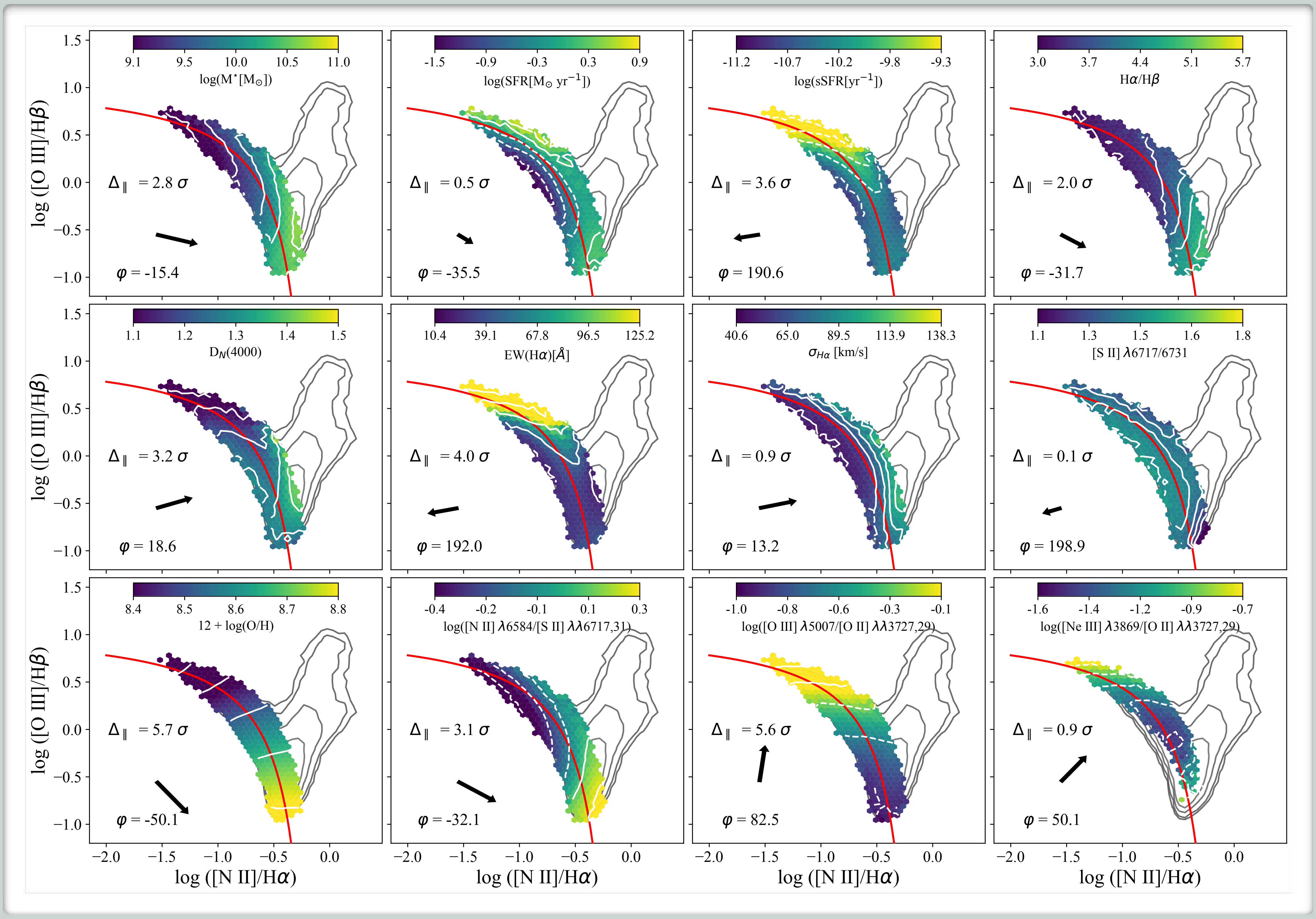
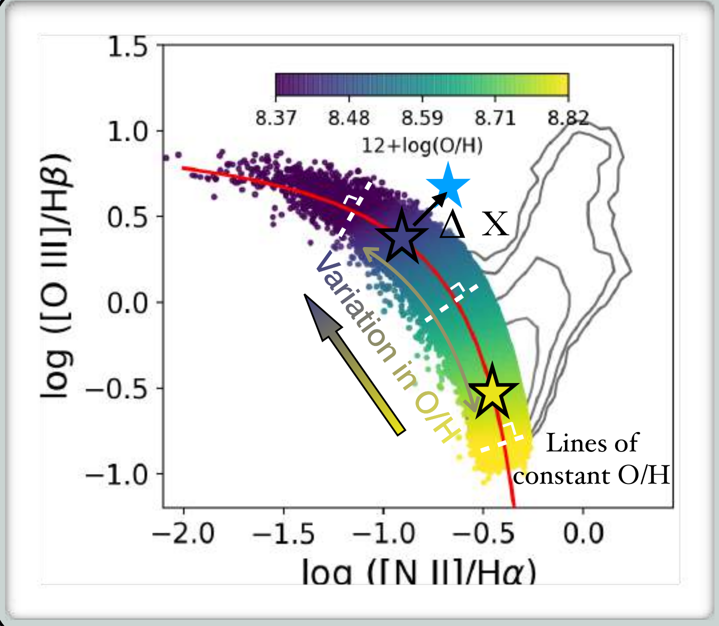
What drives the scatter in the BPT diagrams ? A Machine Learning based analysis

