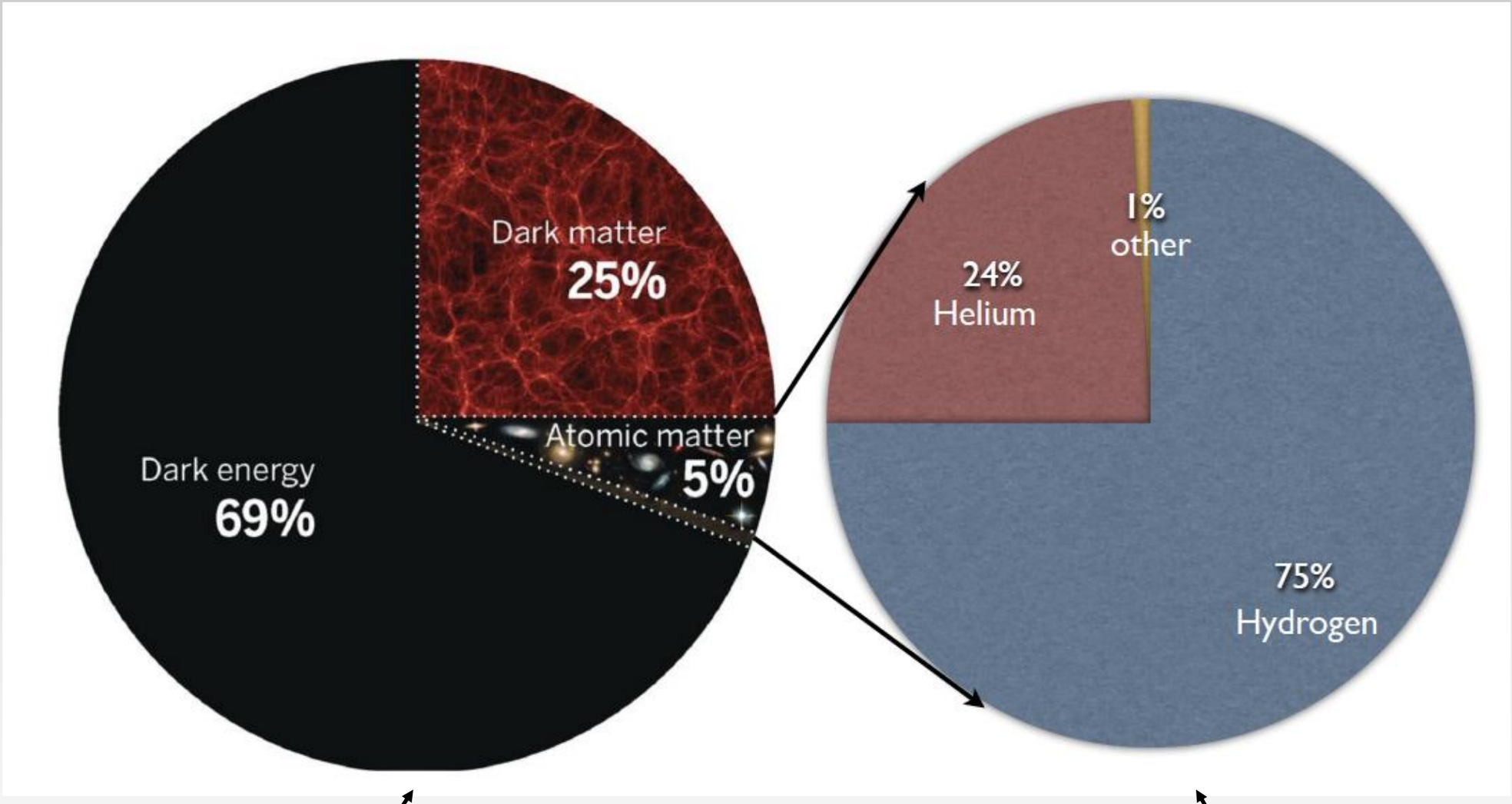
Data Science in Astronomy : Predicting the sources of cosmic reionization

Moupiya Maji, Anne Verhamme, (Observatoire de Geneve)
Maria-Pia Victoria-Feser, Marta Pittavino (Research Center for Statistics)
& the SPHINX collaboration

What is our Universe made of?



