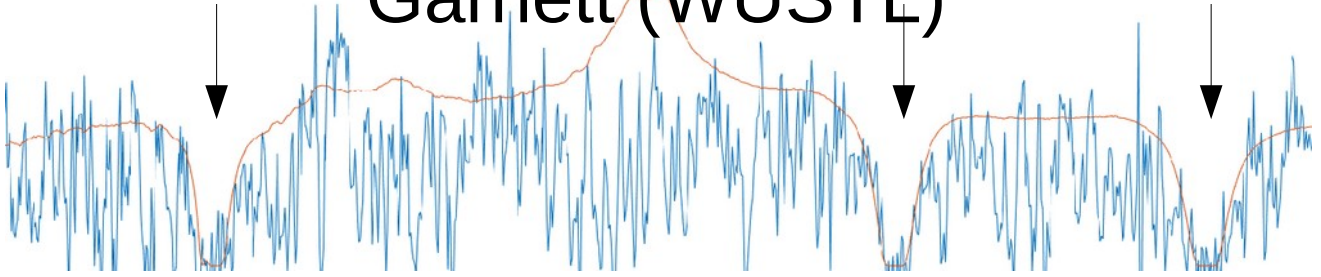
Finding Strong Lyα Absorbers in the Shadows of Quasars with Bayesian Machine Learning

https://arxiv.org/abs/2003.11036 https://github.com/rmgarnett/gp_dla_detection/

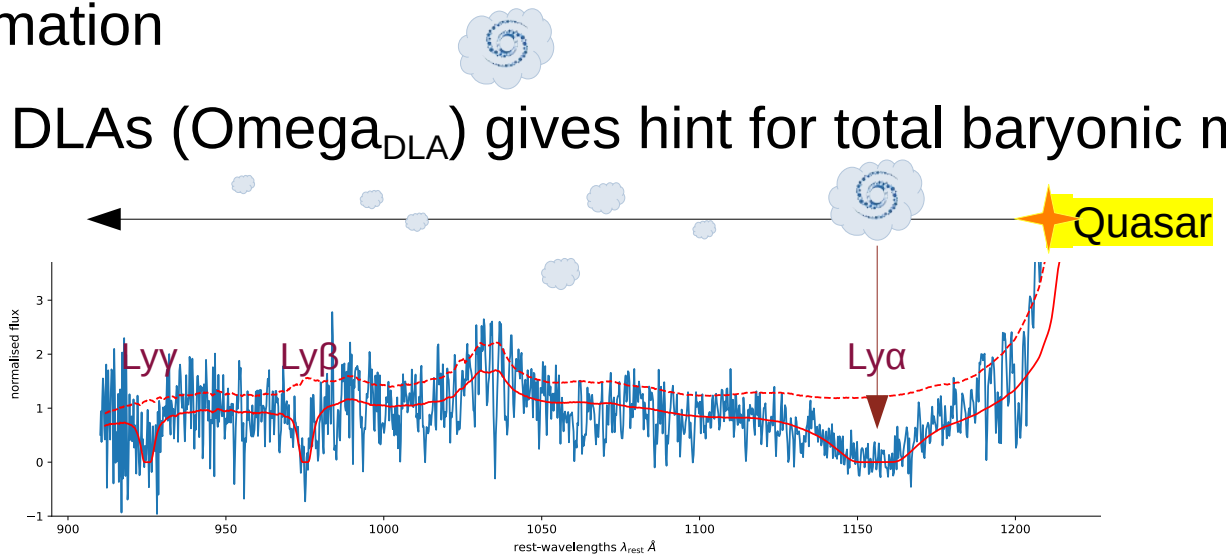
Ming-Feng Ho (Me) (UCR), Simeon Bird (UCR), Roman Garnett (WUSTL)



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

What are Damped Lyman alpha absorbers (DLAs)?