Mirko Curti - *Kavli Institute for Cosmology, University of Cambridge*

The location of star-forming galaxies in the BPT diagrams is strongly sensitive to many different physical properties. Here, we implement machine learning algorithms to assess which parameters are the most relevant in predicting the observed offset from the local star-forming sequence in both the [N II]- and [S II]-BPT diagrams, once metallicity (which primarily determines the position of galaxies along the sequence) is fixed. We find deviations in the N/O abundance as the most predictive parameter for both classifying galaxies above/below the SF-locus and for reproducing the offset from the best-fit line in the [N II]-BPT, whereas deviations in SFR are the primary drivers of the scatter in the [S II]-BPT. The analysis is presented in detail in Curti et al., 2021 (in prep).



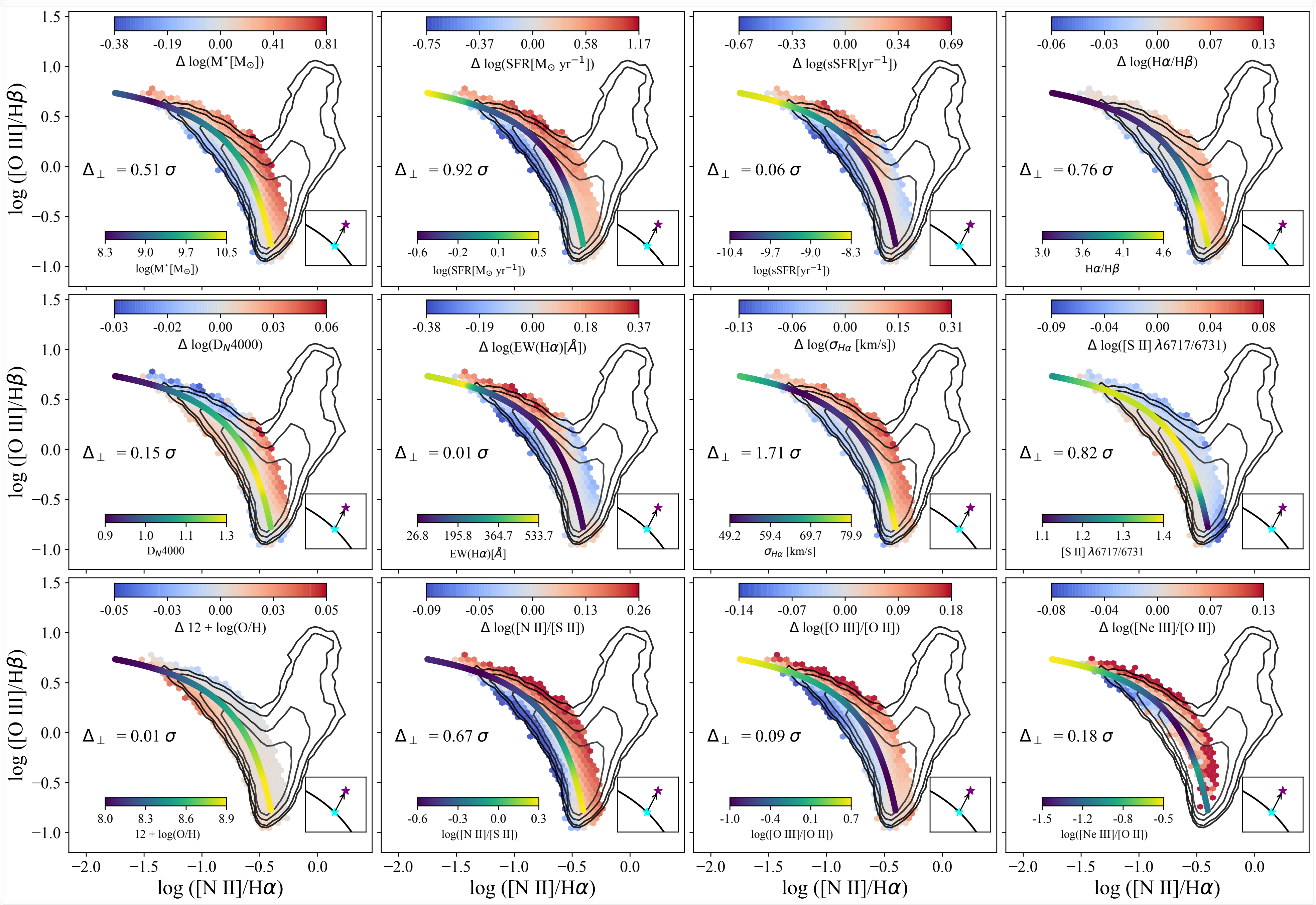
Distribution of star-forming galaxies from the SDSS in the [N II]-BPT diagram. In each panel, the colour-coding reflects the average value computed in small hexagon bins for observational parameters tracing different physical properties. White contours indicate lines of constant values in each parameter, whereas the best-fit to the SF-locus is indicated by the red curve.