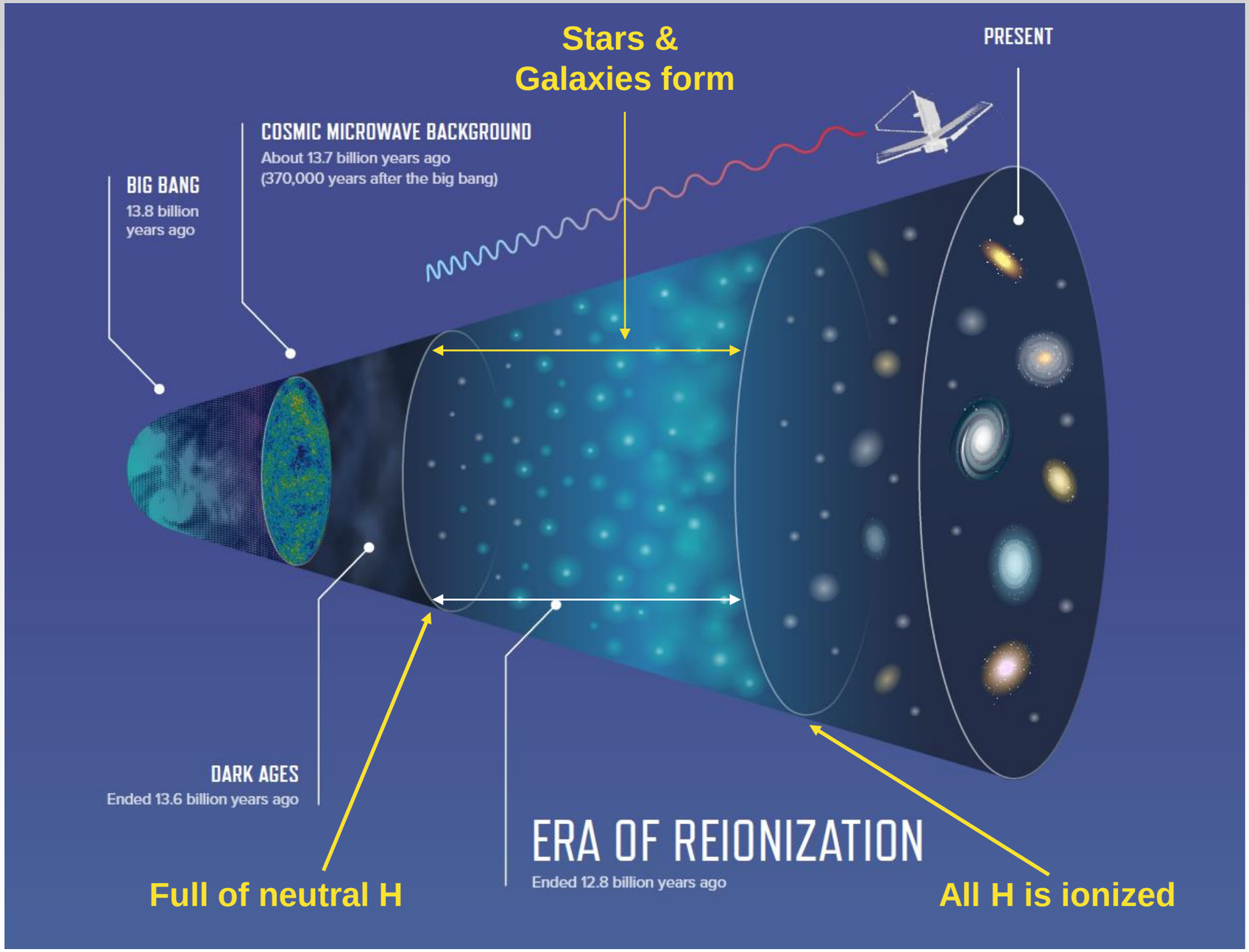
The make-up of
our Universe

75% of all
normal matter is H

Only 5% of the content is
Normal matter

Hydrogen is the most abundant
element in the universe

First galaxies and Reionization



- When our Universe was less than a billion years old (current age 13.8 billion years) galaxies were still just forming and the Universe was full of neutral hydrogen gas.

- Around this time the hot stars in these first galaxies emitted enormous amount of energetic photons and these photons ionised the ubiquitous neutral hydrogen.

- This milestone event is known as reionization of the Universe and is an important part of our cosmic history.

Image: NASA, ESA, Joyce Kang (STScI)

First galaxies and Reionization

- However, it is not possible to directly observe these ionizing photons (also called LyC or Lyman continuum) as they are all destroyed or absorbed on their long journey to us (they would have had to travel for ~13 billion years).

- So, in our project we look for **other signatures** of these ionizing sources to understand how this reionization event unfolded.

First stars , artists conception

How about other Hydrogen emission lines? The brightest line from Hydrogen is Lyman alpha line.

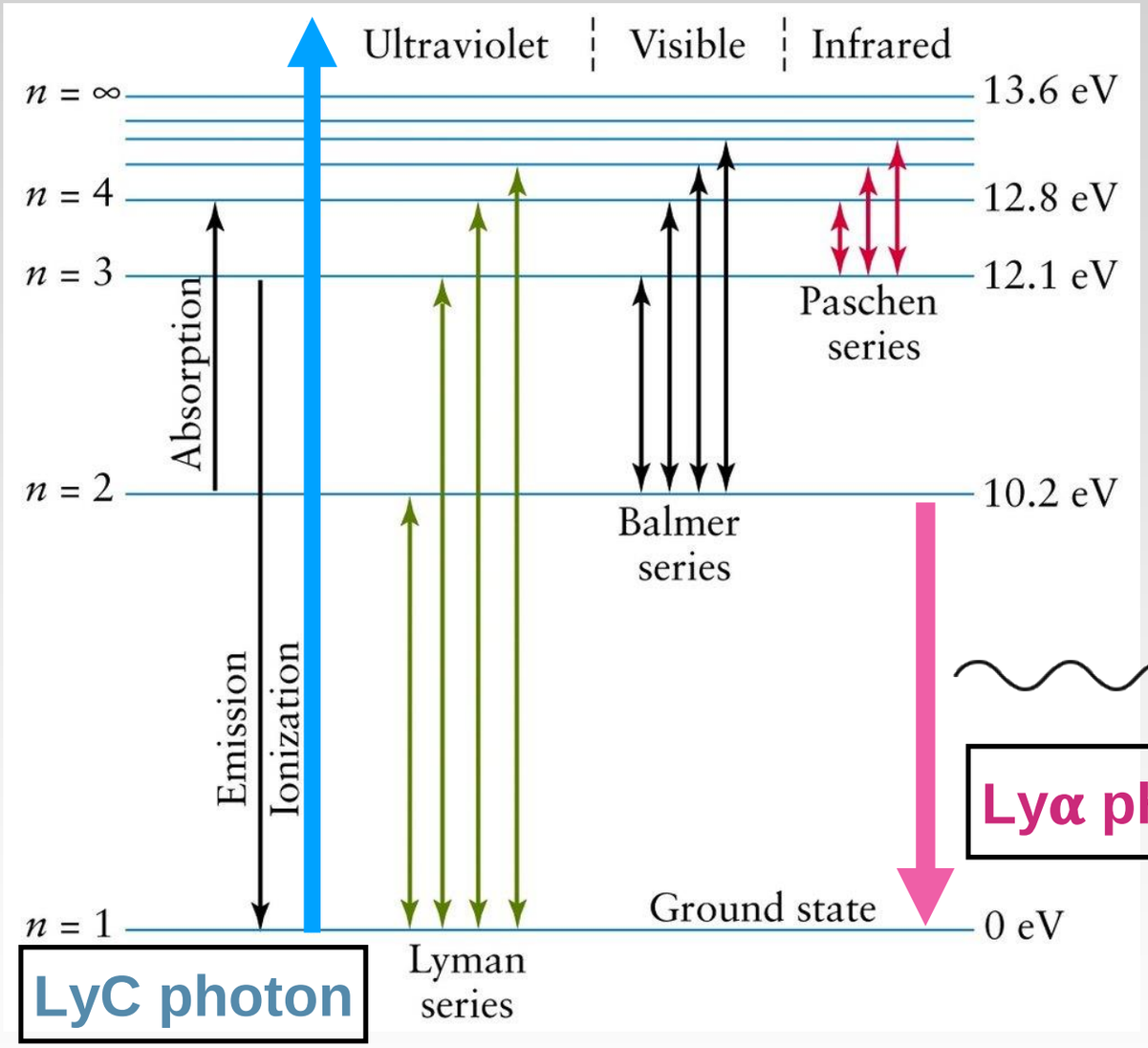
Signature of H ionizing(LyC) photons : Lyman alpha photons

