Constraining Lyman continuum escape using Machine Learning

Sambit Giri

People involved:

Sambit Giri & Garrelt Mellema

Stockholm University

Oskar Klein centre

Erik Zackrisson, Christian Binggeli, Kristiaan Pelckmans & Ruben Cubo

UPPSALA UNIVERSITET

Why is this interesting?

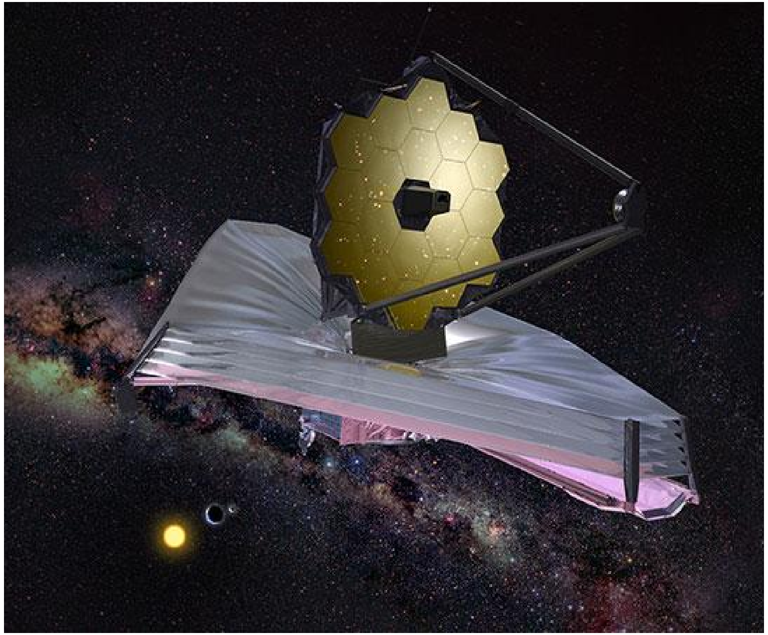
Previous talk by Christian Binggeli

- Understanding the galaxy-based reionization
- Detect more high LyC leakers
- Dealing with Big Data

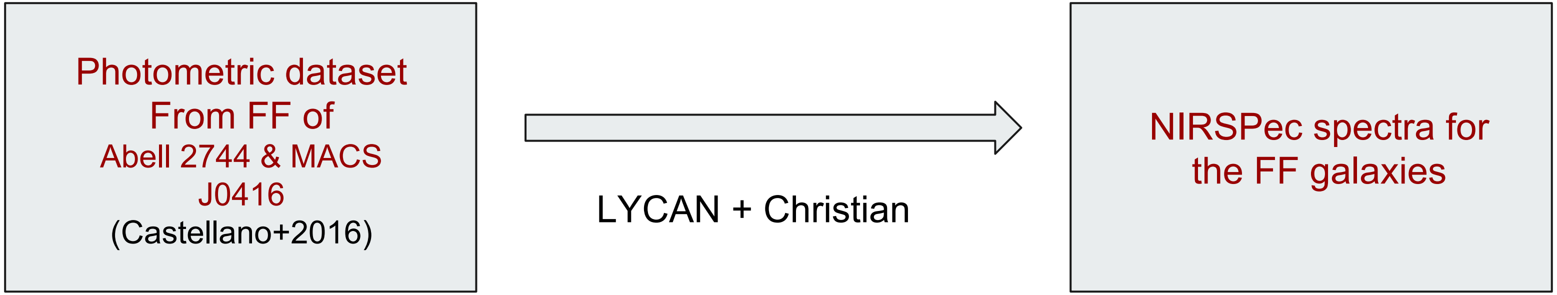
Aim

Machine Learning tools to work on the JWST/NIRSpec spectra

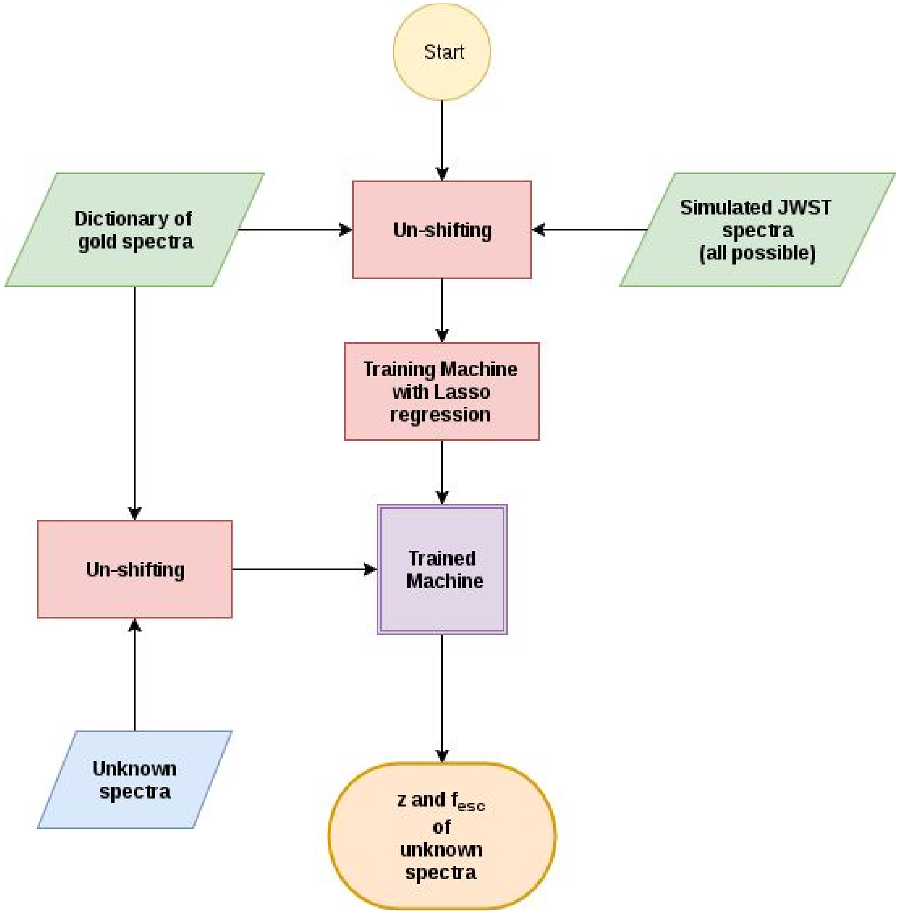
Connect the galaxy simulation to the observed spectra