Star-forming galaxies form a smooth metallicity sequence on the [N II]-BPT, and lines of constant O/H are orthogonal to the best-fit line along the sequence. We assume that the offset from the sequence **i)** occurs orthogonal to the SF-locus **ii)** occurs at fixed metallicity **iii)** is driven by the relative variation in different physical parameters with respect to the average of galaxies lying on the SF-sequence

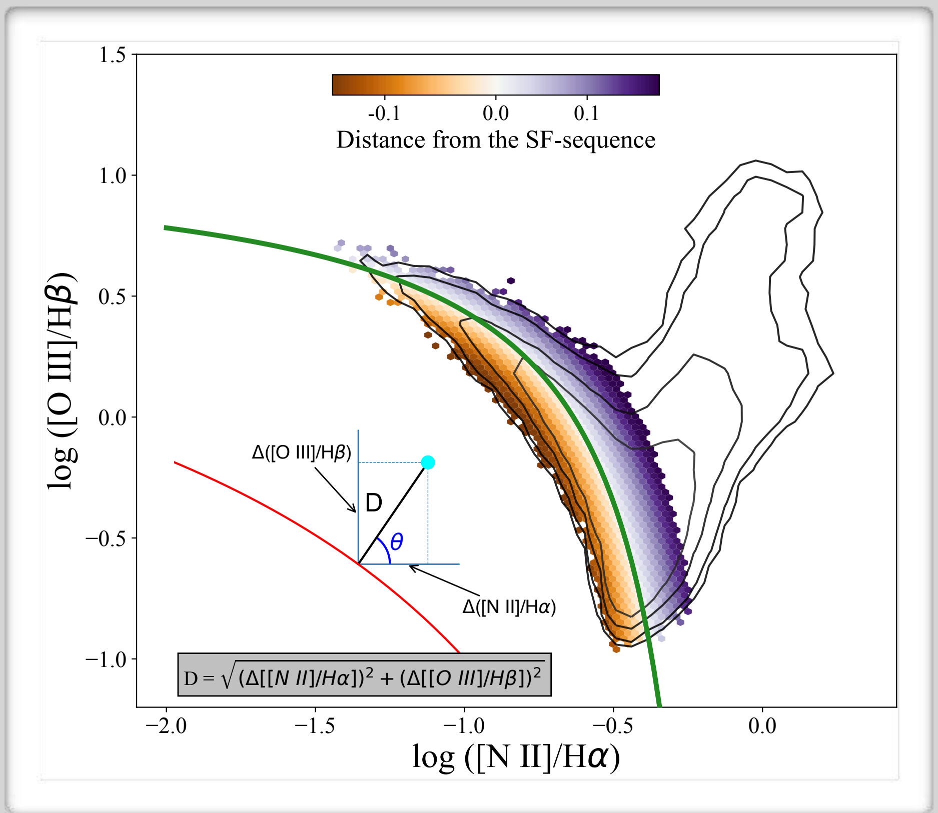
$$\Delta \log X = \log X - \langle \log X(\text{SF-sequence}) \rangle$$

$$X \in [M^\star, \text{SFR}, \text{EW}(H\alpha), D_N4000, [\text{N II}]/[\text{S II}], [\text{O III}]/[\text{O II}], [\text{SII}]6717/31, \sigma(H\alpha), H\alpha/H\beta]$$



