift $z = 2$ and $z = 4$, exhibiting both, Ly$\alpha$ and CIII] (1908 Å) as emission lines. The galaxy systemic redshifts were determined by the observed wavelength of the CIII] lines. Then, authors split their galaxy sample by different observed properties. They found that $\Delta \lambda$ anti-correlated with the Ly$\alpha$ rest frame equivalent width and stellar mass, while it correlated with the galaxy overdensity. These trends are in agreements with our model predictions. Moreover, Guaita et al. (2017) also found that the width of the Ly$\alpha$ stacked line profile anti-correlated with the Ly$\alpha$ rest frame equivalent width and stellar mass, as our model predicts too. Moreover, Muzahid et al. (2019) studied the relation between $\Delta \lambda$ and the SFR in 96 LAE at $z \sim 3$ with spectroscopic observations with MUSE (Bacon et al. 2010). Muzahid et al. (2019), found the an anti-correlation between $\Delta \lambda$ and the Ly$\alpha$ EW, as well as, a correlation between $\Delta \lambda$ and the SFR and Ly$\alpha$ luminosity.
Figure 16. Same as Fig. 15 but implementing the NN:Uniform algorithm.

Figure 17. Quadrupole of LAE samples with (solid) and without (dashed) the peculiar motion of galaxies. The plot structure is the same as Fig. 15. Here the black dashed line indicates zero.
Figure 18. Same as Fig.17 but using the NN:Uniform as Lyα frequency identifier.

Figure 19. Left: Monopole convergence scale defined such as $\xi_0/|\xi_0\cdot R = 0.98$ as a function of the standard deviation of the $\Delta \lambda$ distribution for all our quality configurations and Lyα identification algorithms. The GM are displayed as diamonds, GM-F as squares, NN:Uniform as triangles and NN:Bright as hexagons. Then the S/N of the sample is colored coded. Purple means S/N = 6.0, blue 7.0, orange 10.0 and red 15.0. Additionally we show the linear best fit to these data points in black solid line. The black dashed line indicates the direct conversion from $\Delta \lambda$ to distance shift due to Lyα miss-identification. Right: Same as left, but for the quadrupole. The convergence distance for the quadrupole is defined as $\xi_2/|\xi_2\cdot R = 0.90$. 
7.2 Clustering convergence scale

In Fig. 19 (left) we show the scale where $\xi_0$ and $\xi_{0R}$ converge as a function of the standard deviation of the PDF of $\Delta \ell$ ($\sigma(\Delta \ell)$) for all the algorithms and Ly$\alpha$ line profile quality configuration explored in this work. Here we define the monopole convergence scale as the scale at which $\xi_0/\xi_{0R} = 0.98$. Additionally, we show the best fitting linear relation between this convergence scale and $\sigma(\Delta \ell)$ with a slope of $4.56\text{Mpc}/(h\text{km})$ and null origin. We find that our samples follow quite well this linear relation. In detail, the samples with good S/N tend to cluster at lower $\sigma(\Delta \ell)$ values and vice versa. As a reference, we also show the direct conversion from $\sigma(\Delta \ell)$ to distance (dashed black line), computed using Eq.13 and assuming $\Delta_{\text{LOS}} = 0$ and $V_{\text{LOS}} = 0$. The monopole convergence distance (as defined here) has a slope a factor of $\sim 2$ larger than the conversion from $\sigma(\Delta \ell)$ to distance. Fig. 19 illustrates the strong parallelism between the behaviours of the PDF of $\Delta \ell$ and the observed clustering on small scales (Byrohl et al. 2019).

Then, in the right panel of Fig.19 we show the relation between the convergence scale for the quadrupole (defined as $r(\xi_2/\xi_{2R}) = 0.9$). We find the same trends than in the monopole case. Basically, the lower is the accuracy identifying the Ly$\alpha$ frequency, the larger is the suppression on the observed LAE quadrupole. Additionally, the convergence scale for a fix set $\{\Delta_{\text{pix}}, W_\ell/S/NF\}$ is larger in the quadrupole than in the monopole, even though, the definition of convergence scale is more relaxed in the quadrupole. In fact, following the same procedure as in the monopole we fitted a degree polynomial with null origin ordinate to our samples. We find an slope of $19.2\text{cMpc}/h\text{km}$, which is a factor $\sim 10$ larger than the direct conversion between $\sigma(\Delta \ell)$ and distance.