Lyα photon == 2 ⟶ 1 transition

Wavelength = **1216 Å**
Energy = 10.6 eV

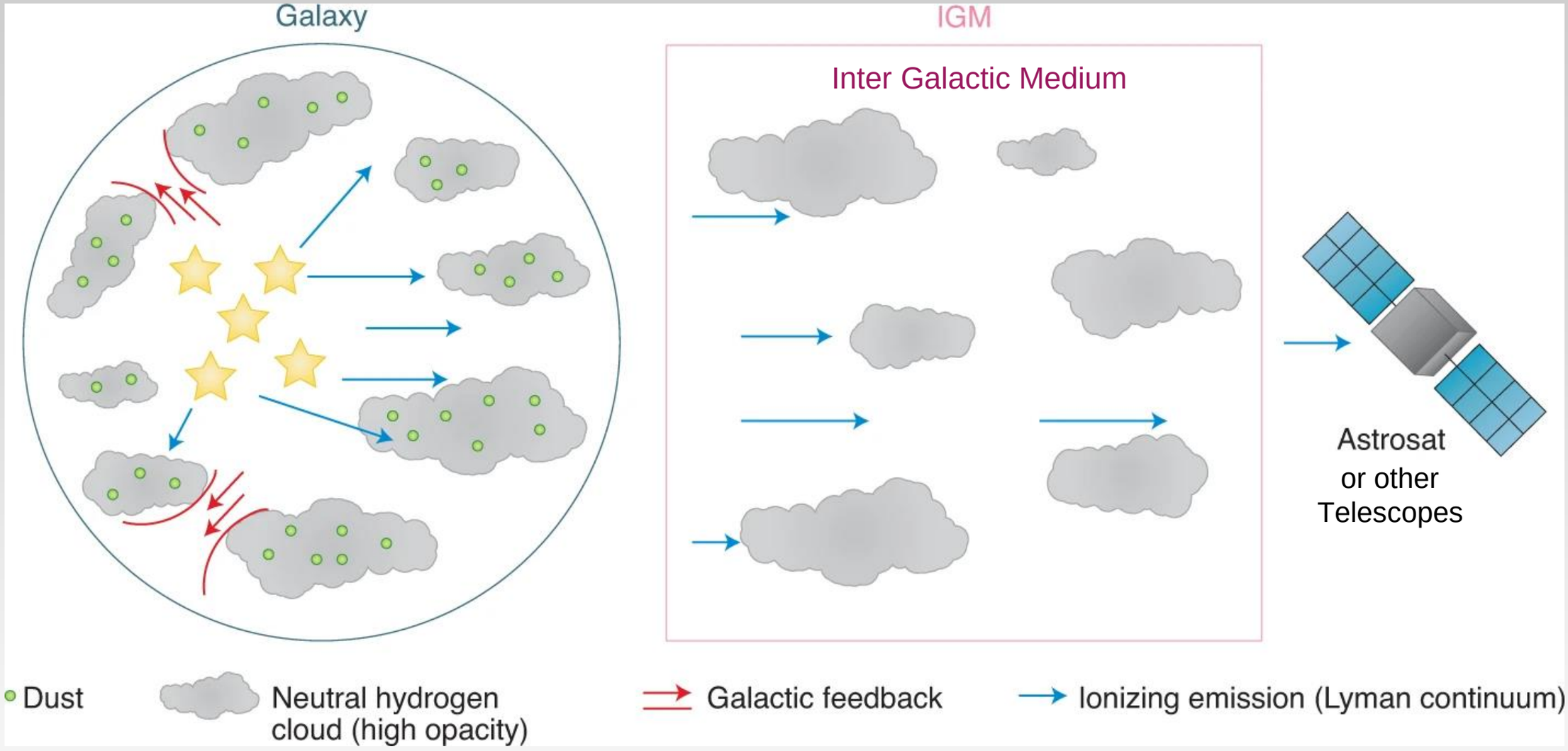
Lyman continuum (LyC) or **ionizing** photon
== 1 ⟶ ∞ transition

Wavelength = **912 Å**
Energy = 13.6 eV



Energy levels of a H atom

First galaxies and Reionization



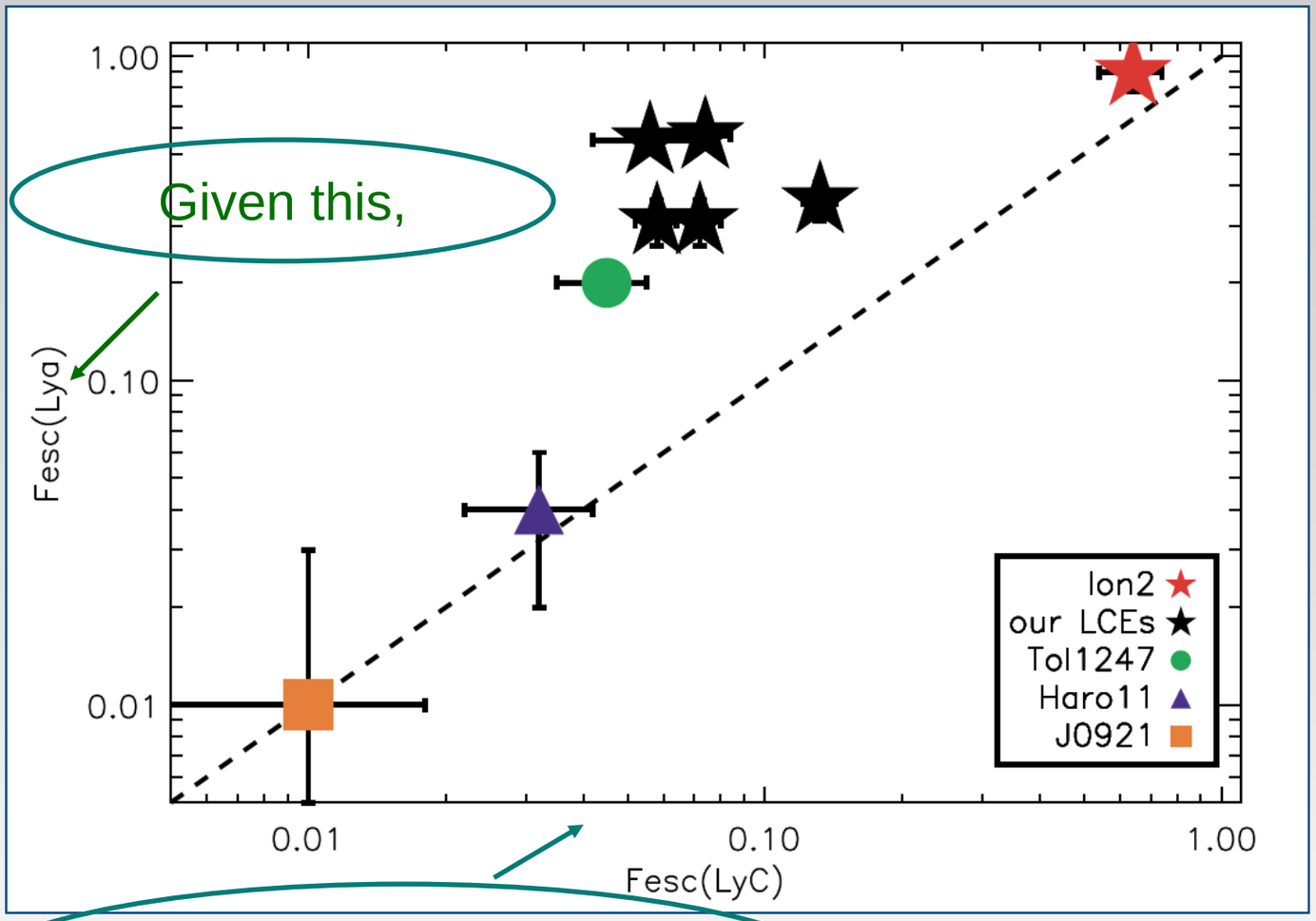
Alavi, A. Escaping photons finally arrested. *Nat Astron* **4,** 1130–1131 (2020).

