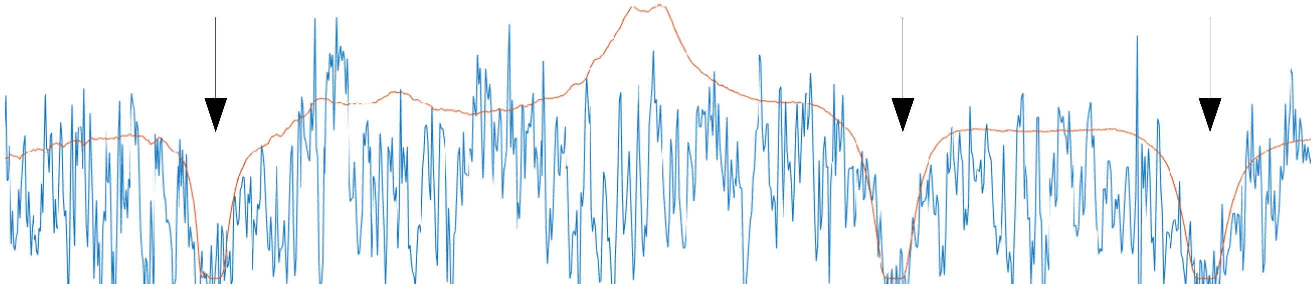
Detecting Multiple DLAs with Bayesian machine learning

Ming-Feng Ho (Me) (UCR), Simeon Bird (UCR), Roman Garnett (WUSTL)

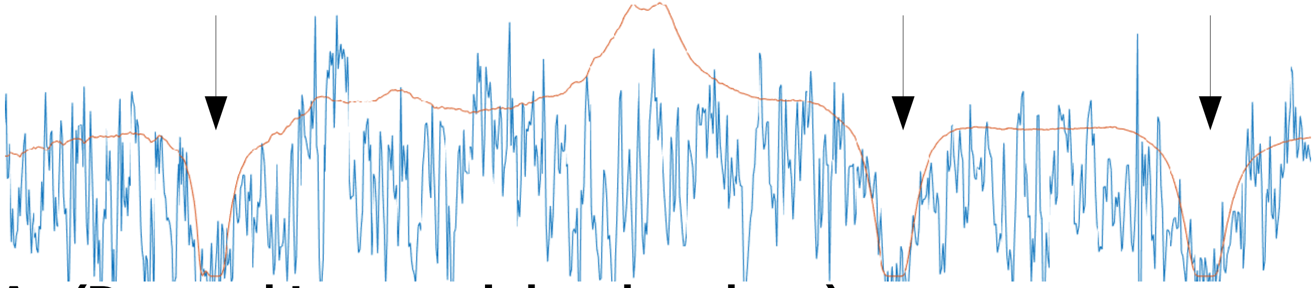


amazon web services

UCR High Performance Computing Center

Detecting Multiple DLAs with Bayesian machine learning

Ming-Feng Ho (Me) (UCR), Simeon Bird (UCR), Roman Garnett (WUSTL)



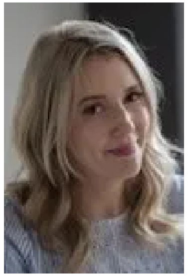
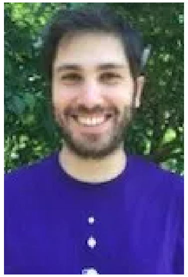
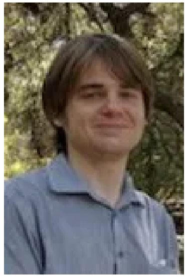
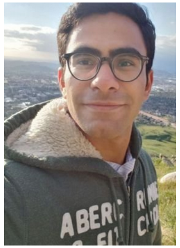
DLAs (Damped Lyman alpha absorbers):
Strong neutral hydrogen absorbers (usually 2<z<5).
Dominate neutral hydrogen budget after reionisation.

High Performance Computing Center

Our group

- PI: Simeon Bird

- Phoebe Upton Sanderbeck (HeII reionisation)
- Martin Fernandez (primordial black holes)
- Bryan Scott (mock catalogue)
- Mahdi Qezlou (DLA metallicity in IllustrisTNG)
- Me (machine learning in Lyman alpha forest)


- **Me:** Hey, I like your ML paper. Are you still working on that project with **Roman**?
- **Simeon:** Glad you enjoyed the paper! I am currently looking for a student in this area. But I actually never met Roman before.
- **Me:** Wait. What did you mean you've never met him before? How did this collaboration work if you did not know him?
- **Simeon:** He is a very secret person who worked for the US government. The only thing I know is his little picture on GitHub.

- **Me:** Huh, that's so cool. So you two just collaborate through GitHub?
- **Simeon:** Yeah.
- **Me:** Does that mean I can also work with you through GitHub without showing up in the group meeting?
- **Simeon:** No, you have to come to the group meetings.