Overall, the performance in determining the Ly$\alpha$ wavelength is fundamentally limited by the spectral quality. A close comparison between the different models implemented in this work shows that the methodology with the lowest monopole suppression on small scales, in general, is NN:Uniform. Then it is followed by NN:Bright for high signal to noise ratios. However, NN:Bright is the methodology the most affected by the reduction on S/NF. In fact, for low S/NF NN:Bright gives the worst results (also depending on the $\Delta_{\text{pix}}$, $W_\ell$ values). Then, in terms of general performance GM-F works better than GM. However, for very low quality Ly$\alpha$ line profiles ($\Delta_{\text{pix}} = 1.0\text{A}$, $W_\ell = 2.0\text{A}$, S/NF = 6.0) the monopole suppression for the LAE samples identified using GM is lowest among its counterparts using different algorithms. This change of trend shows that eventually, for very low quality spectrum, finding the global maximum and setting it as the Ly$\alpha$ wavelength is the most robust proceeding.

7.3 Implications for HETDEX

The Hobby-Eberly Telescope Dark Energy Experiment (Hill et al. 2008; Adams et al. 2011, HETDEX) is a spectroscopic survey chasing LAEs between redshift ~ 1.9 and ~ 3.5. In principle, HETDEX rely only in the measured Ly$\alpha$ line profile to assign a redshift to an LAE. Therefore, the LAE clustering measured by HETDEX would be sensitive to the clustering distortions studied in this work.

The typical spectral resolution and pixel size of HETDEX observations are, respectively, 5\AA{} and 2\AA{} in the observed frame. As the spectral quality is fixed at the observed frame, the spectral quality in the LAE’s rest frame depends on the LAE redshift. Considering HETDEX’s redshift range, the spectral resolution varies between ~ 1.7A and ~ 1.1A and in the LAE’s rest frame. Meanwhile, the pixel size ranges from ~ 0.66A to ~ 0.44A rest frame. For these values of rest frame spectral resolution and pixel size we find that the NN:Uniform algorithm exhibits a better performance than the GM and GM-F algorithms. Thus, in principle, HETDEX would benefit from following the machine learning approach presented in this work.

However, it is challenging to compute the precise impact in the clustering in HETDEX. Since HETDEX LAEs populate smoothly over a given redshift window, the spectral quality is slightly different for every LAE and evolves with redshift. We plan to study this in a follow up paper in which, we will implement LAEs in a simulation lightcone. Meanwhile, here we just give a brief calculation of how much the recovered clustering can improve in HETDEX by using our methodology.

If we consider that HETDEX observations will exhibit an average $W_\ell = 1.5\text{A}$ and $\Delta_{\text{pix}} = 0.5\text{A}$ (rest frame) and S/NF = 6.0. Then, the $\sigma(\Delta \ell)$ values 3 for the GM, GM-F, NN:Uniform and NN:Bright algorithms are 1.39\AA{}, 1.02\AA{}, 0.93\AA{} and 1.43\AA{}. For this particular configuration the NN:Bright algorithm is outperformed by the other algorithms. This is caused by the fact that, NN:Bright is only trained with the brightest LAEs. Also, GM-F performs better than GM, as in general. Meanwhile, the NN:Uniform is algorithm with the highest accuracy, exhibiting a 10% better performance than GM-F. Following our results in the previous subsection, these $\sigma(\Delta \ell)$ values translate into a convergence scale $r(\xi_0/\xi_{0R}) = 0.98$ for the monopole of 6.33cMpc/$h$, 4.65cMpc/$h$, 4.24cMpc/$h$ and 6.52cMpc/$h$ for the GM, GM-F, NN:Uniform and NN:Bright algorithms respectively. Meanwhile, the convergence scale of the quadrupole $r(\xi_2/\xi_{2R}) = 0.98$ are 26.7cMpc/$h$, 19.6cMpc/$h$, 17.9cMpc/$h$ and 27.5cMpc/$h$.