Simulating the JWST/NIRSpec spectra for the Frontier Field

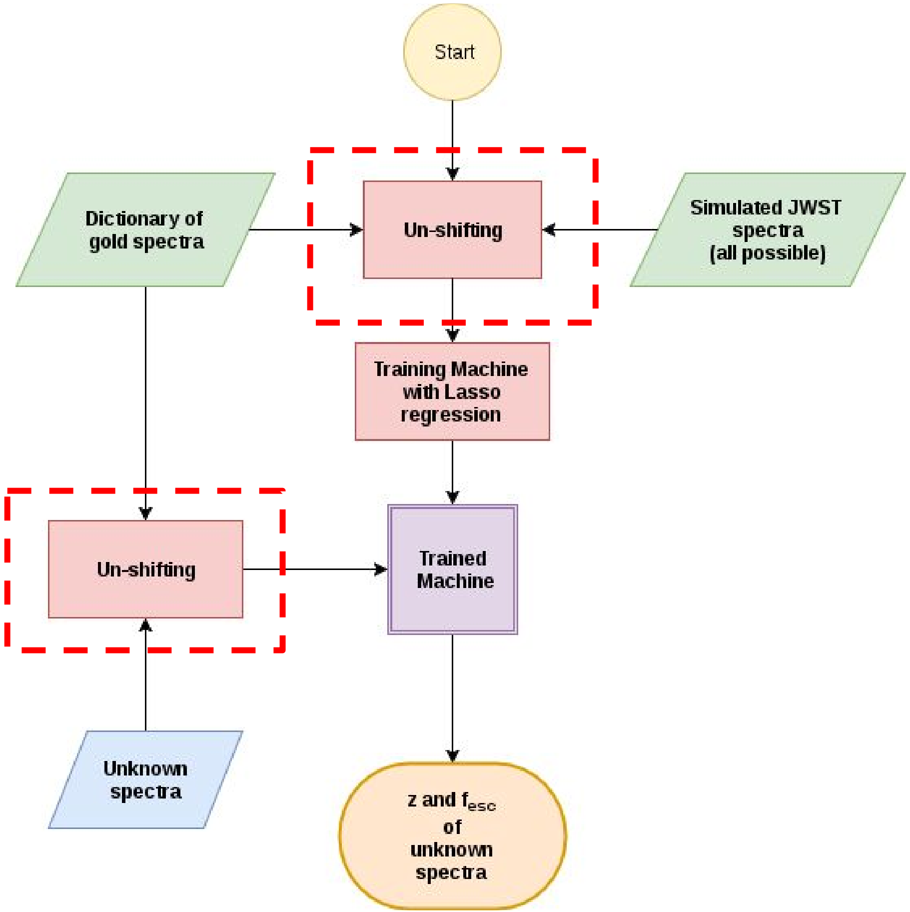


Flowchart



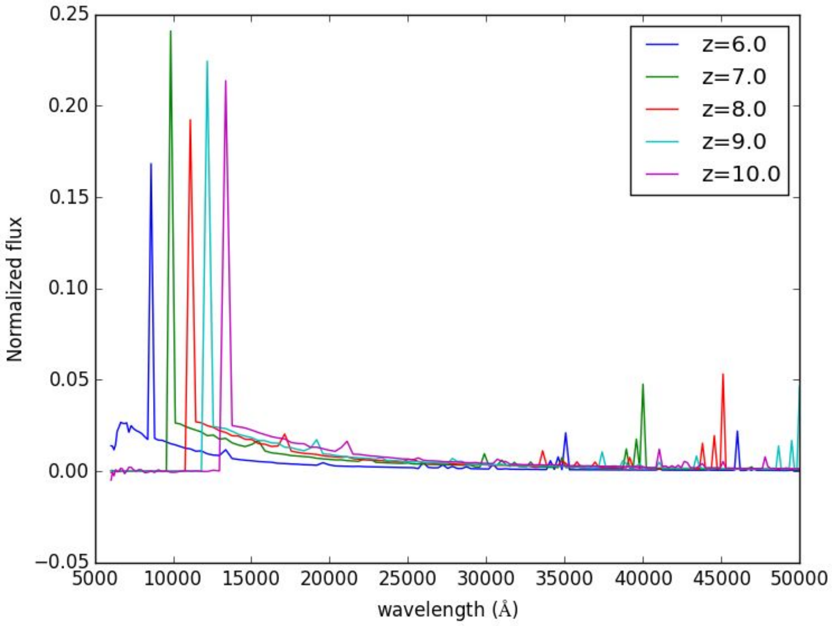
Start
Dictionary of gold spectra
Un-shifting
Simulated JWST spectra (all possible)
Training Machine with Lasso regression
Un-shifting
Trained Machine
Unknown spectra
z and f_{esc} of unknown spectra