Saha, K. *et al. Nat Astron* **4,** 1185–1194 (2020).

Stars in galaxies produce LyC (ionizing) photons. The amount of LyC being produced is called

A lot of these photons are absorbed by dust particles or hydrogen in the galaxy. Only some pho

Lya and LyC from galaxies are correlated



$$\text{Escape fraction} = \frac{\text{Escaping Luminosity}}{\text{Intrinsic Luminosity}}$$

Verhamme et al. A&A, 2017

Observations from local Universe show that the escape fractions of LyC and Lya are correlated. This invites the question,

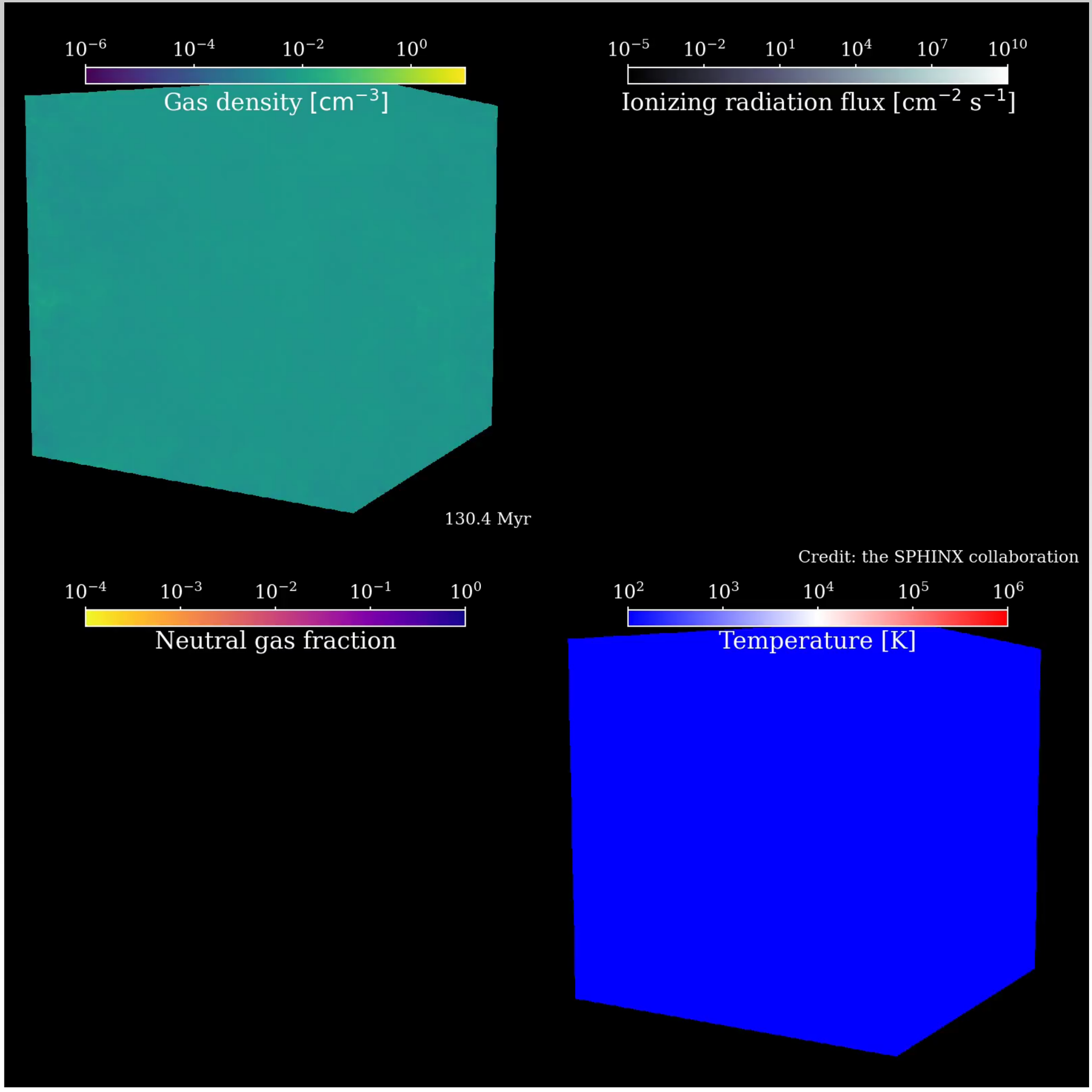
Can we use Lyman alpha emission to estimate the LyC emission from the first galaxies?

Since we can't observe LyC from reionization era, we need to use simulation data to answer this question.