8 CONCLUSIONS

In this work we have addressed the clustering distortions in LAE samples due to the misidentification of the Ly$\alpha$ wavelength in the Ly$\alpha$ line profile and how to mitigate them using neural networks. With this goal, we have analyzed the Ly$\alpha$ line profiles from our previous LAE theoretical model (Gurung-López et al. 2020), that includes the Ly$\alpha$ RT in the ISM and in the IGM. Our LAE model reproduces by construction the observed LAE luminosity function and a bunch of their observables, such us the Ly$\alpha$ escape fraction, the metallicity distribution and clustering amplitude of LAEs (Gurung-López et al. 2019a).

After analysing the stacked Ly$\alpha$ line profiles of our LAE model we find that:

- The stacked Ly$\alpha$ line profile is affected by the IGM. In particular, as the IGM Ly$\alpha$ optical depth increases with

---

3 These values are calculated taking the mean of $\sigma(\Delta \ell)$ for $W_\ell = 1.0\text{A}$ and $W_\ell = 2.0\text{A}$ rest frame.
redshift, the higher the redshift, the more the stacked Lyα line profile is modified by the IGM. At low redshift ($z = 2.2$ and $z = 3.0$), we find that the stacked Lyα line profile changes slightly and only bluewards the Lyα wavelength. Meanwhile, at redshift 5.7, the radiative transfer in the IGM modifies the stacked Lyα line profile up to 2A redder than the Lyα wavelength.

- The stacked Lyα line profile depends on both, galaxy and IGM properties. On one hand, our model predicts that, for the Thin Shell and the Galactic Wind outflow geometries, the Lyα stacked line profile is centered at large wavelength for higher values of SFR and $L_{\text{Ly}\alpha}$. Additionally, LAEs with higher values of Lyα equivalent width exhibit a bluer Lyα stacked line profile. Meanwhile, in the Galactic Wind we find only a small dependence on the stellar mass, while in the Thin Shell, the lower $M_*$, the more redshifted is the peak of the stacked Lyα line profile. Over all, these trends are in good agreement with observational works (Guaita et al. 2017; Muzahid et al. 2019). On the other hand, the stacked Lyα line profiles are more redshifted in high density environments and low IGM large scale line of sight velocity, its gradient and density gradient.

Then, we have introduced a novel approach to measure the systemic redshift of LAEs from their Lyα line using neural networks. In this frame, given a survey that only observed the Lyα line, a fraction of the LAE population could be re-observed to find other features to determine their systemic redshifts. Then, this sub-population would be used to train a neural network that predicts the systemic redshift. In particular, we have explored two different ways of building the training set: i) the re-observed sources are chosen uniformly across their properties (NN:Uniform), and ii) only the brightest LAEs are re-observed (NN:Bright). In order to assess the performance of these methodologies we compare them with others found in the literature. In particular, we use i) GM, that uses the global maximum of the Lyα line profile to assign a systemic redshift (Verhamme et al. 2018; Byrohl et al. 2019; Muzahid et al. 2019) and ii) GM-F, suggested by Verhamme et al. (2018), that uses the width of the Lyα line profile for the redshift due to the Lyα RT in the ISM.

First, we focus on the Lyα line profiles produced by our model, which are ideal in terms of signal to noise and pixelization. We find that the NN:Uniform and NN:Bright algorithms performs better than GM and GM-F. In fact the distributions of the displacement of Lyα (∆α) if the broadest for the GM, followed by the GM-F and NN:Bright, while for the NN:Uniform, it is the thinnest. Then we study how each of these methods impacts the observed clustering of LAEs with ideal line profiles. We find that:

- In general, the power spectrum exhibits a damping at small scales that disappears at large enough distances, as found in Byrohl et al. (2019). This damping is directly linked to the performance recovering the systemic redshift of the LAEs. In fact, the clustering of samples using GM and GM-F exhibits a decrease of 80% of power at $k_{\parallel} = 1.0$. In contrast, the samples using NN:Uniform and NN:Bright exhibit a much shallower damping. Typically, samples using NN:Bright have a damping of 10% or less at $k_{\parallel} = 1.0$. Meanwhile, the samples using NN:Uniform are mostly unaffected at $k_{\parallel} = 1.0$, as the power spectrum damping is of the order of the 1%.