Flowchart


Start
Dictionary of gold spectra
Un-shifting
Simulated JWST spectra (all possible)
Training Machine with Lasso regression
Un-shifting
Trained Machine
Unknown spectra
z and f_{esc} of unknown spectra

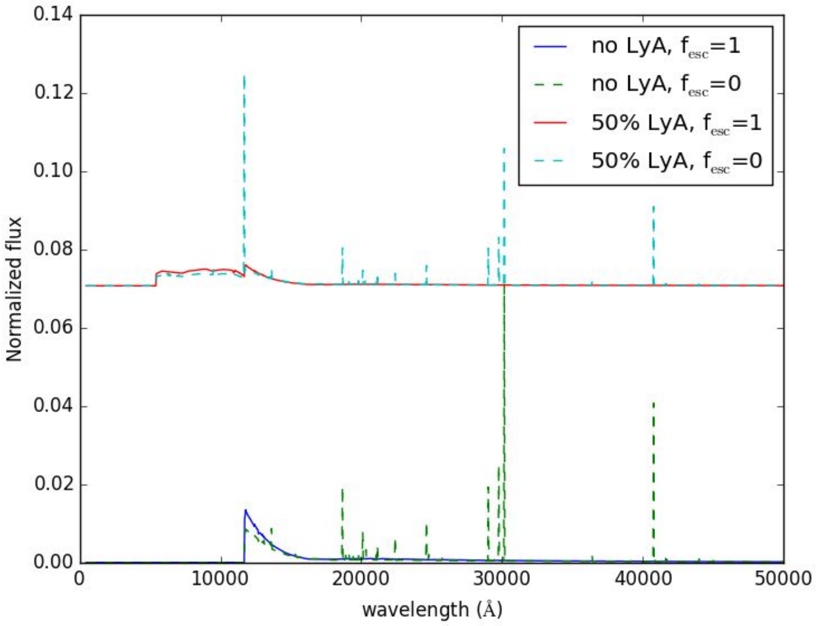
Un-shifting

- The spectra is shifted due to redshifting
- Jensen et al (2016) showed that the f_{esc} can be predicted from the fluxes at a particular redshift z



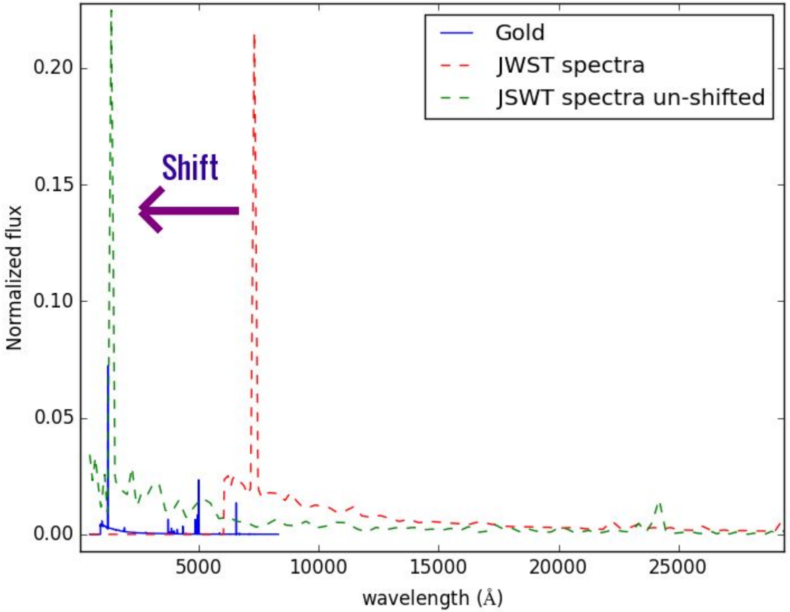
Un-shifting

- We shift the given spectra to the Gold-standard spectra (GSS)
- GSS has the best possible wavelength resolution and no-noise
- GSS dictionary
  - GSS with varying % of Ly-α line and f_{esc}