SPHINX
Simulation

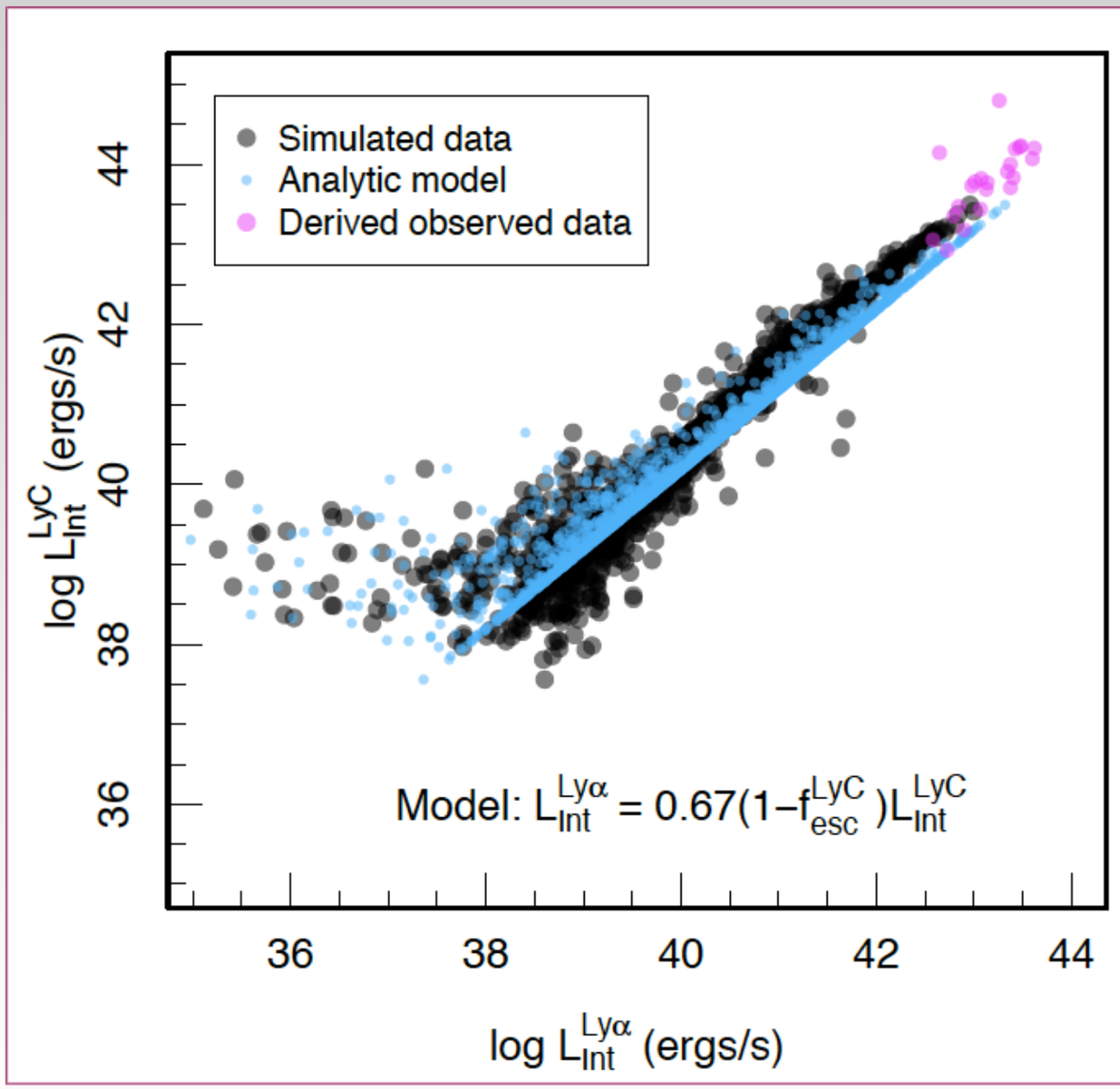
- We use a high resolution cosmological simulation SPHINX simulation for this study.

- We finally have a **sample of 1933 galaxies.** For each of them we have LyC (ionizing radiation), Lyman alpha radiation and other galaxy properties (e.g. their mass, sizes etc.)