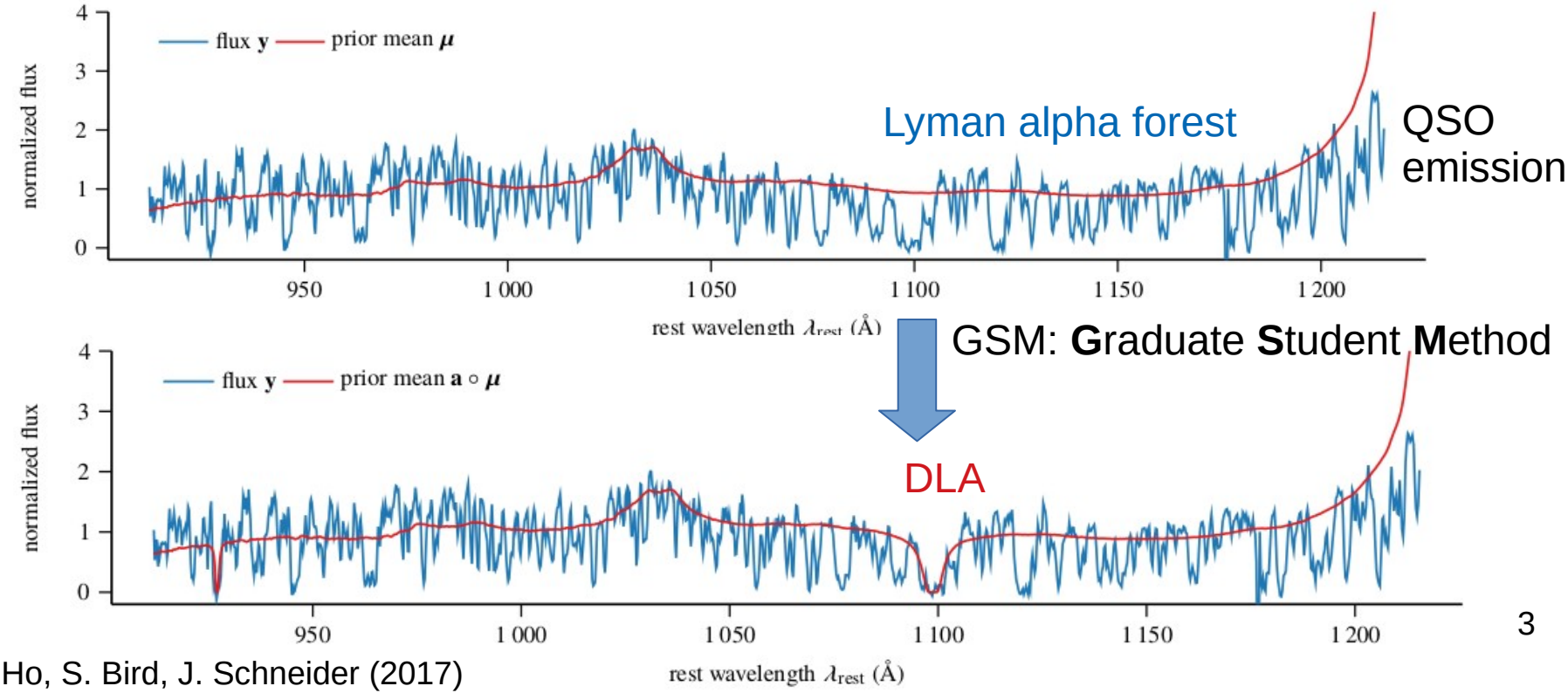
- Neutral hydrogen gas with a **high column density (>$10^{20.3}$ cm^{-2})**
- **Baryonic acoustic oscillation** (BAO): DLAs, uncertainty in Lyman alpha forest power spectrum
- Ultimately accretes onto galactic halos and **fuels star formation**: hint for galaxy formation
- Total mass of DLAs ($Omega_{DLA}$)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



