Highlights of Spanish Astrophysics XII, Proceedings of the XVI Scientific Meeting of the Spanish Astronomical Society held on July 15 - 19, 2024, in Granada, Spain. M. Manteiga, F. González Galindo, A. Labiano Ortega, M. Martínez González, N. Rea, M. Romero Gómez, A. Ulla Miguel, G. Yepes, C. Rodríguez López, A. Gómez García and C. Dafonte (eds.), 2025

ForestFlow: cosmological emulation of Lyman-α forest clustering from linear to nonlinear scales

Chaves-Montero, J.[1], **Cabayol-Garcia, L.**[1,2], **Lokken, M.**[1], **Font-Ribera, A.**[1], **and DESI Lyman α working group**

[1] Institut de Física d'Altes Energies (IFAE), The Barcelona Institute of Science and Technology, 08193 Bellaterra (Barcelona), Spain
[2] Port d'Informació Científica, Campus UAB, C. Albareda s/n, 08193 Bellaterra (Barcelona), Spain

Abstract

On large scales, measurements of the Lyman-α forest offer insights into the expansion history of the Universe, while on small scales, these impose strict constraints on the growth history, the nature of dark matter, and the sum of neutrino masses. This work introduces ForestFlow, a cosmological emulator designed to bridge the gap between large- and small-scale Lyman-α forest analyses. Using conditional normalizing flows, ForestFlow emulates the 2 Lyman-α linear biases (b_δ and b_η) and 6 parameters describing small-scale deviations of the 3D flux power spectrum ($P_{\rm 3D}$) from linear theory. These 8 parameters are modeled as a function of cosmology — the small-scale amplitude and slope of the linear power spectrum — and the physics of the intergalactic medium. Thus, in combination with a Boltzmann solver, ForestFlow can predict $P_{\rm 3D}$ on arbitrarily large (linear) scales and the 1D flux power spectrum ($P_{\rm 1D}$) — the primary observable for small-scale analyses — without the need for interpolation or extrapolation. Consequently, ForestFlow enables for the first time multiscale analyses. Trained on a suite of 30 fixed-and-paired cosmological hydrodynamical simulations spanning redshifts from $z = 2$ to 4.5, ForestFlow achieves 3 and 1.5% precision in describing $P_{\rm 3D}$ and $P_{\rm 1D}$ from linear scales to $k = 5\,{\rm Mpc}^{-1}$ and $k_\parallel = 4\,{\rm Mpc}^{-1}$, respectively. Thanks to its parameterization, the precision of the emulator is also similar for both ionization histories and two extensions to the ΛCDM model — massive neutrinos and curvature — not included in the training set. ForestFlow will be crucial for the cosmological analysis of Lyman-α forest measurements from the DESI survey.

1 Introduction

In this work, we design the first approach to provide consistent predictions for Lyman-α forest clustering from linear to nonlinear scales. To do so, we first compute the best-fitting parameters of a physically-motivated Lyman-α clustering model to measurements from a suite of cosmological hydrodynamical simulations. Then, we emulate these parameters as a function of cosmology using FORESTFLOW, a conditional normalizing flow (cNFs) [1]. In particular, we emulate the 2 Lyman-α linear biases (b_δ and b_η), which completely set the large-scale behavior of $P_{\rm 3D}$ together with the linear power spectrum, and 6 parameters modeling small-scale deviations of $P_{\rm 3D}$ from linear theory. Consequently, this strategy has the potential to make precise $P_{\rm 3D}$ predictions from nonlinear to arbitrarily large (linear) scales even when using simulations with moderate sizes as training data. It also enables predicting any Lyman-α statistic derived from $P_{\rm 3D}$ without requiring interpolation or extrapolation. For instance, we can compute $\xi_{\rm 3D}$ by taking the Fourier transform of $P_{\rm 3D}$ or determine $P_{\rm 1D}$ by integrating its perpendicular modes.

2 Emulator

2.1 Parametric model for $P_{\rm 3D}$

The three-dimensional power spectrum of the Lyman-α forest can be decomposed into three terms

$$P_{\rm 3D}(k,\,\mu) = (b_\delta + b_\eta\, f\,\mu^2)^2 D_{\rm NL}(k,\,\mu) P_{\rm lin}(k), \tag{1}$$

where $f = {\rm d}\log G/{\rm d}\log a$ is the logarithmic derivative of the growth factor G, $(b_\delta + b_\eta\, f\,\mu^2)^2$ accounts for linear biasing and large-scale redshift space distortions, $P_{\rm lin}$ is the linear matter power spectrum, and $D_{\rm NL}$ is a physically-motivated parametric correction accounting for the nonlinear growth of the density field, nonlinear peculiar velocities, thermal broadening, and pressure.