Lyman alpha absorbers

- background: quasars
- Hydrogen absobers

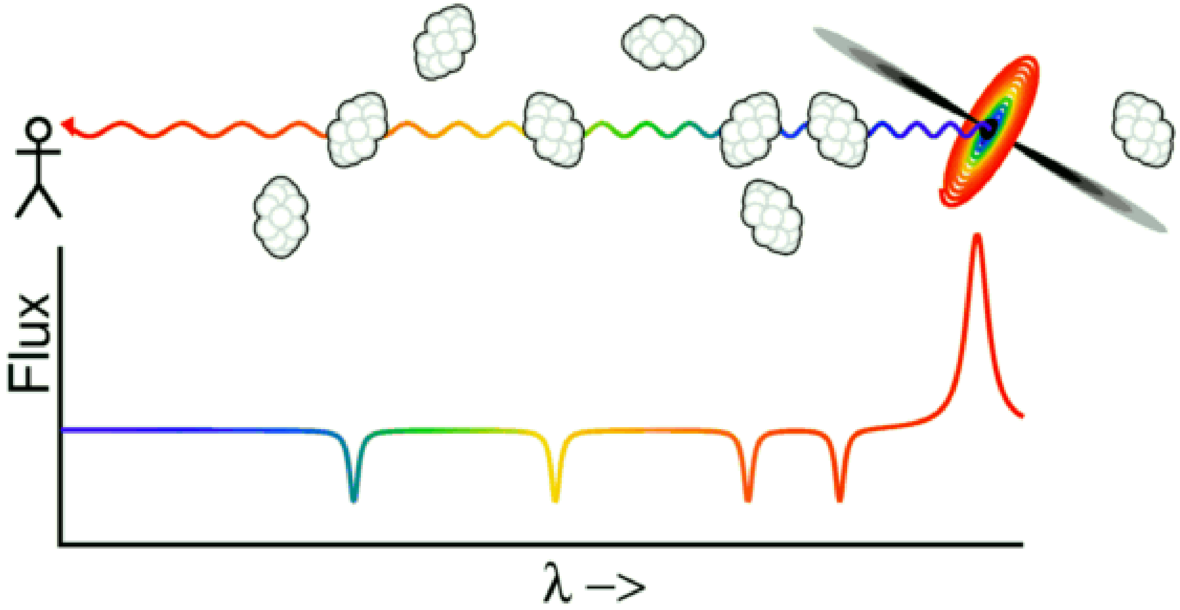
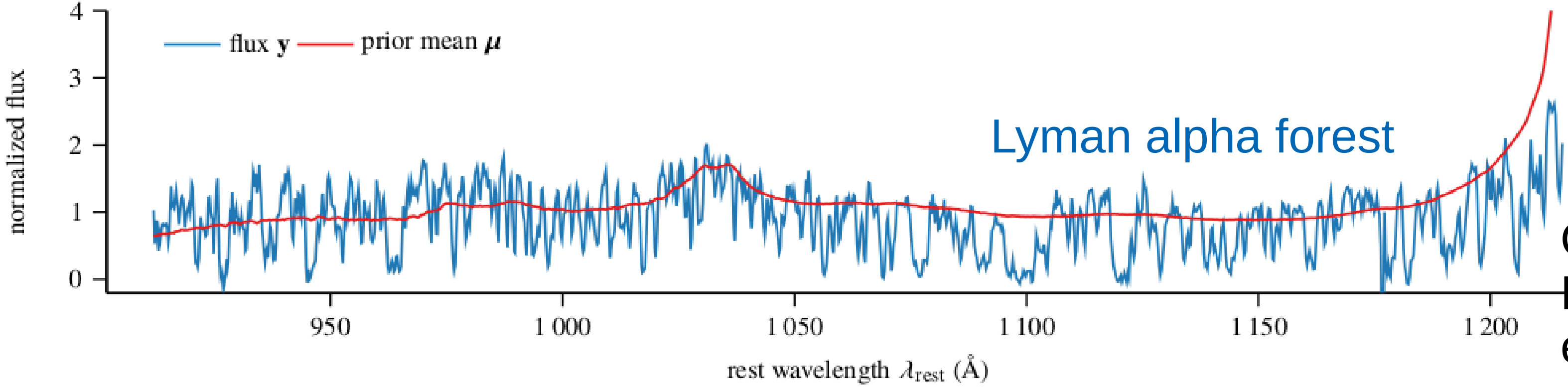


Image: Edward L. Wright

Finding DLAs in Spectra

Currently done by **visual inspection** of spectra

Look for wide dips in the spectrum below (through GSM):