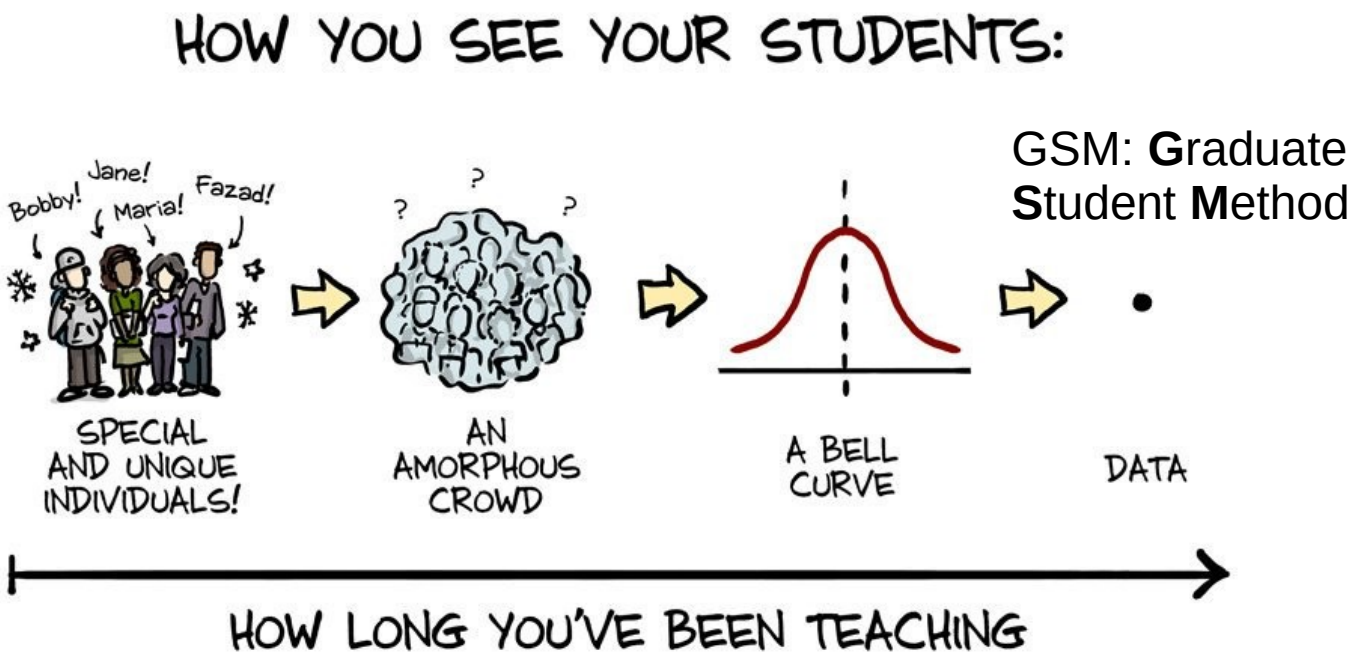
Finding DLAs in Spectra

Currently done by **visual inspection** of spectra

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



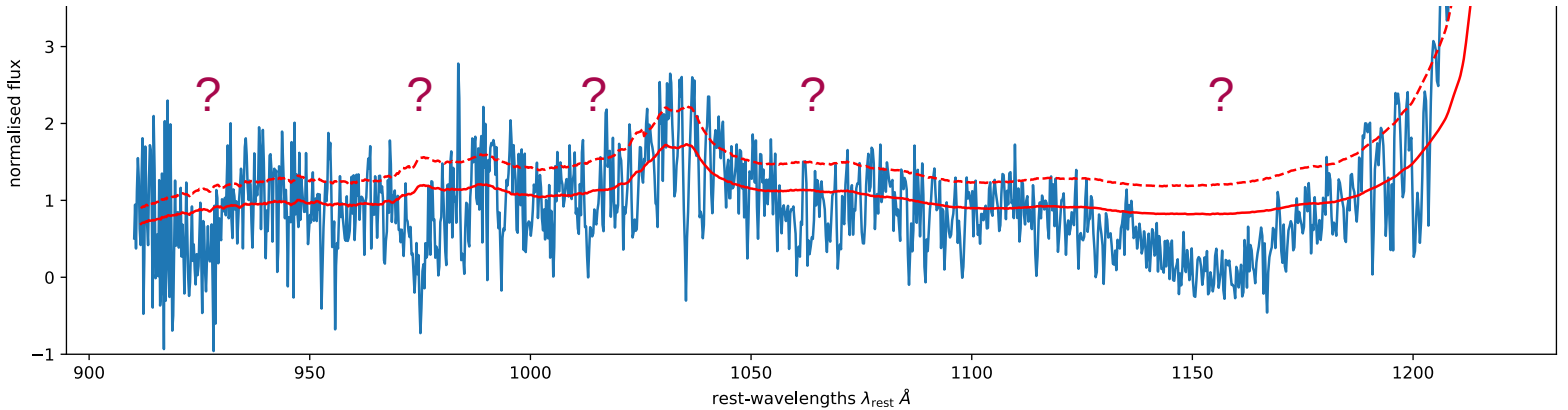
R. Garnett, S. Ho, S. Bird, J. Schneider (2017)



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

Why Machine Learning?

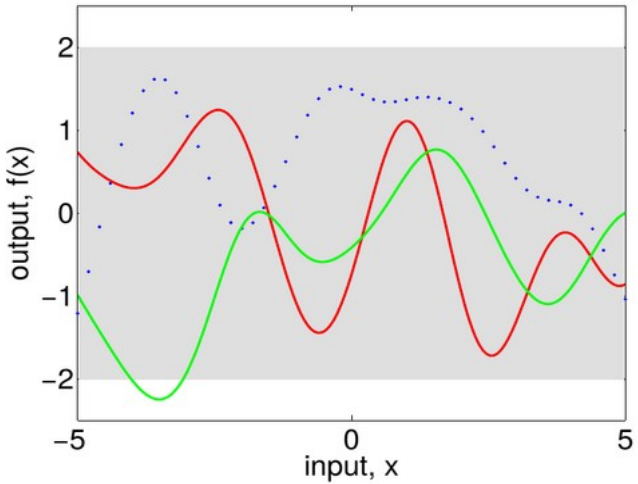
- State-of-art: **visual inspection**, **g**raduate **s**tudent **m**ethod (GSM)
- **No physical model** for quasar emission yet
- Finding DLAs out of weak absorbers in the forest is hard



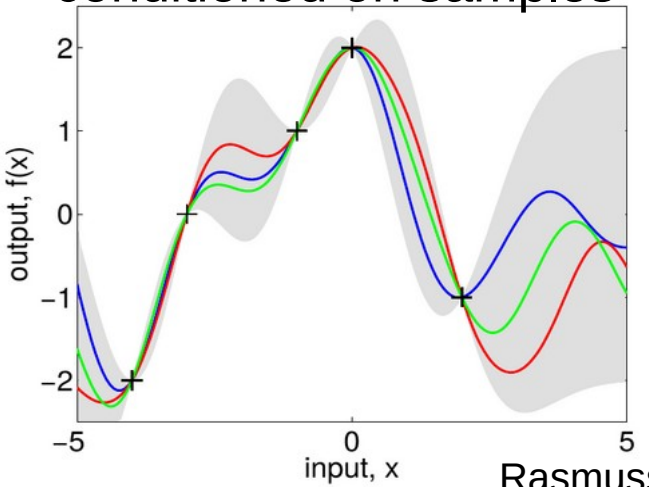
What is a **G**aussian **P**rocess?

- Quasar model is a Gaussian Process
- Bayesian function interpolation, which computes probability distribution of f(x) conditional on input set.

Left: Prior on function



Right: Posterior conditioned on samples

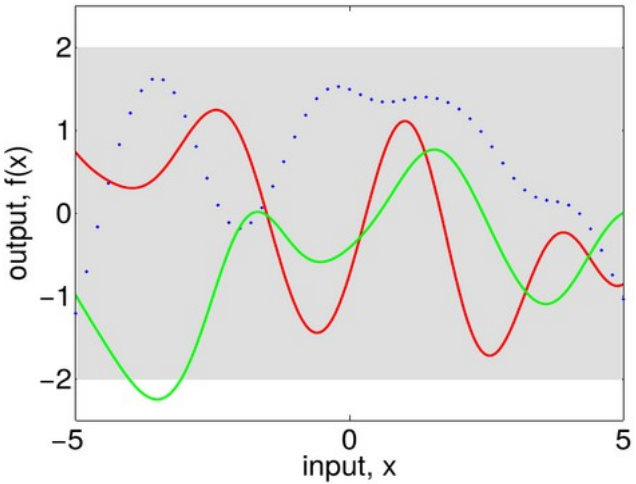


Rasmussen & Williams (GPML)

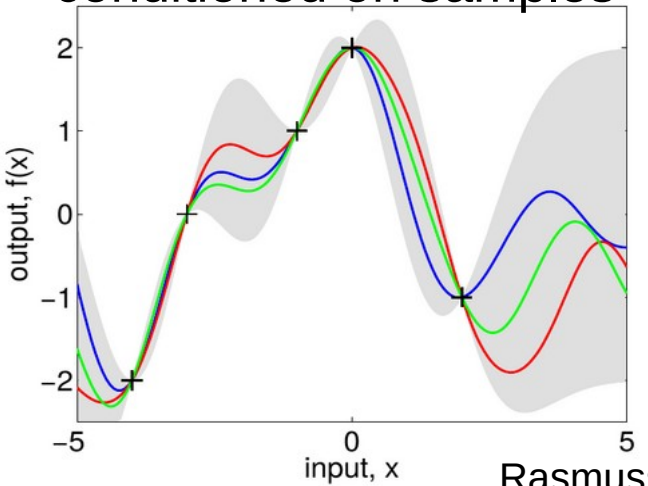
What is a **G**aussian **P**rocess?