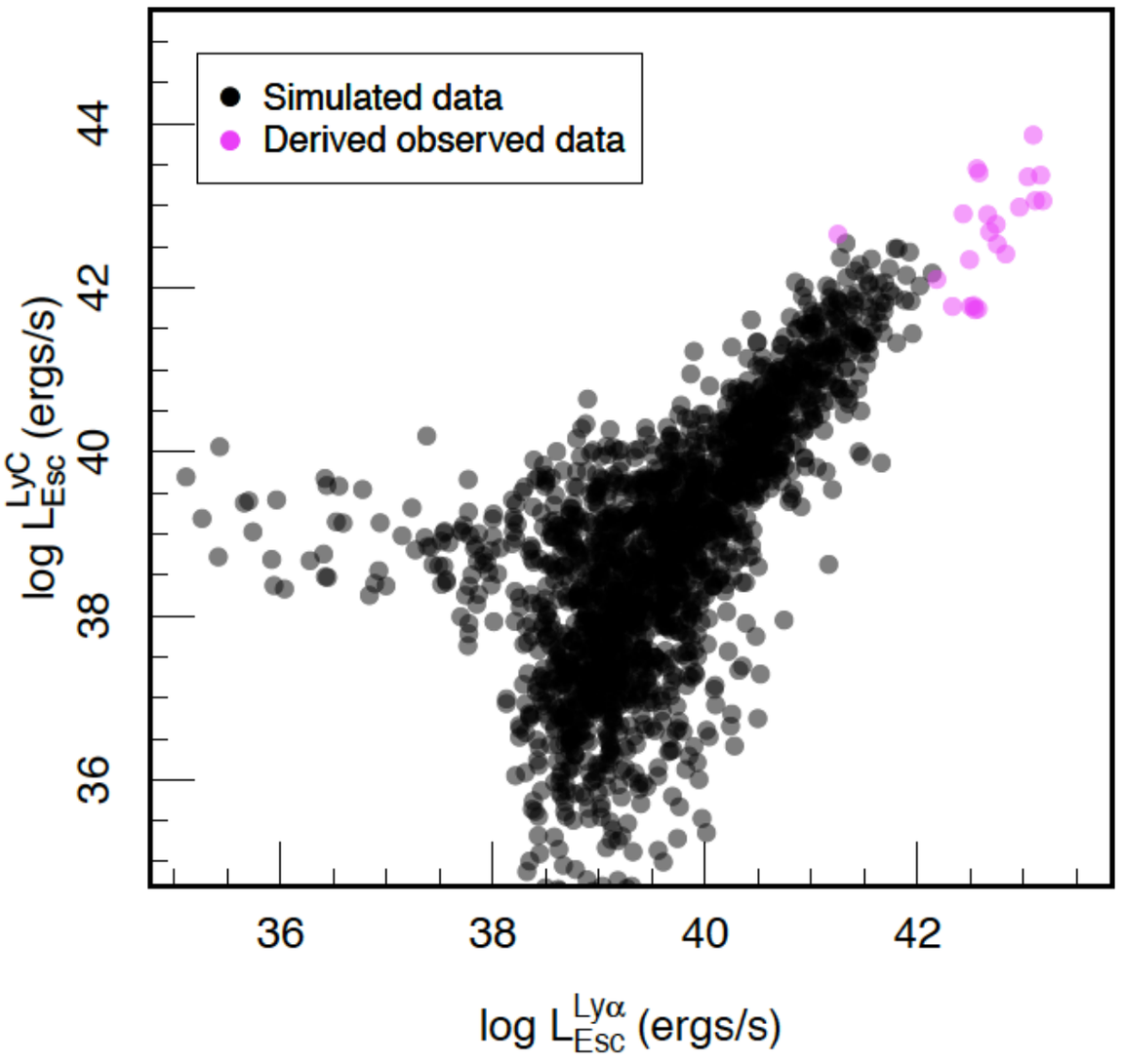


Credit : Sphinx collaboration

Lya and LyC in simulations : Luminosity

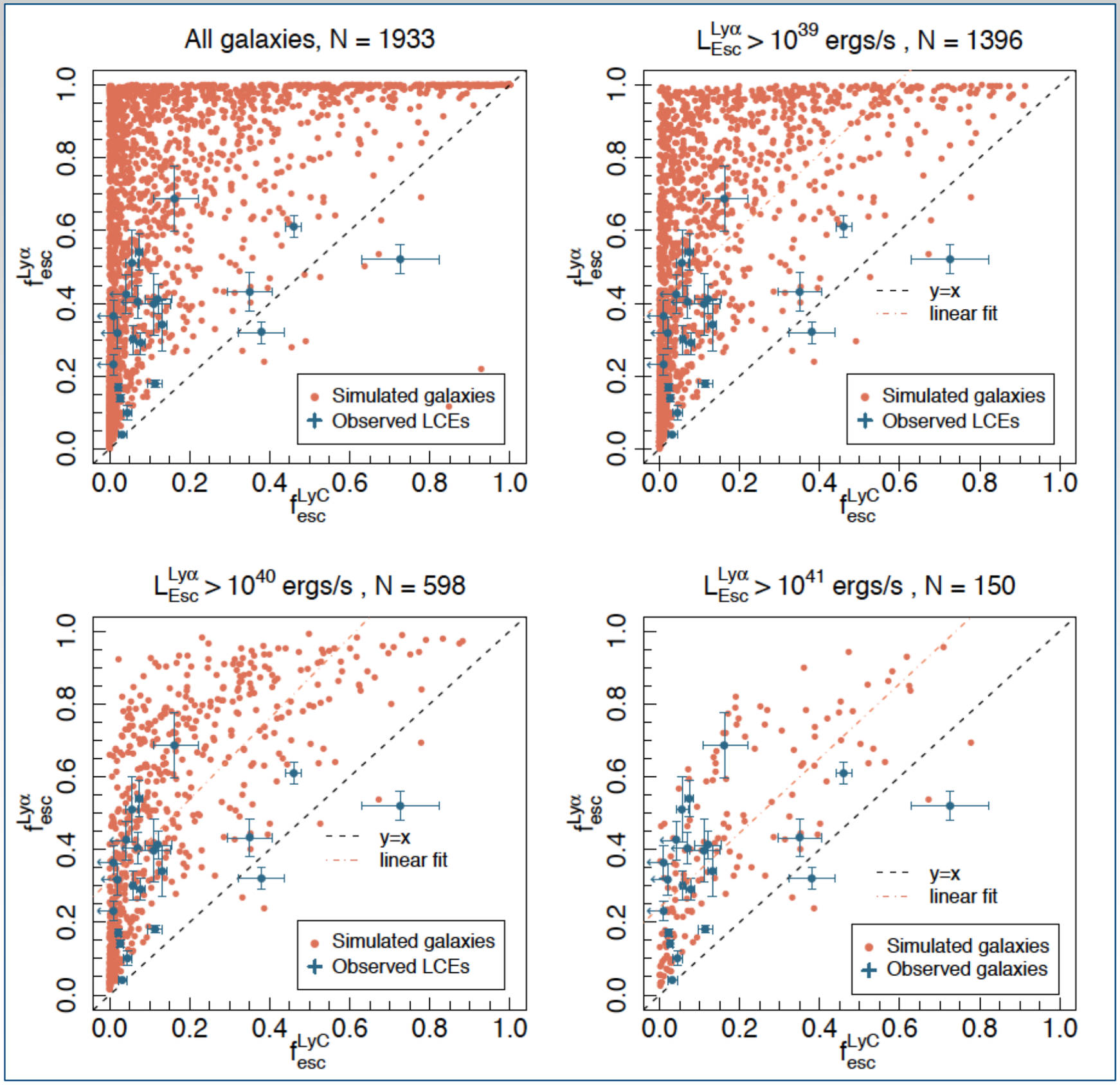


Intrinsic luminosity



Escaping luminosity

Lya and LyC in simulations : escape fraction



Prediction using Multiple Linear Regression

- We didn't find any simple 1:1 relation between Lya or LyC variables and individual galaxy properties. However, for each galaxy we have many more properties e.g. mass, size, star formation rate, stellar age etc.

- So here we employ a *multiple linear regression* method.

- Each variable is standardised using : $x_s = \log(x) - \text{median}(\log(x))$

- The model is then fit by this formula where we predict y using known x's:

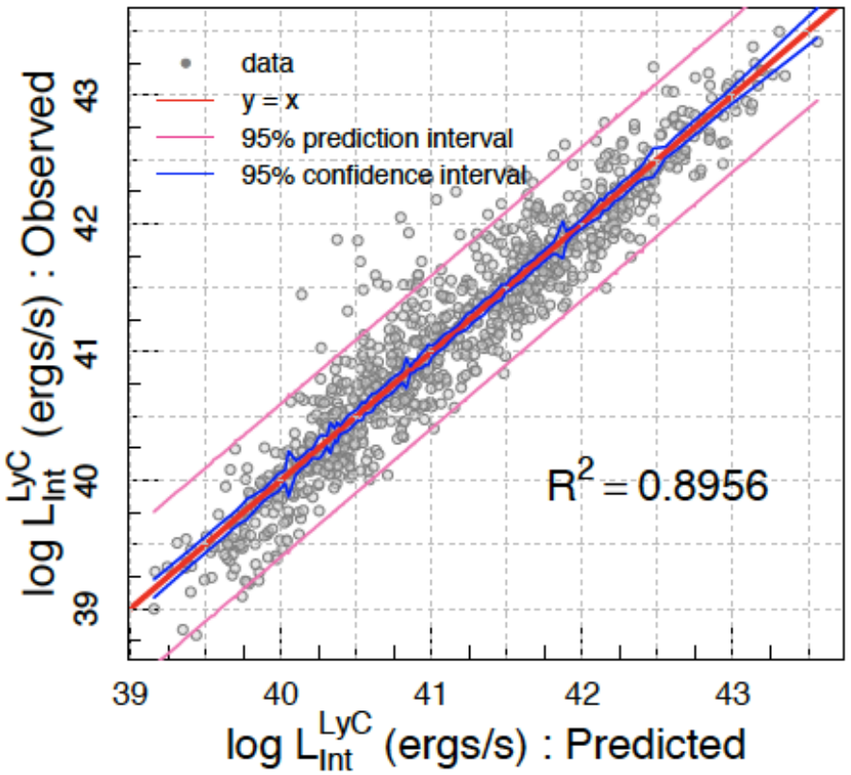
$$y = \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n + \beta$$

- Finally, we determine which x variable holds the most predictive power in determining y by *forward-selection* and *backward selection* method.