- The monopole also exhibits a damping at small scales parallel to the one observed in the power spectrum. In particular, the power suppression in the monopole is of the order of 1% at 1cMpc/h. Meanwhile, for GM and GM-F the monopole damping can go up to the 60% at 1cMpc/h. Also, for the GM and GM-F, the suppression extends up to 10cMpc/h, where, in general, the intrinsic and the observed monopole converge.

- The quadrupole is also sensitive to the systemic redshift determination. In fact, we find that, the lower the accuracy recovering the Lyα wavelength, the more power exhibits the quadrupole between 1cMpc/h and 10cMpc/h. The quadrupole of the samples using NN:Bright and, specially, NN:Uniform are mostly identical to the quadrupole of the underlying LAE population.

Next, we explore the benefits of using NN:Uniform and NN:Bright in comparison with GM and GM-F in realistic line profiles. With this goal, we lower the quality of the Lyα line profile mocking several artifacts in observations. In practice, i) we dilute the line assuming different instrumental FWHM, ii) we reduce the wavelength resolution by pixelizing the line and iii) we include noise in the Lyα line profile. Then, we study the properties of these samples, finding:

- The performance, i.e., the distributions $\Delta$ are tightly connected to the spectral quality. Overall, we find that the spectral quality range covered in this work, the NN:Uniform is the best methodology. Additionally, the NN:Bright is very affected by the noise in the spectrum. This is a result of using only the brightest LAEs as training sample, as this one lacks faint LAEs in which their Lyα line profile is noisy. In this way, NN:Bright is the second best algorithm for high signal to noise lines. Meanwhile, GM-F progressively gets worst as the spectral quality decreases. GM-F is mostly affected by the instrumental FWHM, as this modifies the width of the Lyα line and over-corrects the RT in the ISM. Finally, GM is the algorithm with the lowest performance in most of the spectral quality regime studied here. However, GM is quite insensitive to lowering the spectral quality, specially, increasing the noise. This, in the worst spectral quality considered here, translates into a better accuracy of GM in comparison with NN:Bright and GM-F, while it is similar to NN:Uniform.

- Consequently, the monopole and the quadrupole are affected by the reduction of the spectral quality. We find that NN:Uniform is algorithm that recovers better the clustering of the underlying LAE population. However, as the spectral quality is reduce, the damping of power at small scales increases, as in GM-F and NN:Bright. Meanwhile, the clustering damping at small scales is quite constant for GM through the dynamical range of the spectral quality studied here.

- There is a linear relation between the algorithm performance and the typical scale up to which the clustering power is decreased. This will be useful for the designing of future surveys based on the Lyα line.

Therefore, we conclude that spectroscopic Lyα based surveys such, as HETDEX, might benefit from measuring the systemic redshift of a relative small subsample of LAEs, using other spectral features. And then, using this subsample
to train machine learning algorithms to predict the systemic redshift of the rest of the observed LAE population.

ACKNOWLEDGEMENTS

This work has made used of CEFCAs Scientific High Performance Computing system which has been funded by the Governments of Spain and Aragón through the Fondo de Inversiones de Teruel, and the Spanish Ministry of Economy and Competitiveness (MINECO-FEDER, grant AYA2012-30789). The authors acknowledge the support of the Spanish Ministerio de Economía y Competitividad project No. AYA2015-66211-C2-P-2. We acknowledge also STFC Consolidated Grants ST/L00075X/1 and ST/P000451/1 at Durham University. This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/P002293/1, ST/R002371/1 and ST/S002602/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure. SS was supported in part by the Munich Institute for Astro- and Particle Physics (MIAPP). This work used the DiRAC@Durham facility managed by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). The equipment was funded by BEIS capital funding via STFC capital grants ST/P002293/1, ST/R002371/1 and ST/S002602/1, Durham University and STFC operations grant ST/R000832/1. DiRAC is part of the National e-Infrastructure. SS was supported in part by the Munich Institute for Astro- and Particle Physics (MIAPP) which is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy (EXC-2094-39078311).