- Magic in **Kernel** function: describes how **correlation between function values** depends on parameter distance.
- We trained our kernel from quasar spectra; it describes the correlation between different emission lines.

Left: Prior on function



Right: Posterior conditioned on samples



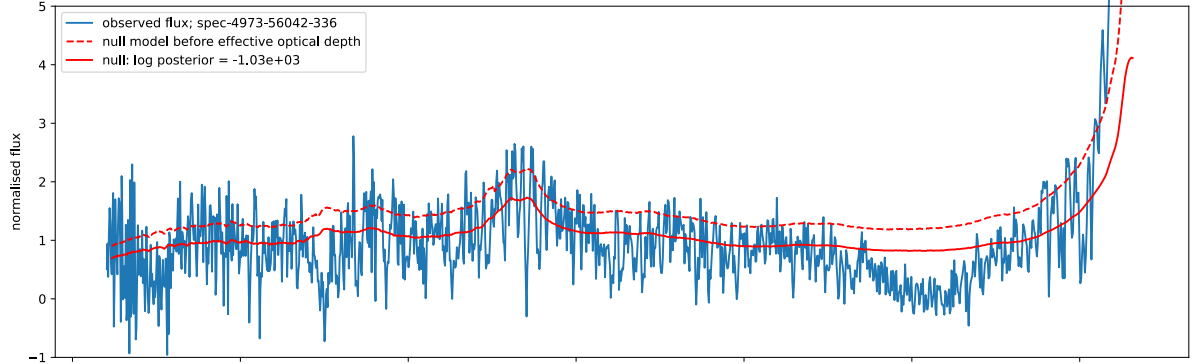
Rasmussen & Williams (GPML)

Bayesian Model Selection

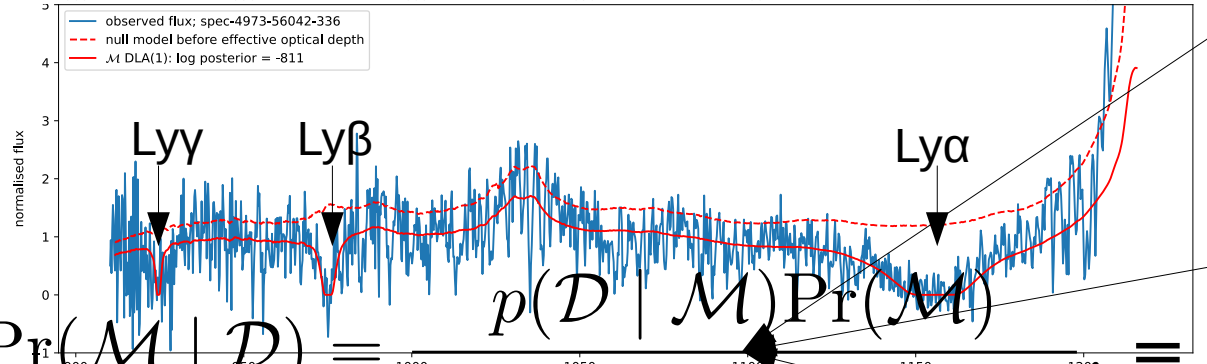
- Trick: train a GP on spectra **without DLAs**
- Build another GP for spectra **with DLAs**
- Evaluate **model posterior**:

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

Bayesian Model Selection



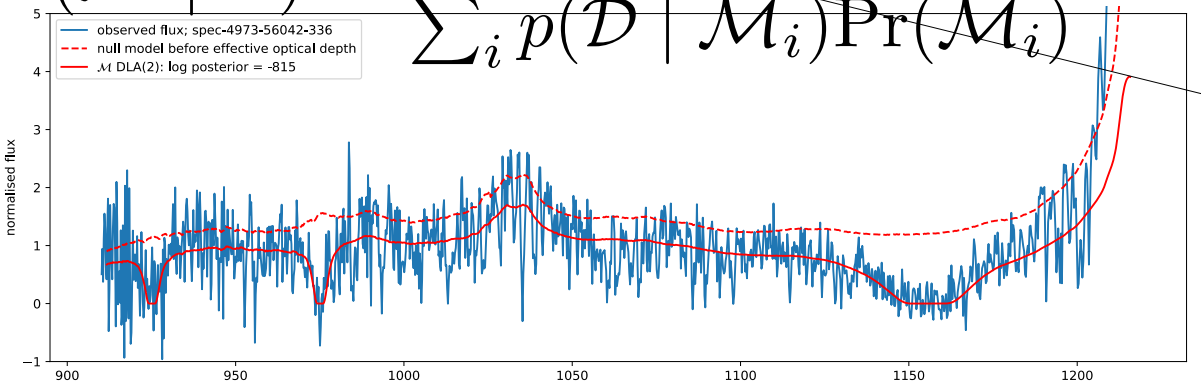
log P(**0 DLA** | D) = -1030



Select!