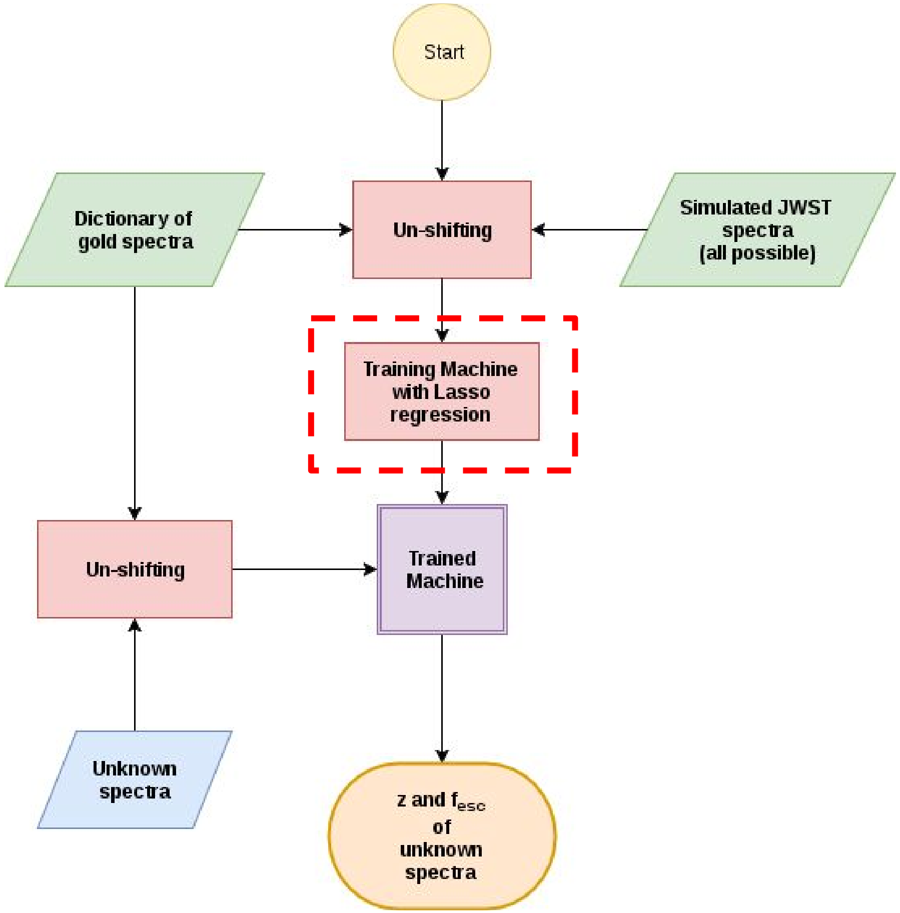


Un-shifting

- The spectra are shifted onto all the GSS using a similarity metric
- Similarity metric: **Cross-relation**, **L2-norm**
- Matching the wavelength resolution
  - **Nearest interpolation** is used to match the wavelength resolution of spectra to GSS



Flowchart


Start
Dictionary of gold spectra
Un-shifting
Simulated JWST spectra (all possible)
Training Machine with Lasso regression
Un-shifting
Trained Machine
Unknown spectra
z and f_{esc} of unknown spectra

Lasso Regression

- Data points: (x_1,y_1), (x_2,y_2),........, (x_m,y_m)
- Machine is trained with a linear model

$$\hat{y} = \beta_0 + \sum_{i=1}^{N} \beta_i x_i$$

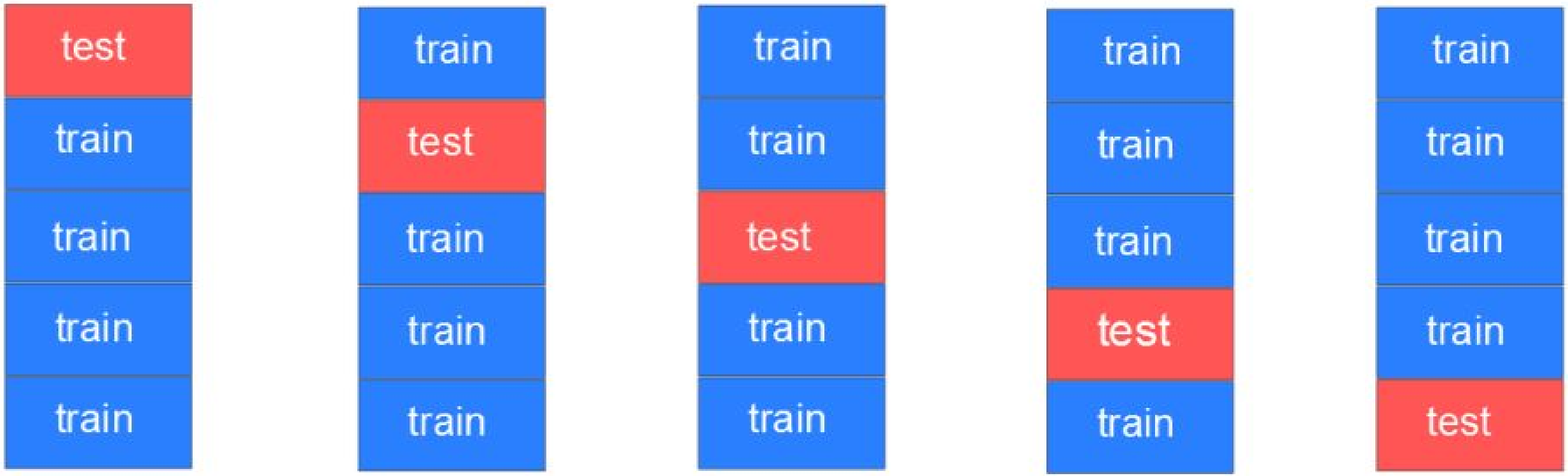
- β_i are the model coefficients to be determined
- *N* is the number of features in x