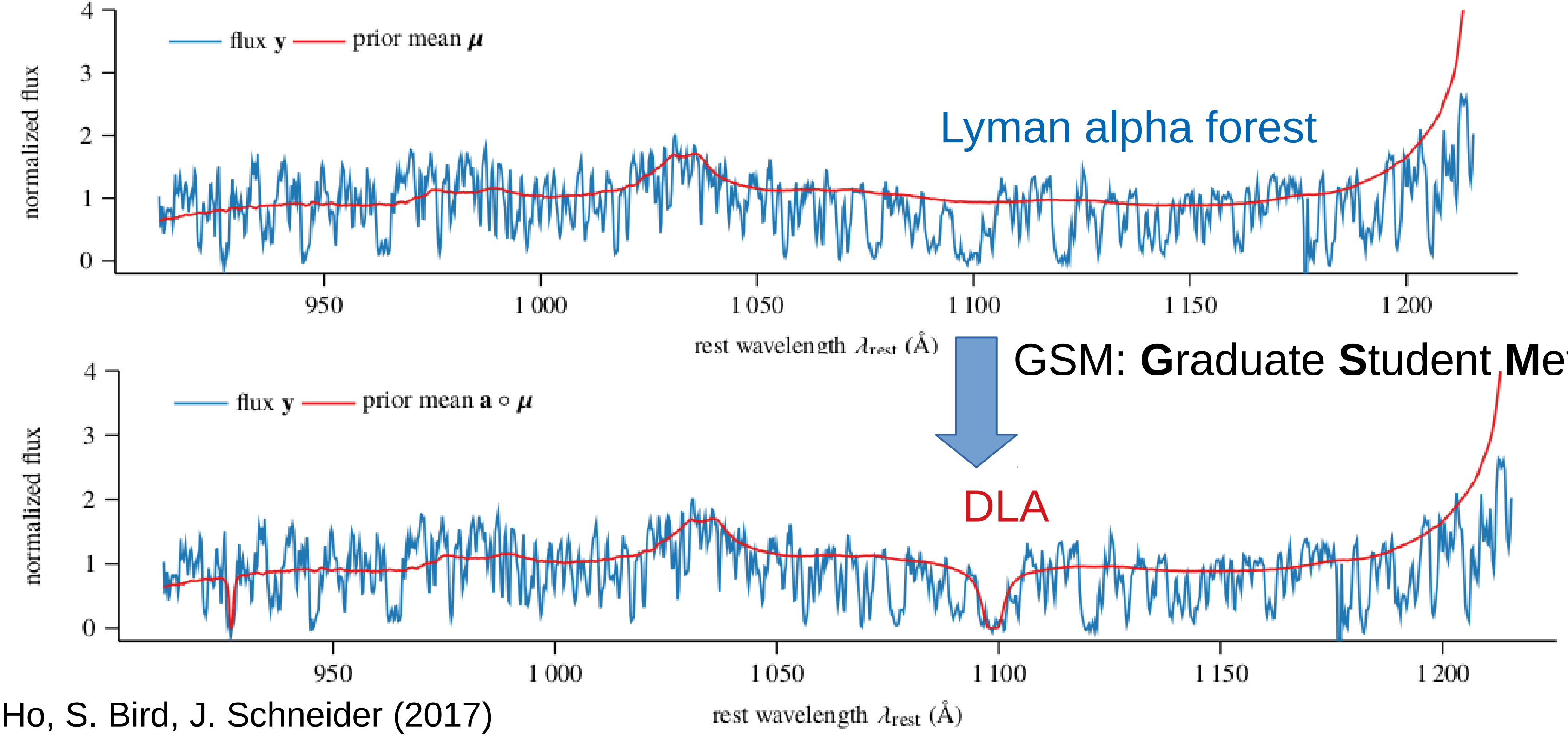
QSO
Lya
emission

R. Garnett, S. Ho, S. Bird, J. Schneider (2017)

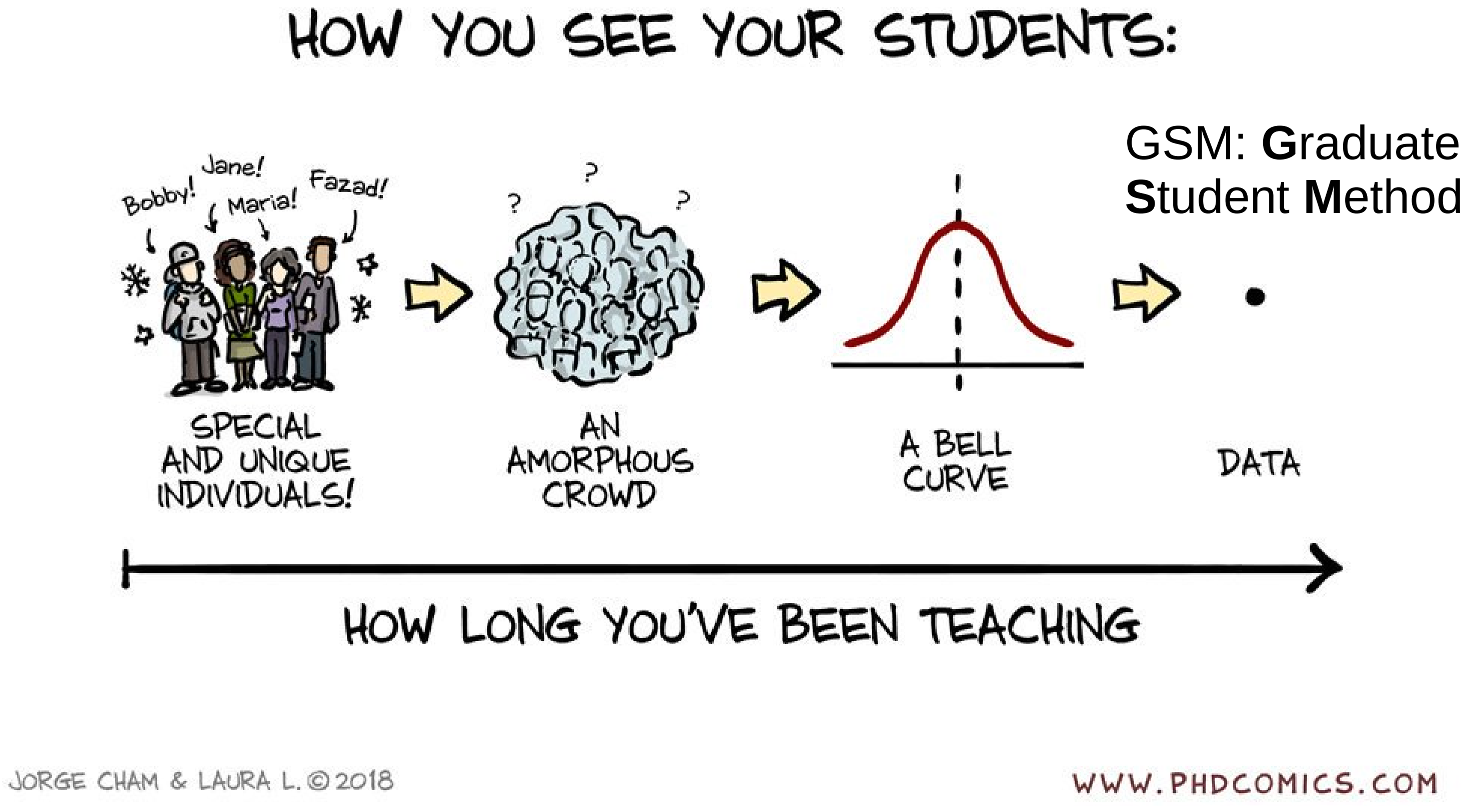
Finding DLAs in Spectra

Currently done by **visual inspection** of spectra

Look for wide dips in the spectrum below (through GSM):



R. Garnett, S. Ho, S. Bird, J. Schneider (2017)



GSM: **G**raduate **S**tudent **M**ethod

Why Machine Learning?