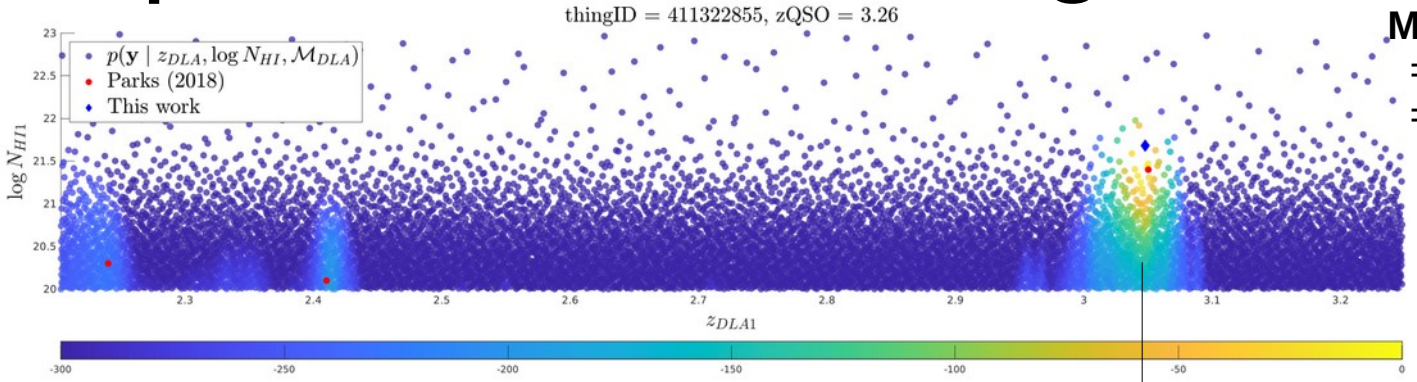
log P(**1 DLA** | D) = -811

$$\Pr(\mathcal{M} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathcal{M})\Pr(\mathcal{M})}{\sum_i p(\mathcal{D} \mid \mathcal{M}_i)\Pr(\mathcal{M}_i)} = 0.989$$