- Likelihood function

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{m} [\hat{y}(\boldsymbol{x}_i) - y_i]^2 + \lambda \sum_{j=1}^{N} |\beta_j| \right\}$$

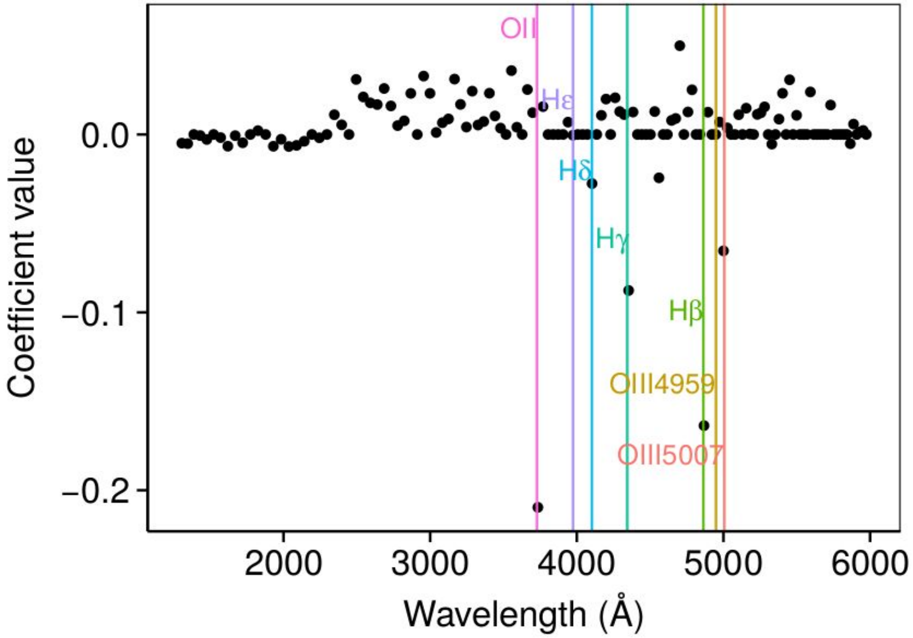
- λ is the penalty term
- It is determined using “cross validation” leaving *n* samples out

k-fold Cross-Validation



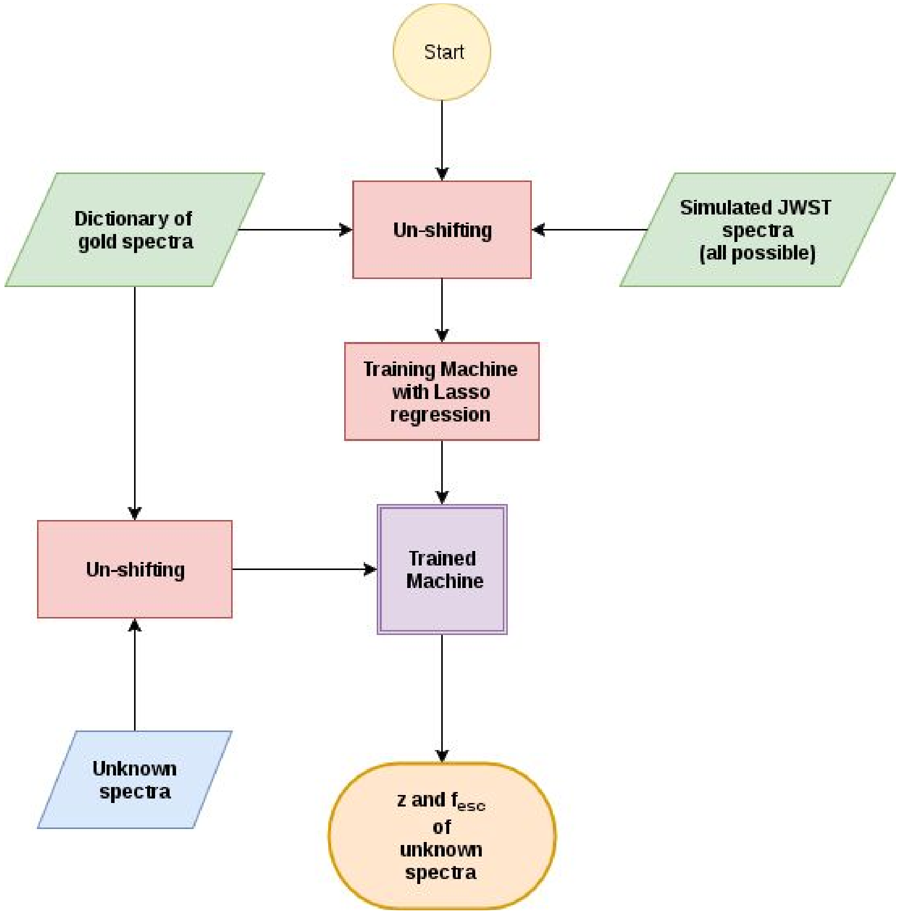
Lasso Regression

- Regularization and Penalty
  - Controls overfitting of data
  - Useful in ill-posed data
  - It suppresses the dependence on unuseful features
- Right image from [Jensen et al (2016)]
  - The features with high coefficients are prominent nebular lines