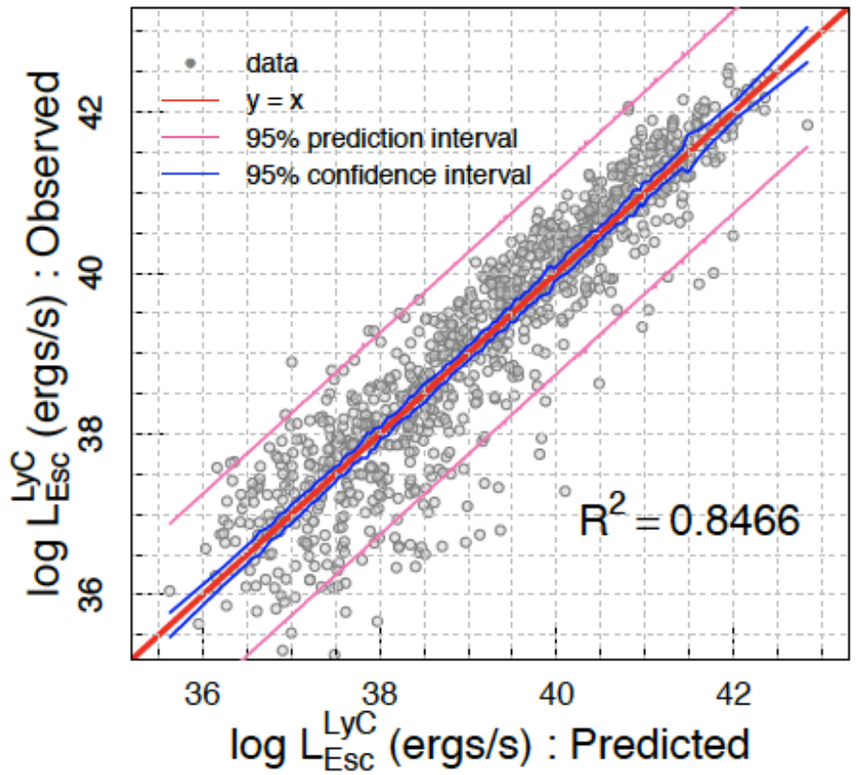
- We want to predict LyC intrinsic luminosity, escaping luminosity and escape fraction.

Predicting LyC

Intrinsic luminosity →

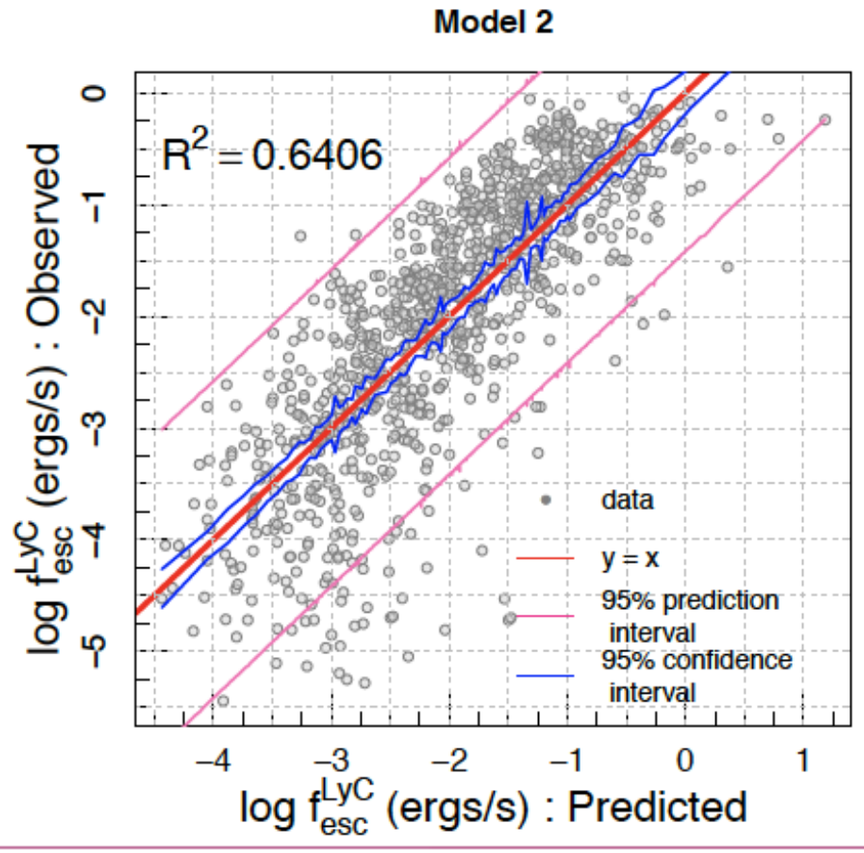


← Escaping luminosity



Performance of our model can be defined by
R^2 ---> A measure of how much of the **data variance** can be explained by our model.

If $R^2 = 1$, model explains everything, if it's 0, it has no explanatory power.



Escape Fraction

Most important x-variables for prediction

- In the model we have supplied 8 galaxy properties for predicting LyC. However, observing and determining many galaxy properties at high redshift can be extremely challenging. So it is necessary to identify which of the x- variables is the most important in predicting y.

- We use forward - selection and backward-selection methods for finding out which x-variables are most important for our prediction.

- Here we find that the first three variables alone explain most of the variance. So knowing the first three properties is practically enough for prediction.