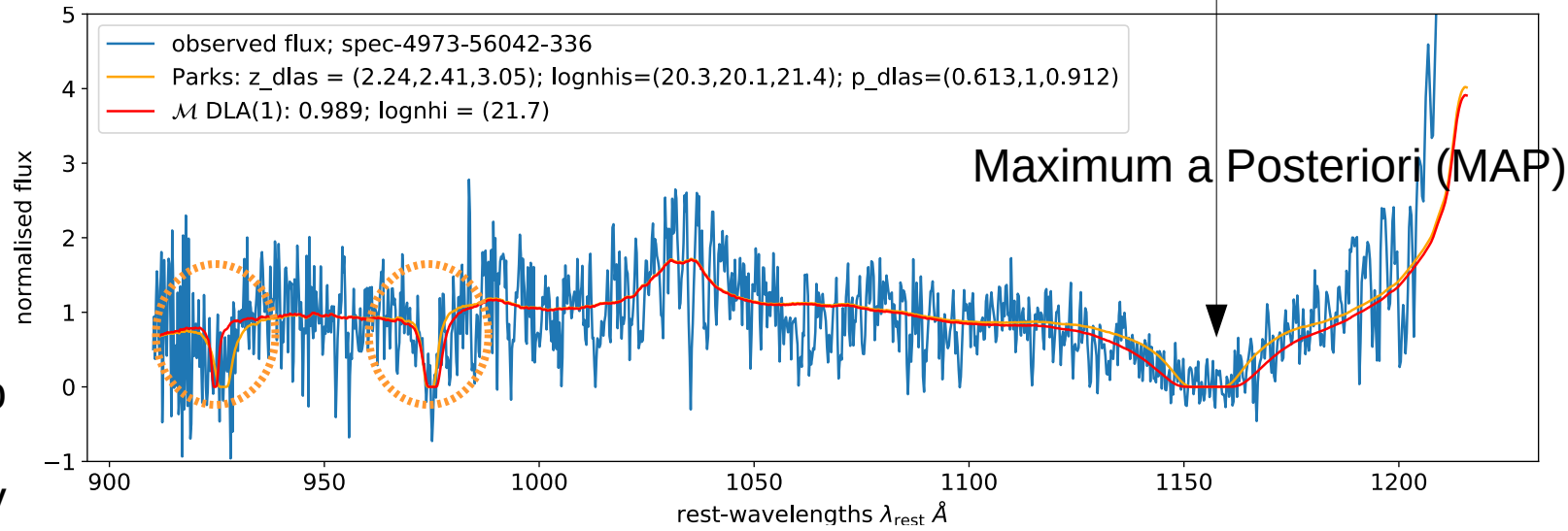


log P(**2 DLAs** | D) = -815

Sum up all possible DLAs → the posterior of having DLAs



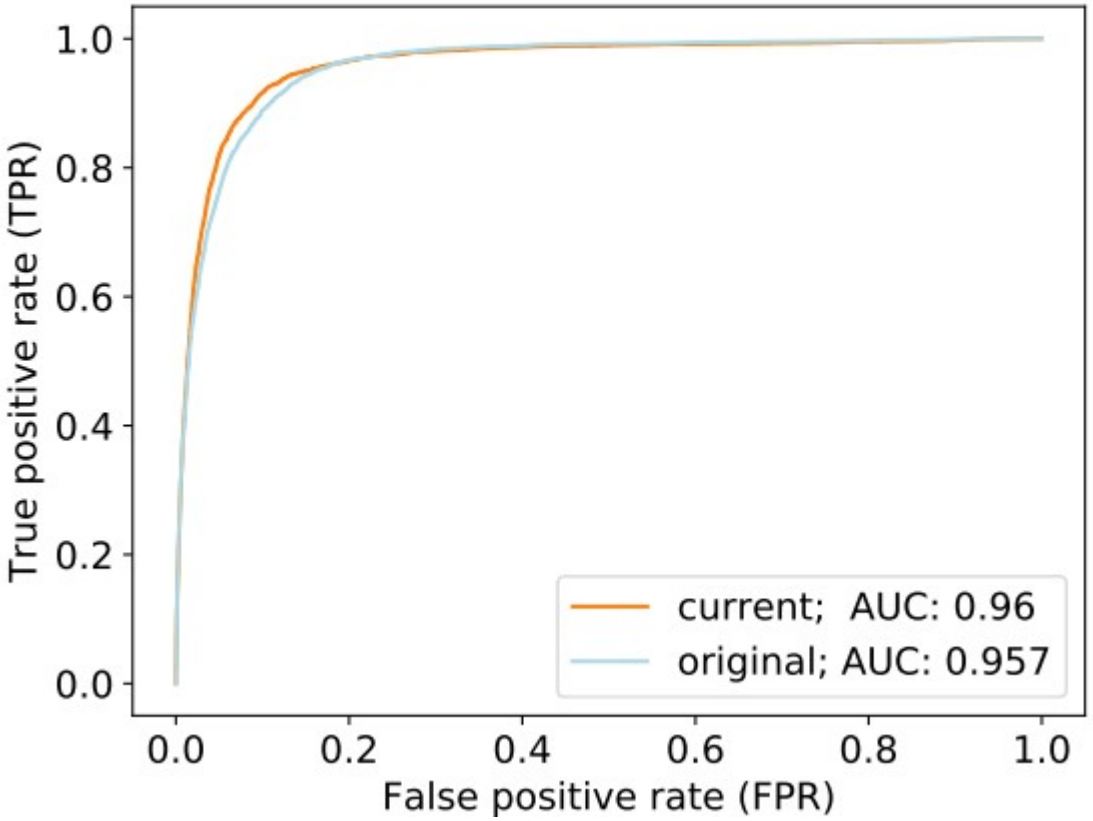
Marginal likelihood
= log P(**D | 1 DLA**)
= -808



Maximum a Posteriori (MAP)

Important to also model Lyβ and Lyγ

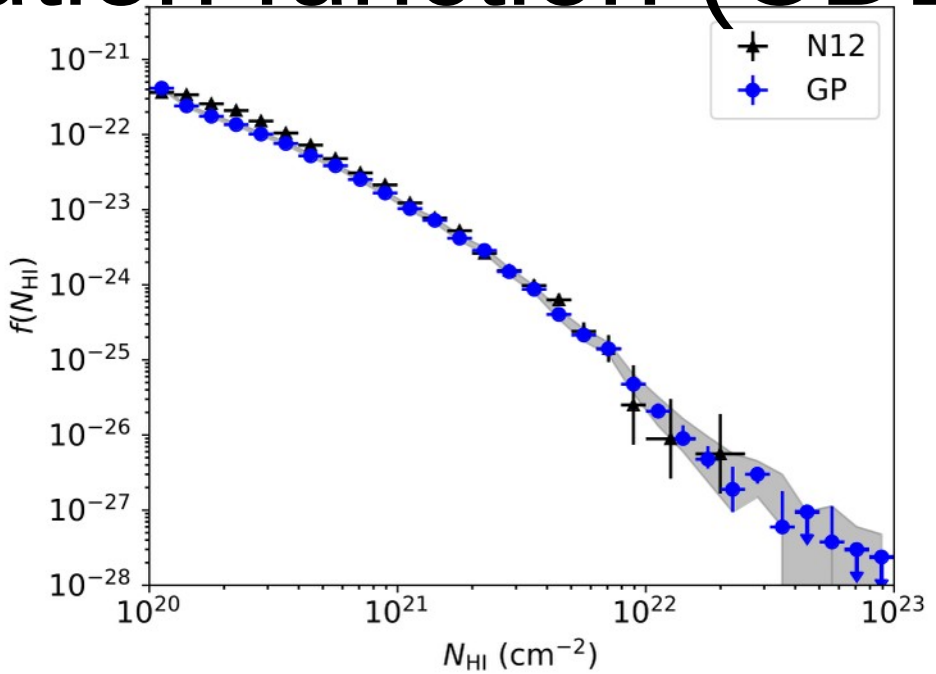
Results: A decent ROC



A better handle for those borderline cases with logNHI~20

M.-F. Ho, S. Bird, R. Garnett (2020)

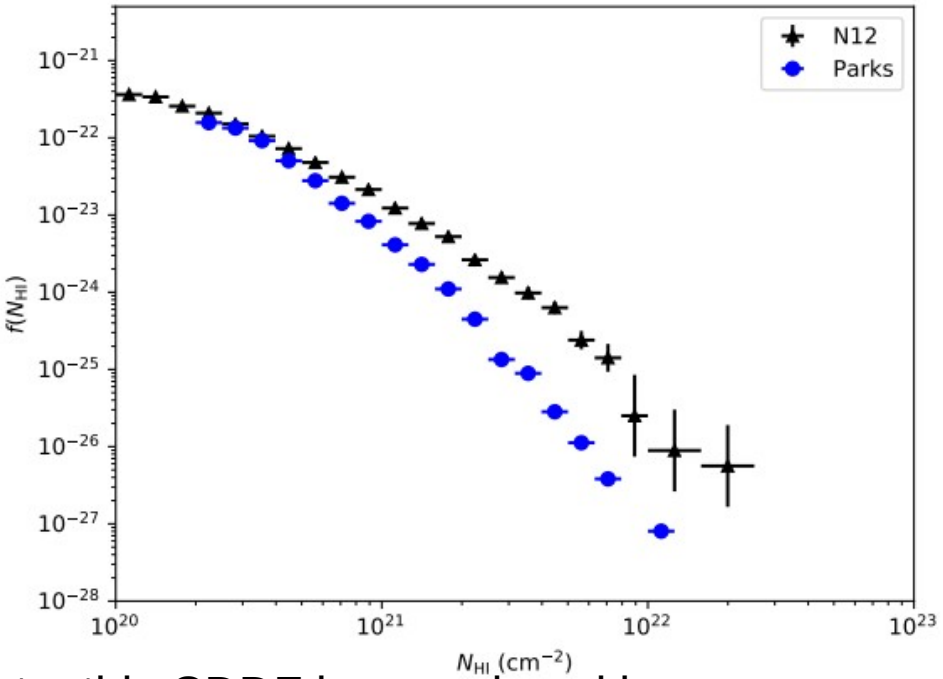
Results: Column density distribution function (CDDF)