- State-of-art: **visual inspection**, **g**raduate **s**tudent **m**ethod (GSM)
- **No physical model** for quasar emission yet
- Finding DLAs out of weak absorbers in the forest is hard

What are Damped Lyman alpha absorbers (DLAs)?

- Neutral hydrogen gas with a **high column density ($>10^{20.3}$ cm^{-2})**
- **Baryonic acoustic oscillation** (BAO): DLAs, uncertainty in Lyman alpha forest power spectrum
- Ultimately accretes onto galactic halos and **fuels star formation**: hint for galaxy formation
- Total mass of DLAs (OmegaDLA) gives hint for total baryonic matter (OmegaM)

Bayesian Model Selection

- Trick: train a GP on spectra **without DLAs**
- Build another GP for spectra **with DLAs**
- Evaluate **model posterior**:

$$\Pr(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\Pr(\mathcal{M})}{\sum_i p(\mathcal{D} \mid \mathcal{M}_i)\Pr(\mathcal{M}_i)}.$$

Model decisions

- Likelihood function without DLAs

$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, \mathcal{M}_{\neg\text{DLA}}) = \mathcal{N}(\boldsymbol{y}; \mu(\boldsymbol{z}), \boldsymbol{K} + \boldsymbol{\Omega} + \boldsymbol{V}) \tag{1}$$

- Likelihood function with a DLA

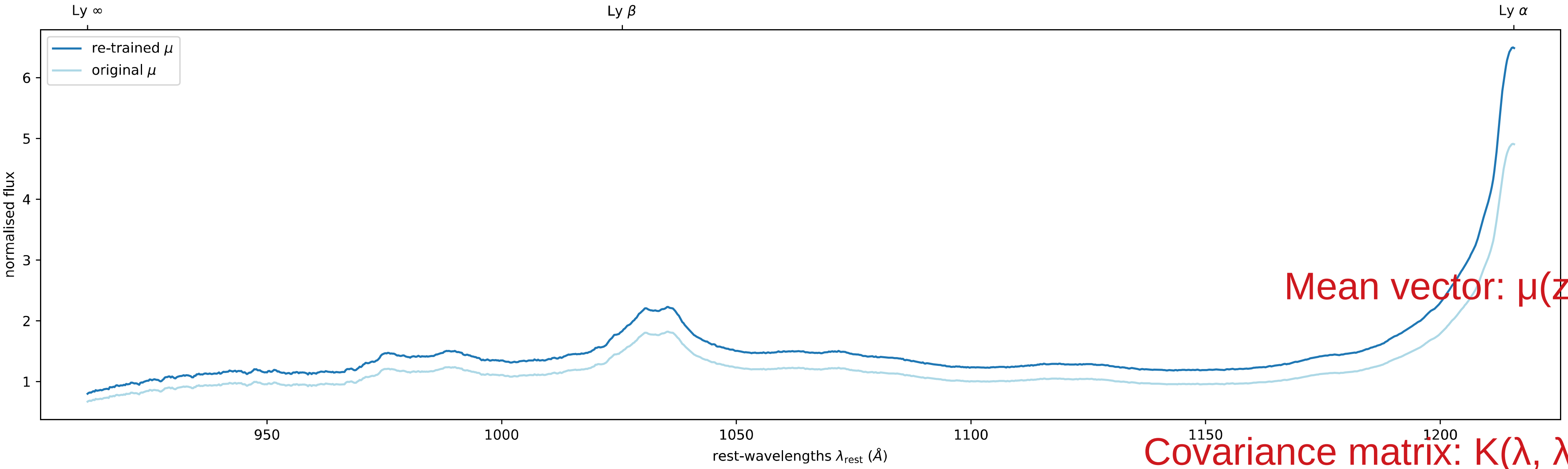
$$p(\boldsymbol{y} \mid \boldsymbol{\lambda}, \boldsymbol{\nu}, z_{\text{QSO}}, z_{\text{DLA}}, N_{\text{HI}}, \mathcal{M}_{\text{DLA}}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{a} \circ \mu(\boldsymbol{z}), \boldsymbol{A}(\boldsymbol{K} + \boldsymbol{\Omega})\boldsymbol{A} + \boldsymbol{V}) \tag{2}$$

with $\boldsymbol{a} = \exp\left(-\tau(\boldsymbol{\lambda}; z_{\text{DLA}}, N_{\text{HI}})\right)$ is the Voigt profile.