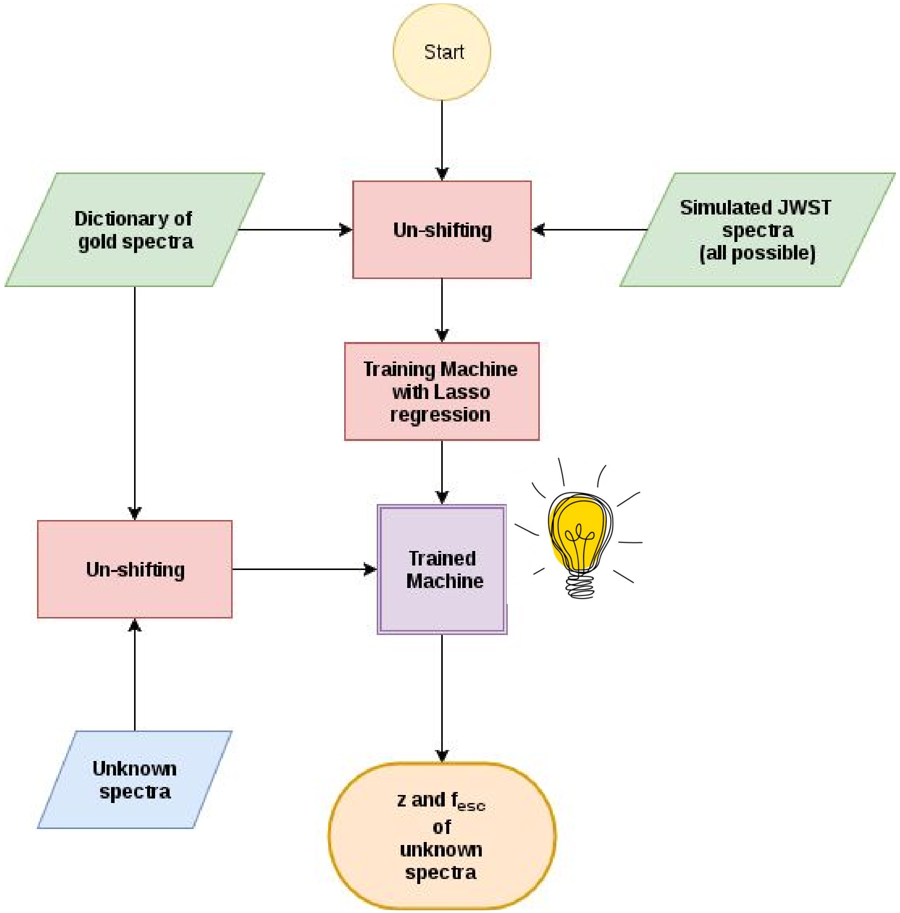
Analysis

- Input feature x: (fluxes of unshifted spectra, shift)
- Output parameter y: (redshift z, escape fraction f_{esc})
- Cross-validation: $k = 10$
- The exposure time for the simulated JWST spectra is 10 h



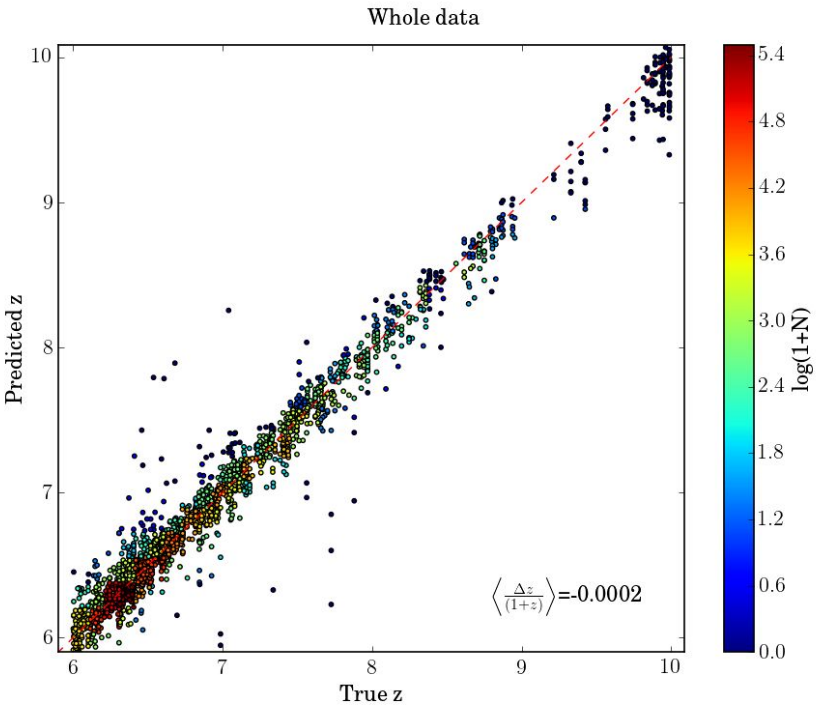
Analysis

- Input feature x: (fluxes of unshifted spectra, shift)
- Output parameter y: (redshift z, escape fraction f_{esc})
- Cross-validation: k = 10
- The exposure time for the simulated JWST data is 10 h