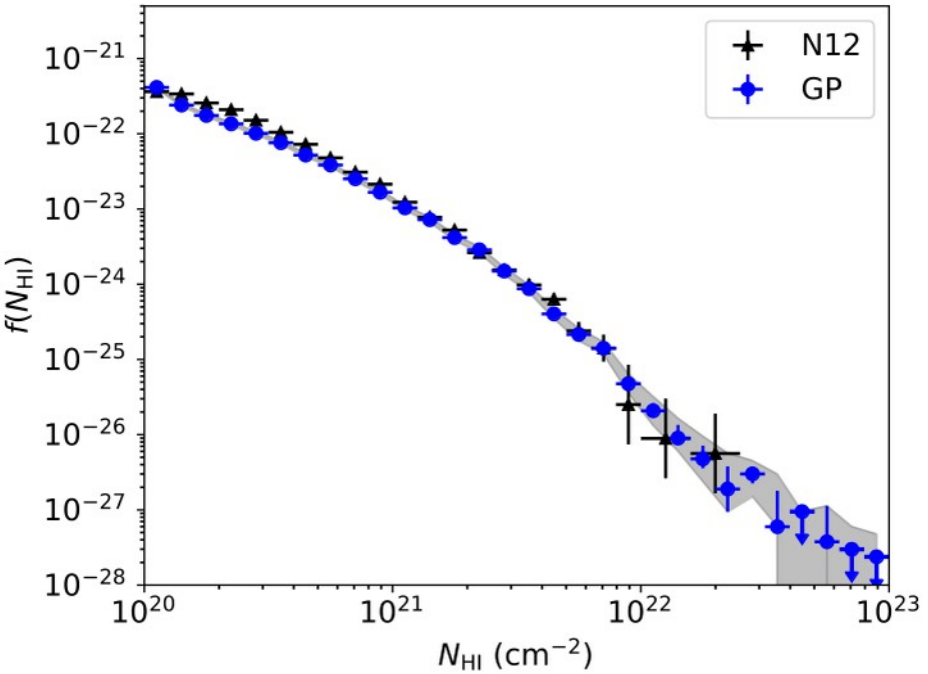


- **Use all data** (DR12), even with SNR < 1
- Properly **Propagate the uncertainty** to NHI
- Reproduce previous catalogue (Noterdaeme 2012)

M.-F. Ho, S. Bird, R. Garnett (2020)

Parks' CDDF





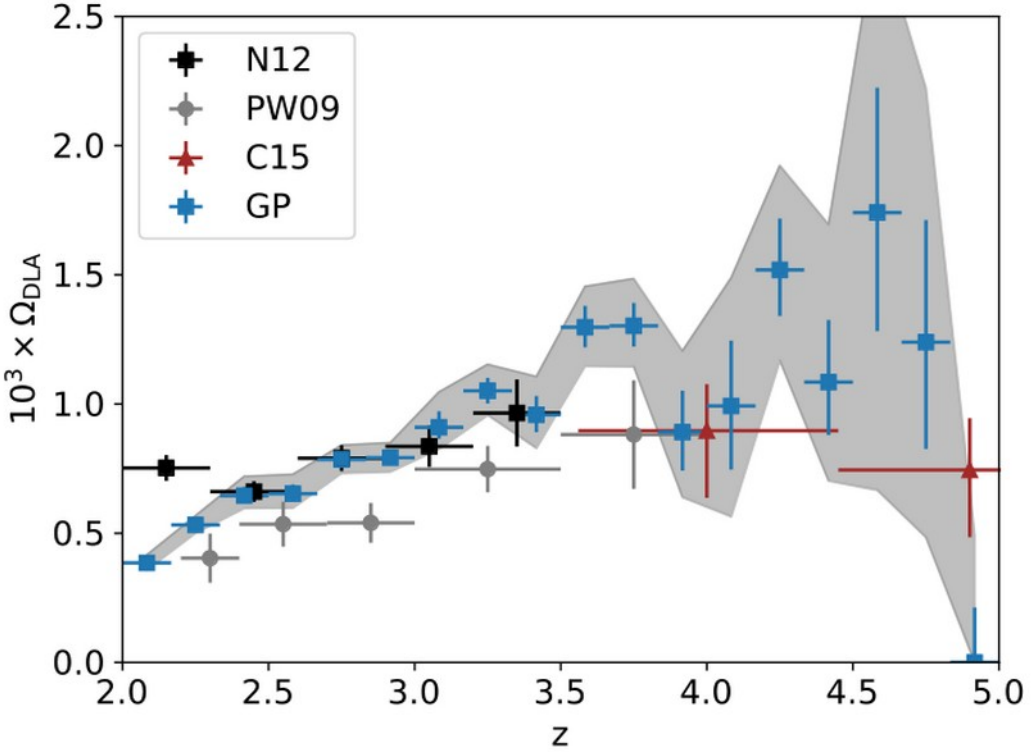
Note: this CDDF is reproduced by us using Parks' catalogue.