The large-scale behavior of $P_{\rm 3D}$ is set by the bias coefficients b_δ and b_η together with the linear power spectrum, and the latter can be computed using a Boltzmann solver. Making predictions for $P_{\rm 3D}$ on small scales is more challenging than on large scales due to the variety of effects affecting this statistic on the nonlinear regime. In this work, we describe small-scale effects using the physically-motivated [2] parameterization

$$D_{\rm NL} = \exp\left\{\left(q_1\Delta^2 + q_2\Delta^4\right)\left[1 - \left(\frac{k}{k_{\rm v}}\right)^{a_{\rm v}}\mu^{b_{\rm v}}\right] - \left(\frac{k}{k_{\rm p}}\right)^2\right\}, \tag{2}$$

where $\Delta^2(k) \equiv (2\pi^2)^{-1}k^3 P_{\rm lin}(k)$ is the dimensionless linear matter power spectrum, μ is the cosine of the angle between the Fourier mode and the line of sight, and the free parameters $k_{\rm v}$ and $k_{\rm p}$ are in $\rm Mpc^{-1}$ units throughout this work. The terms involving $\{q_1,\, q_2\}$, $\{k_{\rm v},\, a_{\rm v},\, b_{\rm v}\}$, and $\{k_{\rm p}\}$ account for nonlinear growth, peculiar velocities and thermal broadening, and gas pressure, respectively.

3 Conditional normalizing flows

Normalizing flows are a class of machine-learning generative models designed to predict complex distributions by applying a sequence of bijective mappings to simple base distributions. A natural extension to this framework is conditional NFs, a type of NFs that condition the mapping between the base and target distributions on a series of input variables. Given an input $\mathbf{x} \in X$ and target $\mathbf{y} \in Y$, cNFs predict the conditional distribution $p_{Y|X}(\mathbf{y}|\mathbf{x})$ by applying a parametric, bijective mapping $f_\phi : Y \times X \to Z$ to a base distribution $p_Z(\mathbf{z})$ as follows

$$p_{Y|X}(\mathbf{y}|\mathbf{x}) = p_Z(f_\phi(\mathbf{y}, \mathbf{x})|\mathbf{x}) \left| \frac{\partial f_\phi(\mathbf{y}, \mathbf{x})}{\partial \mathbf{y}} \right|, \tag{3}$$

where ϕ are the parameters of the mapping, while the last term of the previous equation is the Jacobian determinant of the mapping. In FORESTFLOW, the input is given by the parameters capturing the dependence of the Lyman-α forest on cosmology and IGM physics, $\mathbf{x} = \{\Delta^2_{\mathrm{p}}, n_{\mathrm{p}}, \bar{F}, \sigma_{\mathrm{T}}, \gamma, k_{\mathrm{F}}\}$ [3], the target by the parameters of the P_{3D} model, $\mathbf{y} = \{b_\delta, b_\eta, q_1, q_2, k_{\mathrm{v}}, a_{\mathrm{v}}, b_{\mathrm{v}}, k_{\mathrm{p}}\}$, and the base distribution is an 8-dimensional Normal distribution $N^8(0, 1)$, where the dimension is determined by the number of P_{3D} model parameters.

3.1 Training and testing data

We train FORESTFLOW using measurements from the suite of cosmological hydrodynamical simulations presented in [3], which consists of 30 fixed-and-paired hydrodynamical simulations of 67.5 Mpc on a side. We evaluate different aspects of the emulation strategy using 6 fixed-and-paired simulations with cosmological and astrophysical parameters not considered in the TRAINING simulations: CENTRAL to test of the emulator's performance at the center of the parameter space, SEED to evaluate the impact of cosmic variance in the training set on FORESTFLOW predictions, and GROWTH with a different expansion history, NEUTRINOS with massive neutrinos, and CURVED with curvature to evaluate the performance of the emulation strategy for cosmologies not included in the training set. Finally, we use REIONISATION to tests the emulator's performance for thermal histories not considered in the TRAINING simulations.

To generate the training and testing data for our emulator, we compute the best-fitting parameters of Eqs. 1 and 2 to measurements from the simulations described in §3.1. We fit the model using P_{3D} measurements from $k = 0.09$ to $5\,\mathrm{Mpc}^{-1}$ and P_{1D} measurements from $k_\parallel = 0.09$ to $4\,\mathrm{Mpc}^{-1}$. The size of our simulation boxes determines the largest scales used, while the maximum wavenumbers are set by the smallest scales measured by DESI.