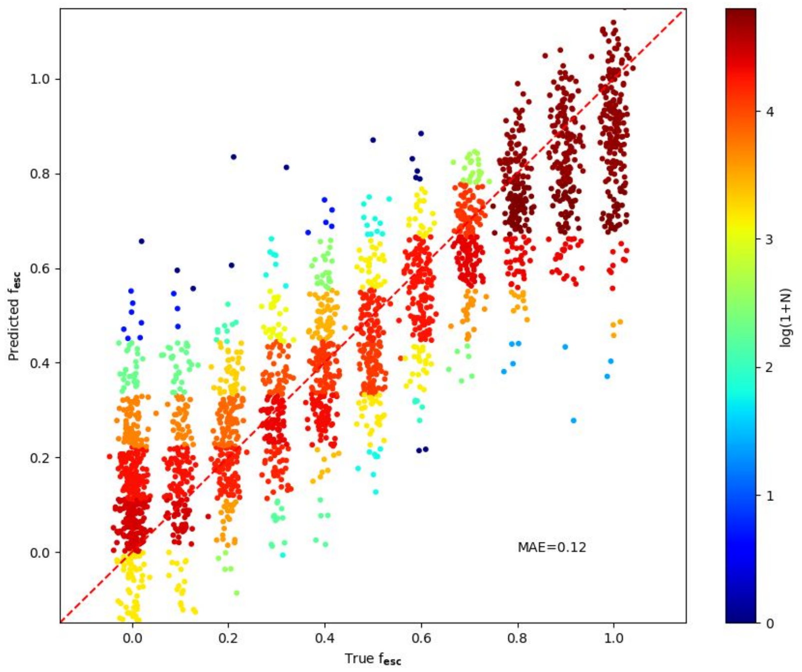
Results

- Scatter plot of true and predicted z
- The accuracy in estimating the “shift” affects the prediction
- The spread in the prediction is more at high z
  - Due to noisier spectra from high z



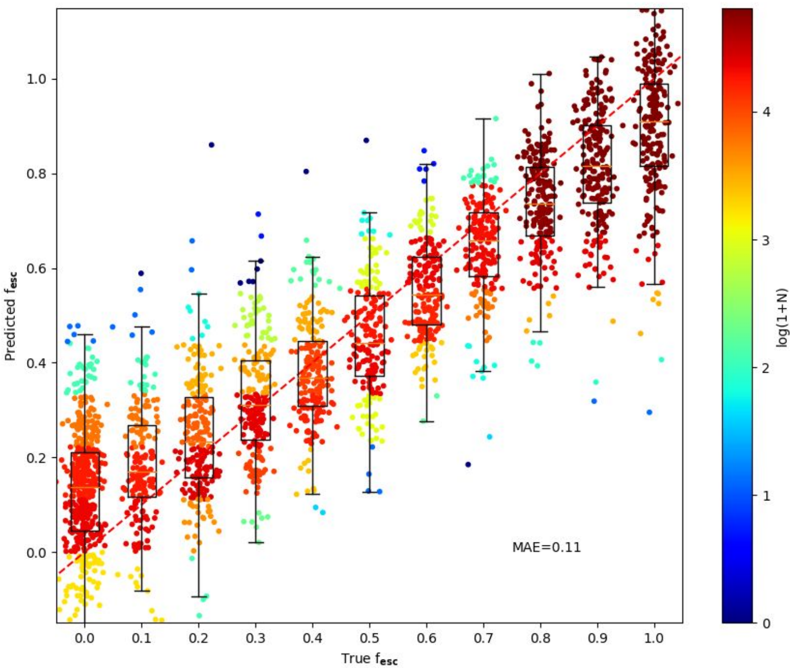
Results

- Scatter plot of true and predicted f_{esc}
- The mean absolute error (MAE) is similar to what is shown by [Jensen et al (2016)](#) for single z