REFERENCES

Adams J. J., et al., 2011, ApJS, 192, 5
Ahn S., Lee H.-W., Lee H. M., 2000, Journal of Korean Astronomical Society, 33, 29
Bacon R., et al., 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 8, doi:10.1117/12.856027
Baugh C. M., et al., 2016, MNRS, 462, 3854
Benitez N., et al., 2014, preprint, (arXiv:1403.5237)
Beutler F., et al., 2014, MNRS, 443, 1065
Byrohl C., Saito S., Behrens C., 2019, MNRS, 489, 3472
Caruana J., et al., 2018, MNRS, 473, 30
Cole S., Lacey C. G., Baugh C. M., Frenk C. S., 2000, MNRS, 319, 168
Croom S., Dijkstra M., McCourt M., Oesch P. A., 2015, ApJ, 836, L6
Garrington L., et al., 2017, ApJ, 843, 133
Steidel C. C., Giavalisco M., Pettini M., Dickinson M., Adelberger K. L., 1998, ApJ, 505, L17
Tasca L. A. M., et al., 2017, A&A, 600, A110
Trainor R. F., Steidel C. C., Strom A. L., Rudie G. C., 2015, ApJ, 809, 89
Verhamme A., Schaerer D., 2018, A&A, 616, A60
Verhamme A., Schaerer D., Maselli A., 2006, A&A, 460, 397
Zhang Z., Miralda-Escudé J., 2002, ApJ, 578, 33
Zheng Z., Miralda-Escudé J., 2010, ApJ, 716, 574
Zheng Z., Cen R., Trac H., Miralda-Escudé J., 2011, ApJ, 726, 809
Zheng Z., Cen R., Trac H., Miralda-Escudé J., 2012, ApJ, 768, 52

APPENDIX A: WAVELENGTH IMPORTANCE

In this section we quantify which wavelength of the line profile contribute the most to the determination of the $\lambda_{obs}^\text{Ly}\alpha$ in the neural networks in this work. With this goal in mind, for a given wavelength bin centered at the wavelength $\lambda_{pix}$ we define its importance as

$$I(\lambda_{pix}) = \sigma(\omega(\lambda_{pix})).$$  \hspace{1cm} (A1)

where $\omega(\lambda_{pix})$ is each weight for the pixel centered at $\lambda_{pix}$ and $\sigma(\omega(\lambda_{pix}))$ is the standard deviation of that pixel across the training set.

In Fig. 20 we show the importance as a function of wavelength for the NN:Uniform at different redshifts and for the two outflow geometries. In general, we find that the wavelength range with a significant importance is wider than the
Lya stacked line profiles at \( z = 2.2 \) and 3.0, while at \( z = 5.7 \) both are similar. On one hand, at \( z = 5.7 \) the IGM absorbs mostly all the flux bluewards Lya Zheng et al. (2010); Laursen et al. (2011); Gurung-López et al. (2020). Therefore, \( \lambda_{\text{Lya}} \) is extracted mainly from the information closer than 2Å to \( \lambda_{\text{Lya}} \). Meanwhile, the regions with \( \lambda < \lambda_{\text{Lya}} \) and \( \lambda > \lambda_{\text{Lya}} + 5 \)Å contain little information in comparison. On the other hand, at lower redshifts (\( z = 2.2 \) and 3.0), the IGM absorption is not that strong and the neural networks need the information from a broader wavelength range. In fact, at these redshifts, most of the importance is located between \( \lambda_{\text{Lya}} \) and 5Å redwards. Additionally, the spectral region with a significant importance spawns from \( \sim 1 \)Å bluewards \( \lambda_{\text{Lya}} \) to \( \sim 15 \)Å redwards \( \lambda_{\text{Lya}} \). This broad wavelength range exhibits the large variety of line profiles in our LAE population, as the presence of very broad lines (FWHM\( \sim 10 \)Å) extend the importance range up to \( \sim 15 \)Å.

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