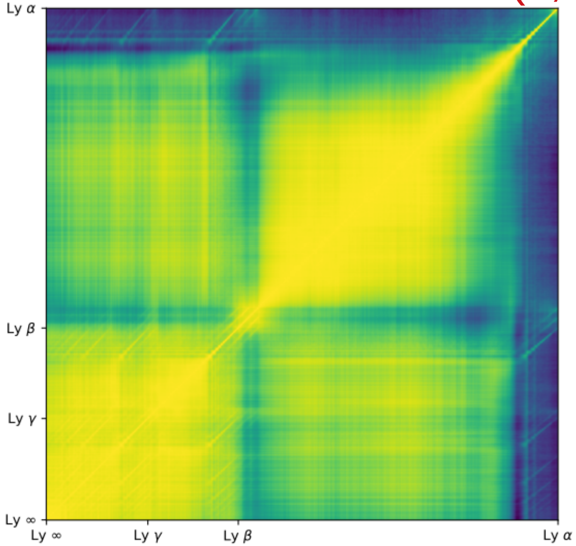
- DLA model evidence: integrate out $\theta = (N_{\text{HI}}, z_{\text{DLA}})$ with parameter priors learned from training data

- Prior for two DLAs in one spectrum $\mathcal{M}_{\text{DLA}(2)}$:

$$\begin{aligned} p(\theta_1, \theta_2 \mid z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(2)}) &= \\ &= p(\theta_1 \mid z_{\text{QSO}}, \mathcal{D}, \mathcal{M}_{\text{DLA}(1)}) p(\theta_2 \mid z_{\text{QSO}}, \mathcal{M}_{\text{DLA}(1)}) \end{aligned} \tag{3}$$



Covariance matrix: K(λ, λ')



- Maximise the null model log likelihood $\mathcal{L}(\boldsymbol{K}, \boldsymbol{\Omega})$ on the training set (SDSS DR9)

  where $\boldsymbol{K} = \boldsymbol{M}\boldsymbol{M}^\top$ and $\boldsymbol{M}$ is an $(N_{\text{pixel}} \times k)$ matrix, with k is the number of eigenspectra.

- You can think it's an fancy way to find principal components $\boldsymbol{M}$ with the consideration of absorption noise Ω.