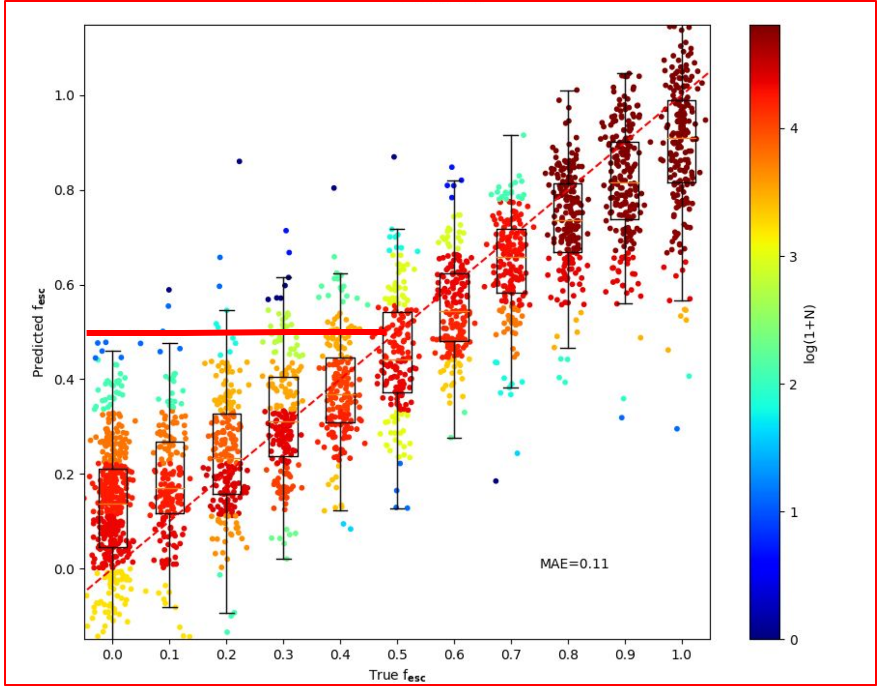
Results

- Scatter plot of true and predicted f_{esc}
- The mean absolute error (MAE) is similar to what is shown by [Jensen et al (2016)] for single z
- **Identify high leakers**



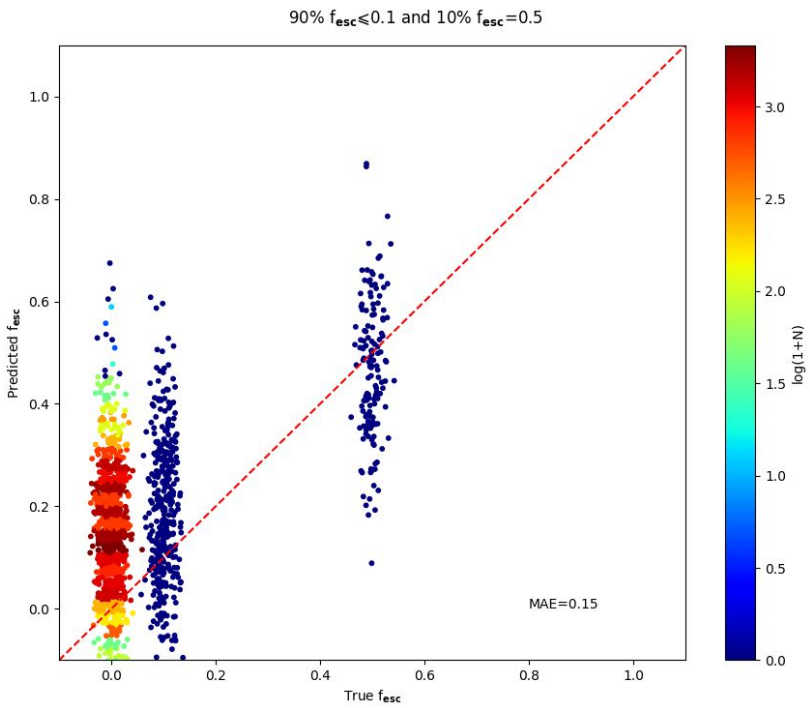
Results

- Scatter plot of true and predicted f_{esc}
- The mean absolute error (MAE) is similar to what is shown by Jensen et al (2016) for single z
- **Identify high leakers**



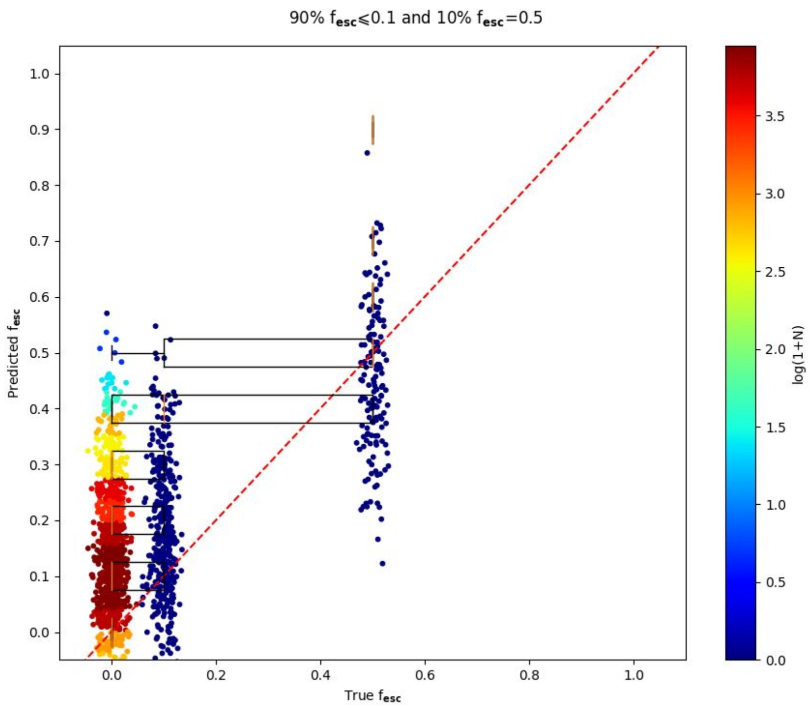
Results

- The real dataset will be skewed
- With very few high leakers



Results

- The real dataset will be skewed
- With very few high leakers
- The identified high leaker will have a certain probability of being a high leaker



Summary

- The procedure will be useful to analyse the large JWST dataset
- It estimates the redshift and the escape fraction of the high redshift galaxies
- It is useful in detecting the high Lyman continuum leaking galaxies