Intrinsic luminosity

$L_{\rm int}^{\rm LyC}$		
Rank	Variable	Adjusted R^2
1	$\rm SFR_{10}$	0.7764
2	$L_{\rm esc}^{\rm Ly\alpha}$	0.8856
3	$\rm SFR_{100}$	0.8911
4	M_{Gas}	0.8919
5	Stellar Age	0.8949
6	$M_\star$	0.8955
7	R_{Gal}	0.8956
8	Metals	0.8956

Escaping luminosity

$L_{\rm esc}^{\rm LyC}$		
Rank	Variable	Adjusted R^2
1	$L_{\rm esc}^{\rm Ly\alpha}$	0.7877
2	M_{Gas}	0.8243
3	$\rm SFR_{10}$	0.8355
4	$M_\star$	0.8436
5	$\rm SFR_{100}$	0.8460
6	Stellar Age	0.8463
7	Metals	0.8467
8	R_{Gal}	0.8466

Escape Fraction

$f_{\rm esc}^{\rm LyC}$		
Rank	Variable	Adjusted R^2
1	$L_{\rm esc}^{\rm Ly\alpha}$	0.2983
2	$\rm SFR_{10}$	0.5397
3	M_{Gas}	0.6269
4	$M_\star$	0.6392
5	Metals	0.6411
6	R_{Gal}	0.6411
7	$\rm SFR_{100}$	0.6409
8	Stellar Age	0.6406

Our collaboration

- In the **CCSD kick-off meeting, early September 2020**, where Marta and Maria-Pia presented statistical methodologies that were linked with the statistical models we were using for your research project and Anne made the first contact with Marta and Maria-Pia.

- Then, we decided to present our work to one of the **data science clinic**s organised by the CCSD, to get feedback on the choice of the statistical method. We presented in the data clinic of **December 2020**, asked our key questions on the model used and a more active collaboration started.

- Some of the questions that we had were:
  - *How can we validate or establish if the linear regression is a good choice for our model?*
  - *Can we use it to predict a quantity which would not follow a normal distribution (fesc) ?*
  - *Can the predictors be non-gaussian ?*
  - *How can we deal with interlopers/outliers ?*
  - *How can we deal with extreme values (half of the sample has SFR=0)?*

- They gave us some feedback during the meeting, and we also set up longer meeting sessions separately. We sent them the draft afterwards, and then again we met together in **Summer 2021** to discuss and implement their comments.

- **Their inputs were extremely useful to improve our statistical model. Marta and Maria-Pia were added as co-authors of the work and warrants of the validity of our modelling.**

Summary

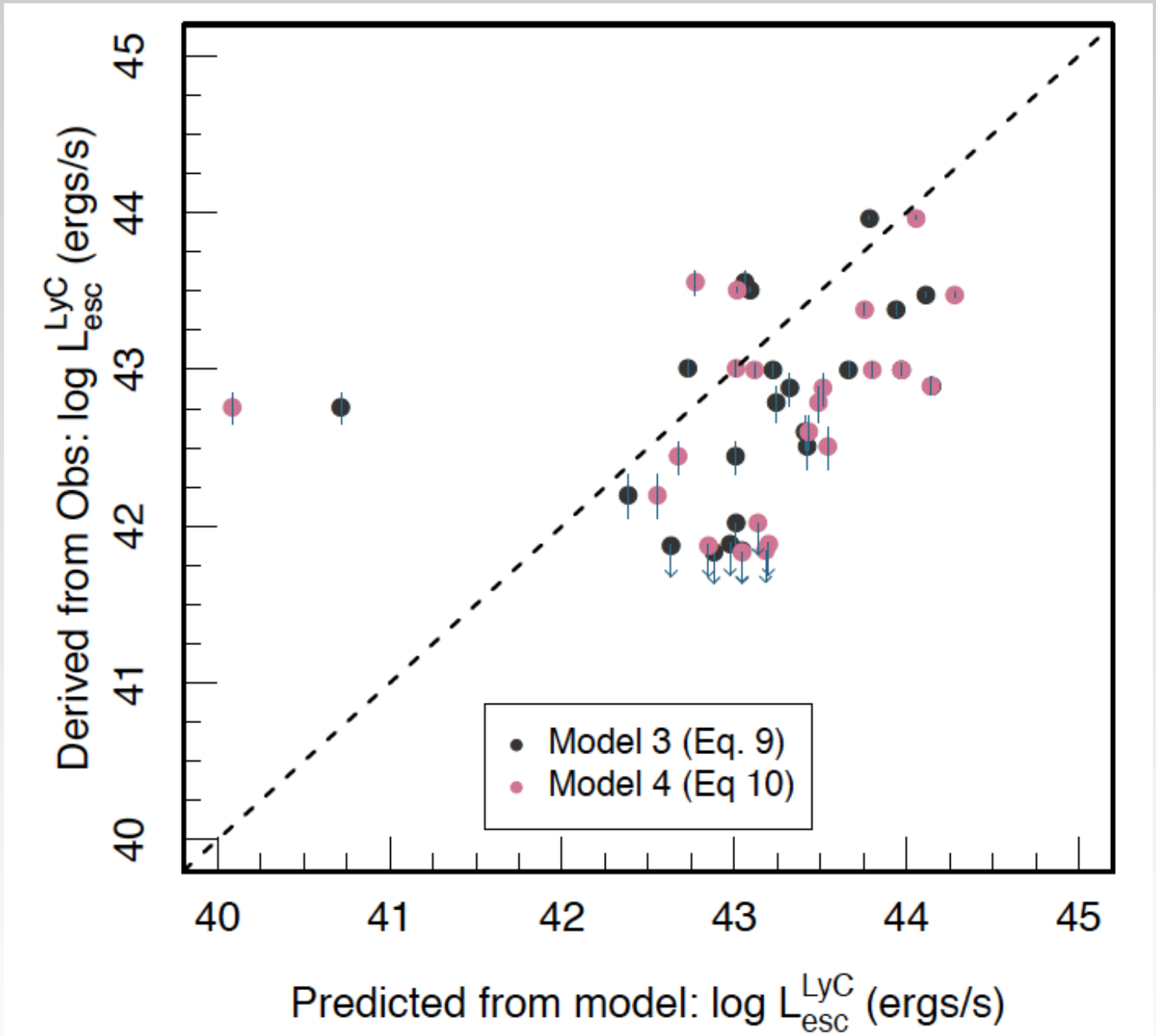
- It is impossible to observe the ionizing radiation of reionization era directly as these photons get absorbed on their way to us. So we need indirect tracer for LyC radiation. Would Lyman alpha help?

- We analyse 1933 galaxies in the SPHINX simulation and calculate Lya emission from both recombination and collision. We find that LyC and Lya luminosities are somewhat correlated but has high dispersion and the escape fractions are not correlated.

- In absence of any clear 1:1 correlations, we employ a multivariate linear regression modelling and find that it is possible to predict LyC radiation of galaxies given their physical properties and Lya emission.

- Our model can be very helpful in identifying which galaxies contributed most towards reionization and can be applied to plan future astronomical surveys.

Applying our model to observations



Escaping luminosity

r model to 23 observed galaxies and find that our predictions match with the observed luminositie