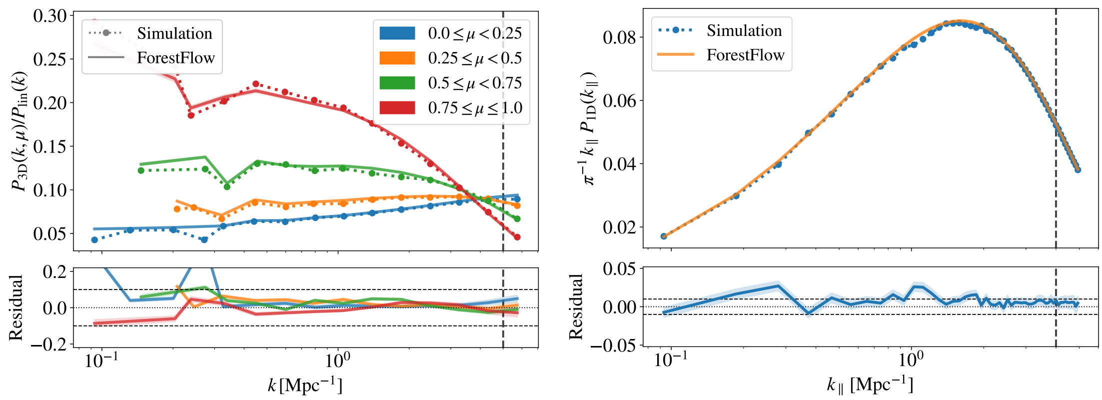

Figure 1: Accuracy of the emulator in recovering P_{3D} and P_{1D} measurements from the CENTRAL simulation at $z = 3$. Dotted lines show measurements from simulations, solid lines and shaded areas display the average and 68% credible interval of FORESTFLOW predictions, respectively, and vertical dashed lines indicate the minimum scales considered for computing the training data of the emulator. The overall performance of the emulator in recovering P_{3D} and P_{1D} is 2.0 and 0.6%.

4 Emulator performance

4.1 Cosmologies in the training set

In Fig. 1, we compare measurements of P_{3D} and P_{1D} from the CENTRAL simulation at $z = 3$ with FORESTFLOW predictions. Dotted lines show simulation measurements, while solid lines and shaded areas display the average and 68% credible interval of FORESTFLOW predictions, respectively. As we can see, the emulator captures the amplitude and scale-dependence of P_{3D} and P_{1D} precisely. To better characterize the emulator's performance, we compute the average accuracy of FORESTFLOW in recovering measurements from CENTRAL across redshift. We find that it is 1.2 and 0.3% for b_δ and b_η, respectively, which translates into 1.1 and 1.2% for perpendicular and parallel P_{3D} modes on linear scales, and 2.6 and 0.8% for P_{3D} and P_{1D}. Note that cosmic variance hinders our ability to test the performance of the model; however, this does not necessarily indicate a decrease in the model's accuracy for P_{3D} on the largest scales sampled by our simulation.

4.2 Cosmologies and IGM histories outside the training set

In Fig. 2, we examine the accuracy of FORESTFLOW reproducing P_{3D} and P_{1D} measurements from simulations not included in the training set. Lines indicate the redshift average of the relative difference between model predictions and simulation measurements. The first two rows show the results for the CENTRAL and SEED simulations, whose only difference is their initial distribution of phases. Consequently, the predictions of FORESTFLOW are the same for both. As we can see, the performance of FORESTFLOW is practically the same for both

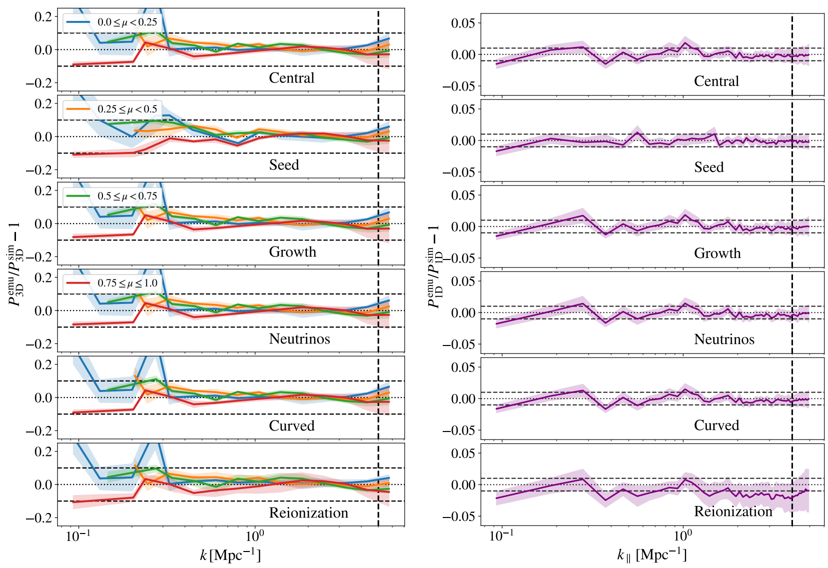

Figure 2: Performance of the emulator for simulations with cosmologies not included in the training set. Lines and shaded areas display the average and standard deviation of the results for 11 snapshots between $z = 2$ and 4.5, respectively. From top to bottom, the rows show the results for the CENTRAL, SEED, GROWTH, NEUTRINOS, CURVED, and REIONISATION simulations.

simulations.