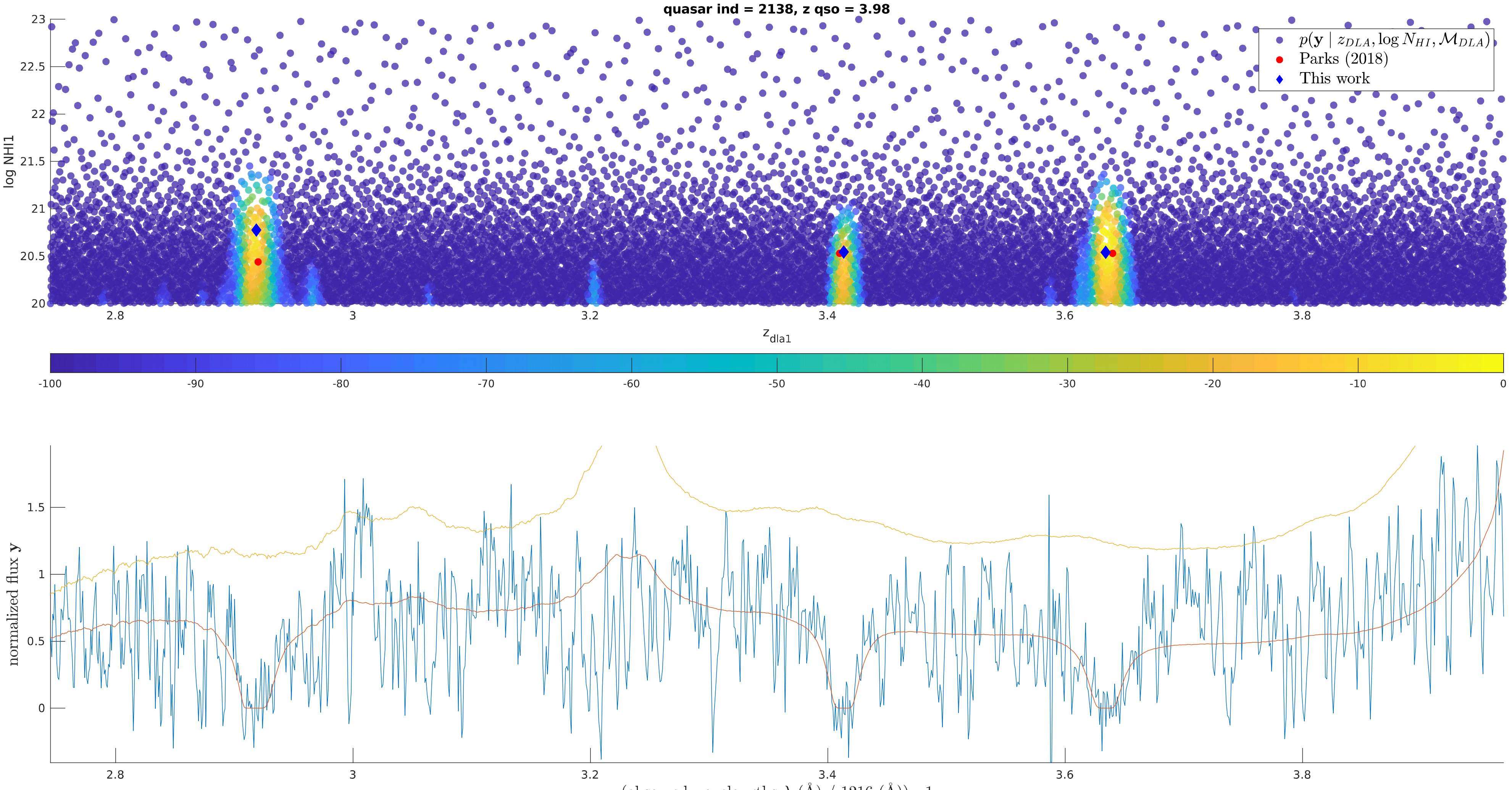
Bayesian Model Selection

- Evaluate model posteriors of each model with Bayesian model selection

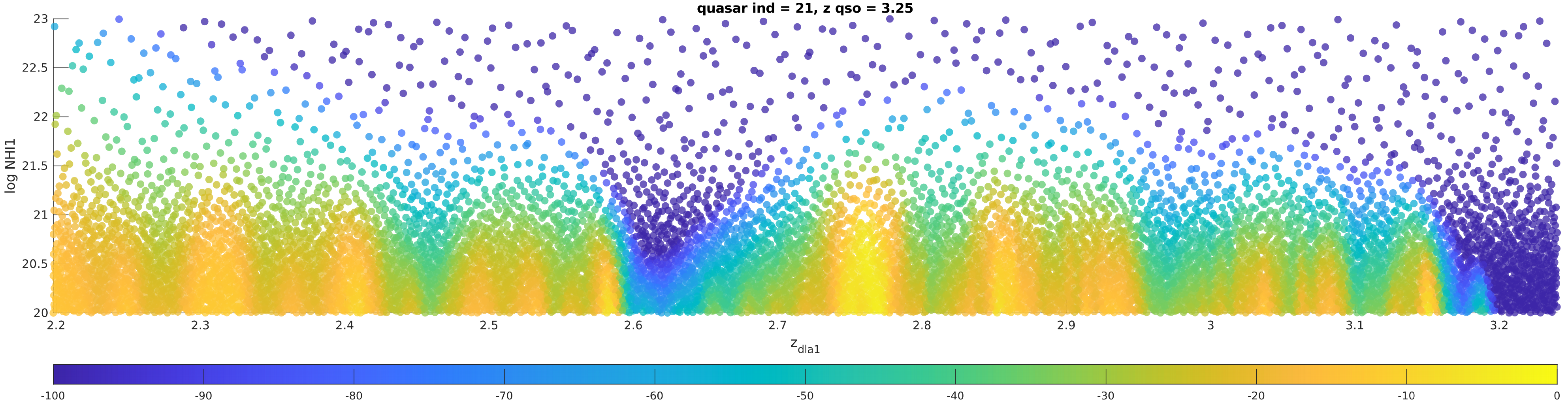
$$\Pr(\mathcal{M}_{\mathrm{DLA}\,(i)} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M}_{\mathrm{DLA}\,(i)})\Pr(\mathcal{M}_{\mathrm{DLA}\,(i)})}{p(\mathcal{D} \mid \mathcal{M}_{\mathrm{sub}})\Pr(\mathcal{M}_{\mathrm{sub}}) + \sum_{i=0}^{k} p(\mathcal{D} \mid \mathcal{M}_{\mathrm{DLA}\,(i)})\Pr(\mathcal{M}_{\mathrm{DLA}\,(i)})}.$$

- The alternative model $\mathcal{M}_{\mathrm{sub}}$ is used for regularisation.