- More **precise measurement** than Parks on the **high-end**

D. Parks, J. X. Prochaska, S. Dong, Z. Cai (2018)

M.-F. Ho, S. Bird, R. Garnett (2020)

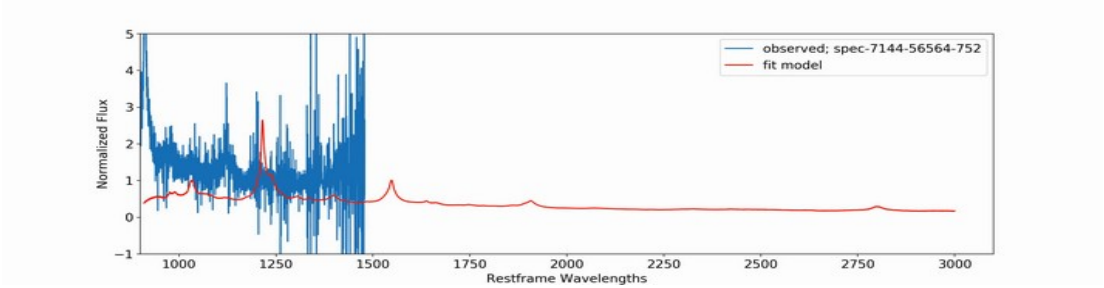
Results: Total Mass of DLA



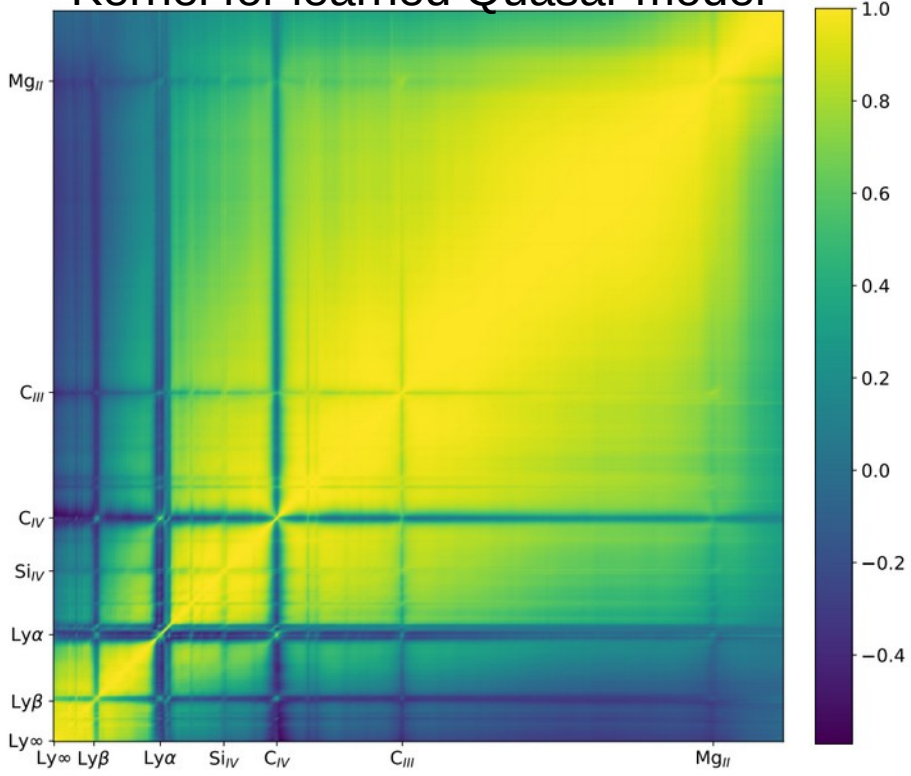
- Count all **weak detections**
- Extend measurements to **z > 4.0**

M.-F. Ho, S. Bird, R. Garnett (2020)

Bonus: Quasar redshift estimation using a GP



Kernel for learned Quasar model



with UCR CS people:
Leah Fauber, Chrisitian Shelton, Ishita Korde

https://arxiv.org/abs/2006.07343 Fauber et al. (2020)

Conclusion

- **Automated** the detection of **DLAs**
- We get a **posterior density** **per spectrum**
- Decent accuracy; better NHI estimations
- Can also **estimate zQSO**

https://arxiv.org/abs/2003.11036

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

Me:
https://github.com/jibanCat

Redshift estimation code with GP:
https://github.com/sbird/gp_qso_redshift

Thank Reza Mondai & Madi Qezlou(UCR Astro):
for valuable comments.
Thank Yongda Zhu and Marie Wingyee Lau (UCR, Astro):
for useful discussions on quasar continua.
Thank Julia Plank (UAuckland, Psych):
for ~~listening to my complaints~~ oversea psychological supports during pandemic

How do we count expected number of DLAs at a given redshift bin?

How we count DLAs on column density bins?

- Having posterior density: larger sample size and smaller error bars
- Compute Pr(N) with Poisson-Binomial process

$$\Pr(N) = \sum_{\mathrm{DLA} \in F_N} \prod_{i \in \mathrm{DLA}} p^i_{\mathrm{DLA}}(\{\mathcal{M}_{\mathrm{DLA}}\} \mid Q) \prod_{j \in \mathrm{DLA}^c} (1 - p^j_{\mathrm{DLA}}(\{\mathcal{M}_{\mathrm{DLA}}\} \mid Q))$$

where F_N corresponds to all subsets of N integers can be selected from the sequence $\{1, 2, ..., n\}$. Bin Q is an interval in the parameter space of column density or DLA redshift $Q \in \{N_{\mathrm{HI}}, z_{\mathrm{DLA}}\}$

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