In the third, fourth, and fifth rows of Fig. 2, we use the GROWTH, NEUTRINOS, and CURVED simulations to evaluate the accuracy of FORESTFLOW for three different scenarios not contemplated in the training set: different growth history, massive neutrinos, and curvature. As we can see, the performance of FORESTFLOW for all these simulations is approximately the same as for the CENTRAL simulation. These results support that using the small-scale amplitude and slope of the linear power spectrum to capture cosmological information enables setting precise constraints on growth histories and ΛCDM extensions not included in the training set [3].

In the last row of Fig. 2, we examine the accuracy of FORESTFLOW for the REIONISATION simulation, which employs a HeII reionization history significantly different from those used by the TRAINING simulations. The performance of the emulator for this and the CENTRAL simulation is similar, which is noteworthy given that the performance of P_{1D} emulators for the REIONISATION is significantly worse than for the CENTRAL simulation [5]. The outstanding performance of FORESTFLOW is likely because the relationship between IGM physics and the parameters of the P_{3D} model is more straightforward than with P_{1D} variations.

5 Conclusions

We present FORESTFLOW, a cosmological emulator that predicts Lyman-α clustering from linear to nonlinear scales. Using an architecture based on conditional normalizing flows, FORESTFLOW emulates the 2 linear Lyman-α biases (b_δ and b_η) and 6 physically-motivated parameters capturing small-scale deviations of the three-dimensional flux power spectrum ($P_{\rm 3D}$) from linear theory. We summarize the main results of this work below:

- FORESTFLOW predicts Lyman-α clustering on arbitrarily large (linear) scales when combined with a Boltzmann solver, and makes predictions for any statistics derived from $P_{\rm 3D}$ without interpolation or extrapolation.
- To train the emulator, we use the best-fitting value of the 8 model parameters to $P_{\rm 3D}$ and $P_{\rm 1D}$ measurements from a suite of 30 fixed-and-paired cosmological hydrodynamical simulations spanning 11 equally-spaced redshifts between $z = 2$ and 4.5. We emulate these parameters as a function of cosmology and IGM physics.
- The accuracy of FORESTFLOW in predicting $P_{\rm 3D}$ from linear scales to $k = 5\,{\rm Mpc}^{-1}$ is 3% and 1.5% for $P_{\rm 1D}$ down to $k_\parallel = 4\,{\rm Mpc}^{-1}$. It also displays similar performance as before for two extensions to the ΛCDM model — massive neutrinos and curvature — and ionization histories not included in the training set. We find that the size and number of training simulations have a similar impact on the emulator's performance as uncertainties arising from the limited flexibility of the 8-parameter model.

The release of FORESTFLOW is timely for Lyman-α forest analyses with the ongoing Dark Energy Spectroscopic Instrument (DESI) survey. It enables a series of novel multiscale studies with DESI data, including connecting large- and small-scale analyses as well as extending three-dimensional analyses towards smaller scales.

References

[1] Winkler, C., Worrall, D., Hoogeboom, E., and Welling, M. (2019). Learning likelihoods with conditional normalizing flows. *arXiv e-prints*, 1912.00042.

[2] Arinyo-i-Prats, A., Miralda-Escudé, J., Viel, M., and Cen, R. (2015). The non-linear power spectrum of the Lyman alpha forest. *JCAP*, *2015*, 017–017.

[3] Pedersen, C., Font-Ribera, A., Rogers, K. K., McDonald, P., Peiris, H. V., Pontzen, A., and Slosar, A. (2021). An emulator for the Lyman-α forest in beyond-ΛCDM cosmologies. *JCAP*, *2021*(5), 033.

[4] Bird, S., Rogers, K. K., Peiris, H. V., Verde, L., Font-Ribera, A., and Pontzen, A. (2019). An emulator for the Lyman-α forest. *JCAP*, *2019*(2), 050.

[5] Cabayol-Garcia, L., Chaves-Montero, J., Font-Ribera, A., and Pedersen, C. (2023). A neural network emulator for the Lyman-α forest 1D flux power spectrum. *MNRAS*, *525*(3), 3499–3515.