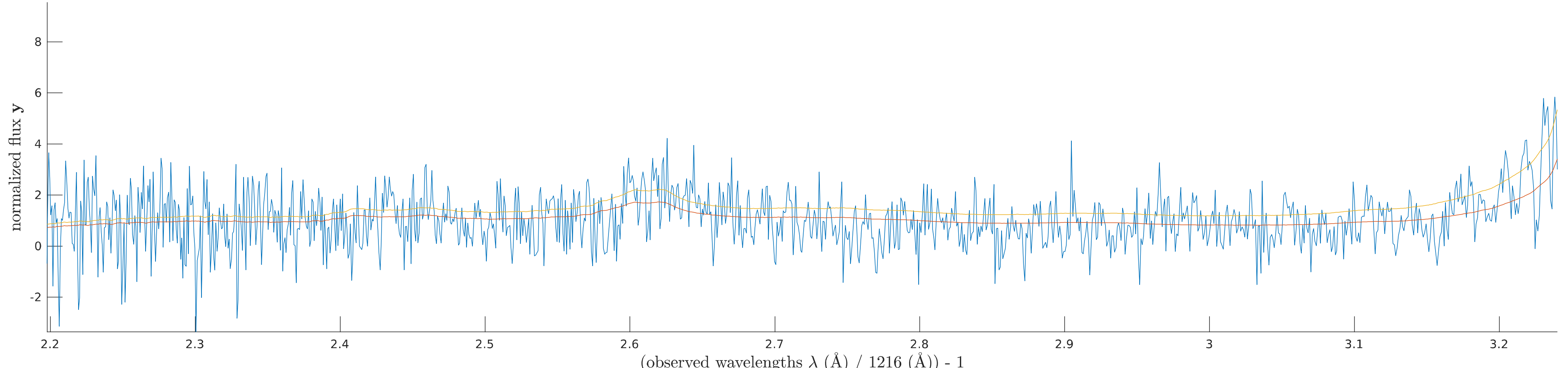
Example: 3 DLAs



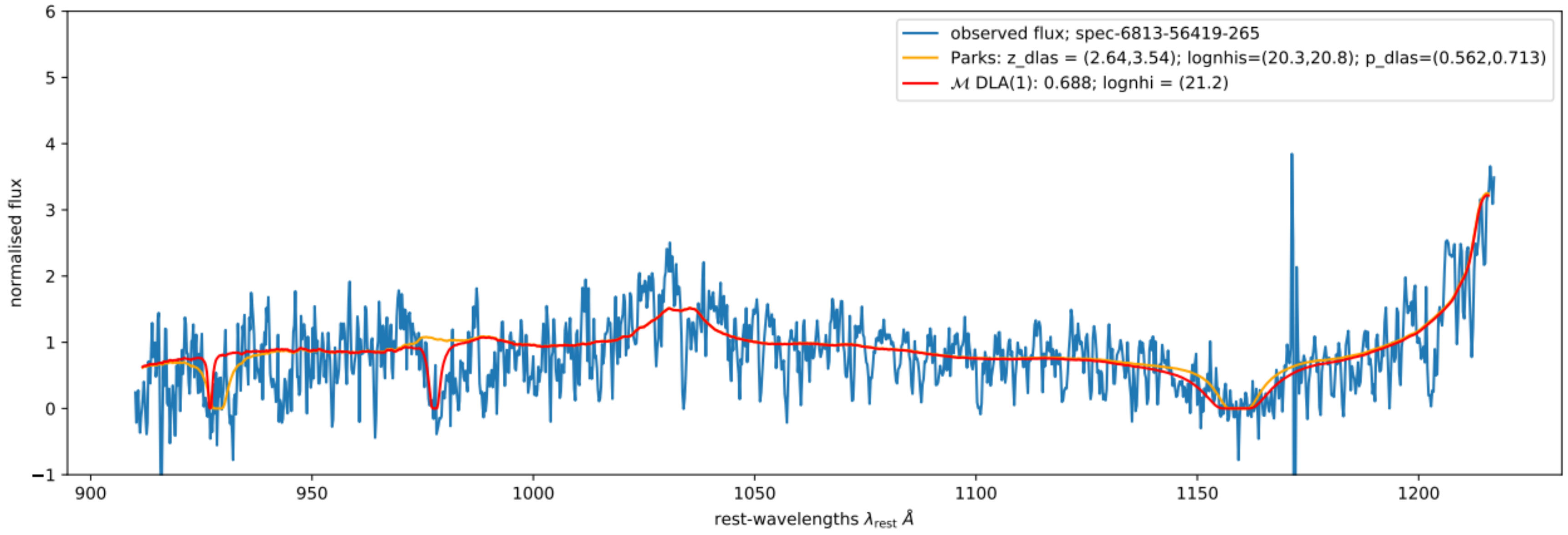
Estimate uncertainty in a fully-Bayesian way



As a Bayesian, you don’t give an answer. You give a bunch of answers.

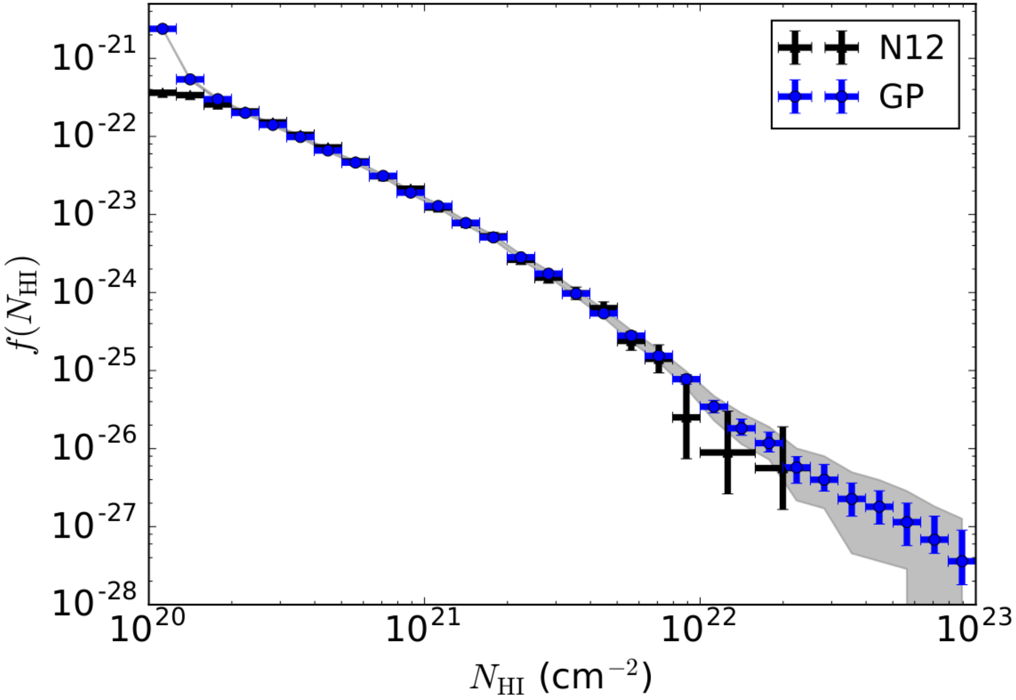


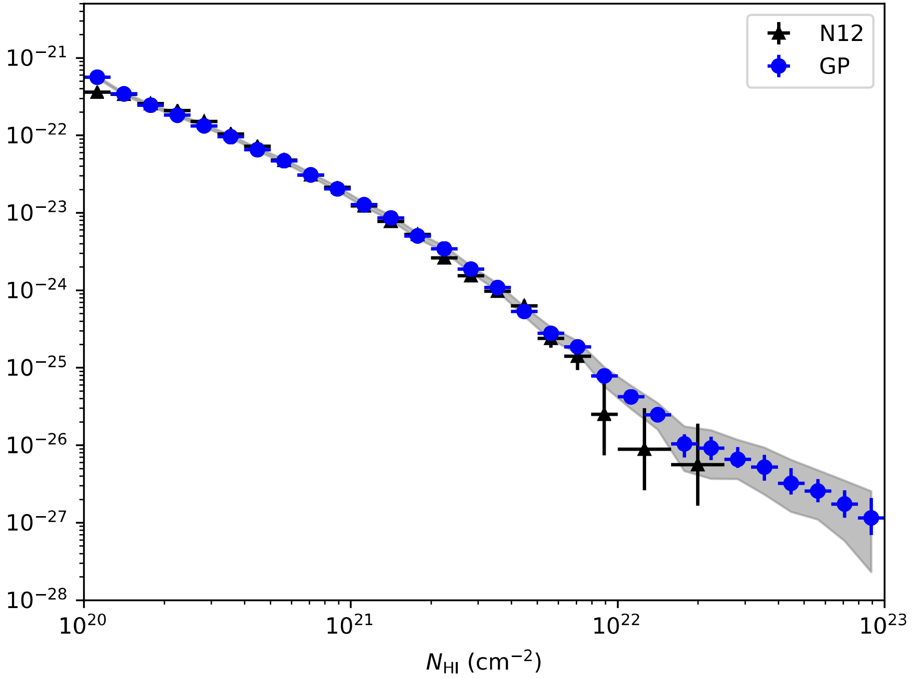
Our model includes beta absorption of DLAs



Orange curve: Parks (2018), A convolutional neural network model for finding DLAs

Results: CDDF



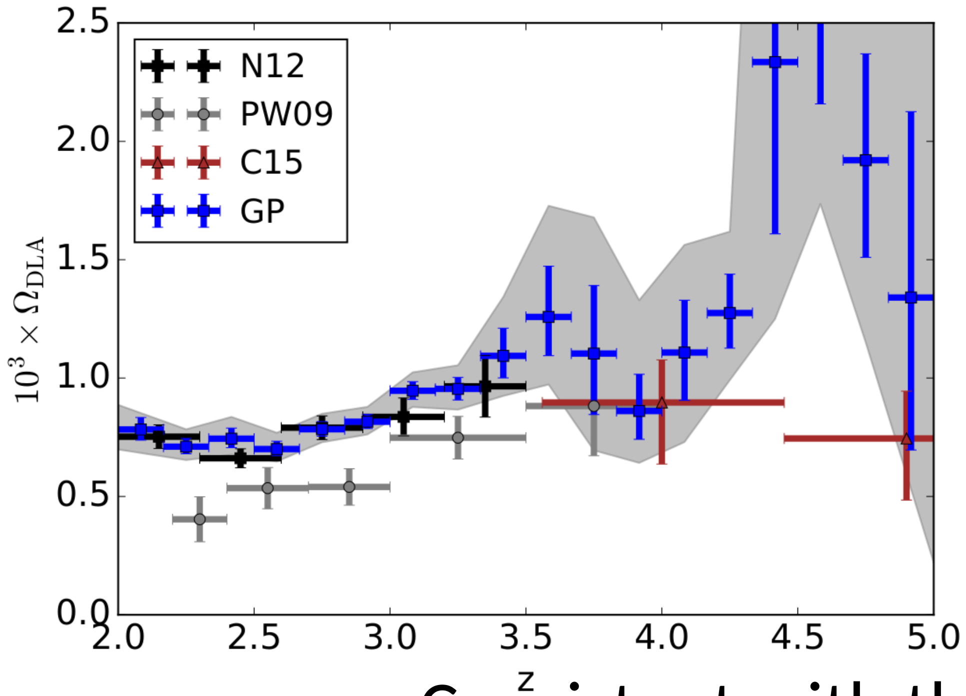


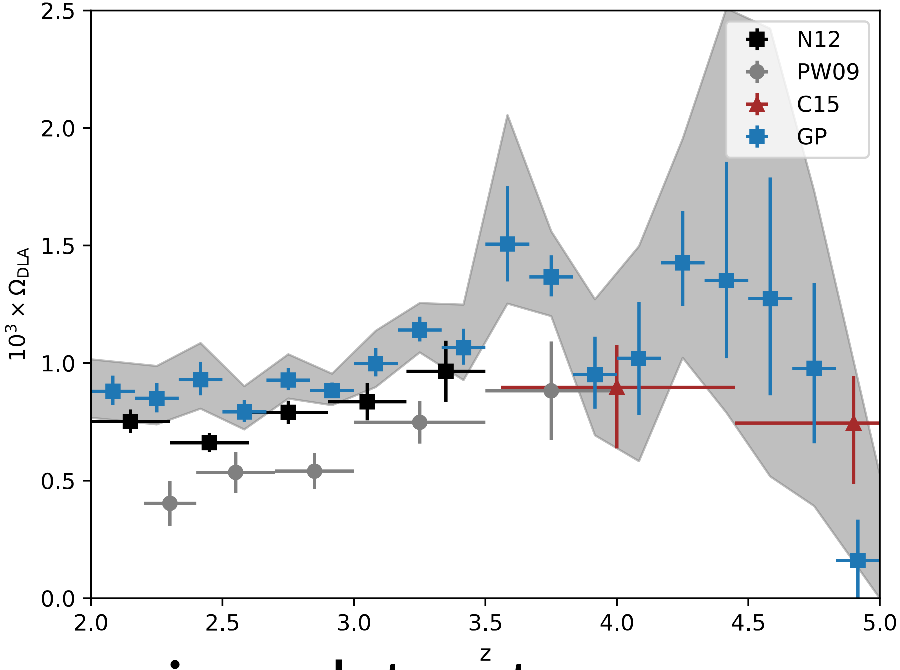
- **Use all data** (DR12), even with SNR < 1
- Precise measurement of HI

S. Bird, R. Garnett, S. Ho (2017)

M.-F. Ho, S. Bird, R. Garnett (in prep)

Results: Total Mass of DLA





- Consistent with the previous dataset, with predictions for z > 4

S. Bird, R. Garnett, S. Ho (2017)

M.-F. Ho, S. Bird, R. Garnett (in prep)

Conclusion

- Automated detection of DLAs
- We get a posterior density per spectrum
- Regularise our previous model: extend to arbitrary number DLAs without overfitting

Our previous model is publicly available :
https://github.com/rmgarnett/gp_dla_detection/

Me:
https://github.com/jibanCat