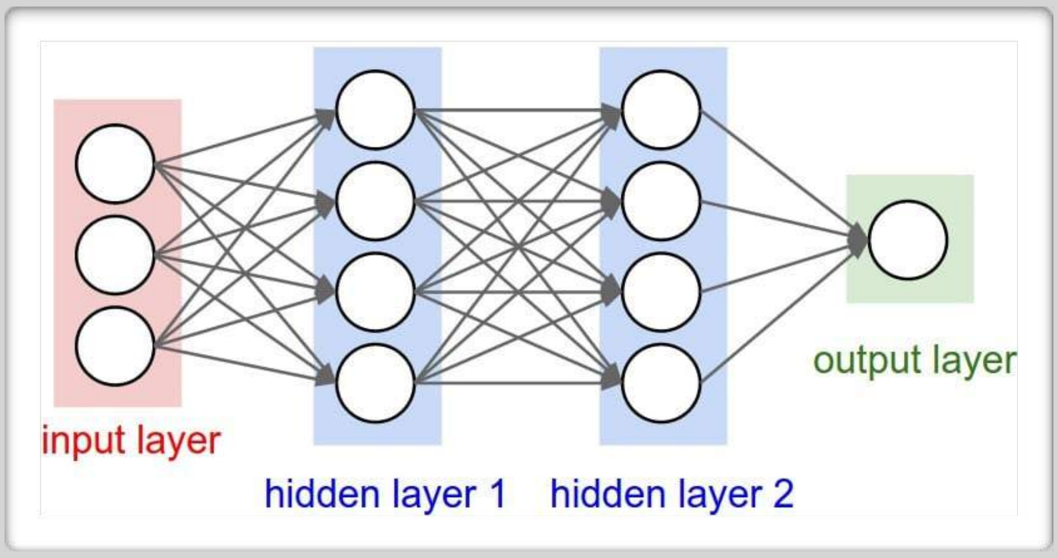
The distribution of galaxies in the [N II]-BPT diagram is colour-coded by the mean logarithmic deviation in each parameter with respect to the average of galaxies along the SF-sequence, assuming a purely orthogonal offset vector. These quantities represent the input features for our machine learning analysis, aimed at both classifying galaxies as lying above/below the best-fit line and predicting the exact amount of offset (i.e., the amplitude of the orthogonal offset-vector).

Hidden layers : [24;12] Neurons - **Activation Function** : ReLu - **Loss Function** : Binary Crossentropy (classification); MSE (regression) - **Optimizer** : Adam (0,001 learn rate) **Metrics** : [Accuracy/AUC] (classification); [MSE/Improvement over Random] (regression) - Trained over **200 epochs**, for each individual parameter and for the full set altogether

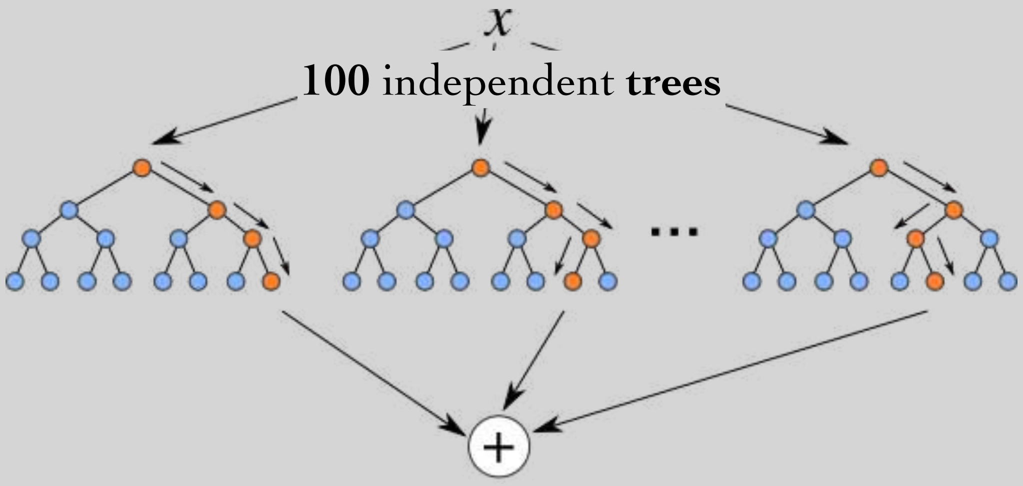


Galaxies in the [N II]-BPT are colour coded by their distance **D** from the SF-sequence, the target label for the ML regression

Artificial Neural Network



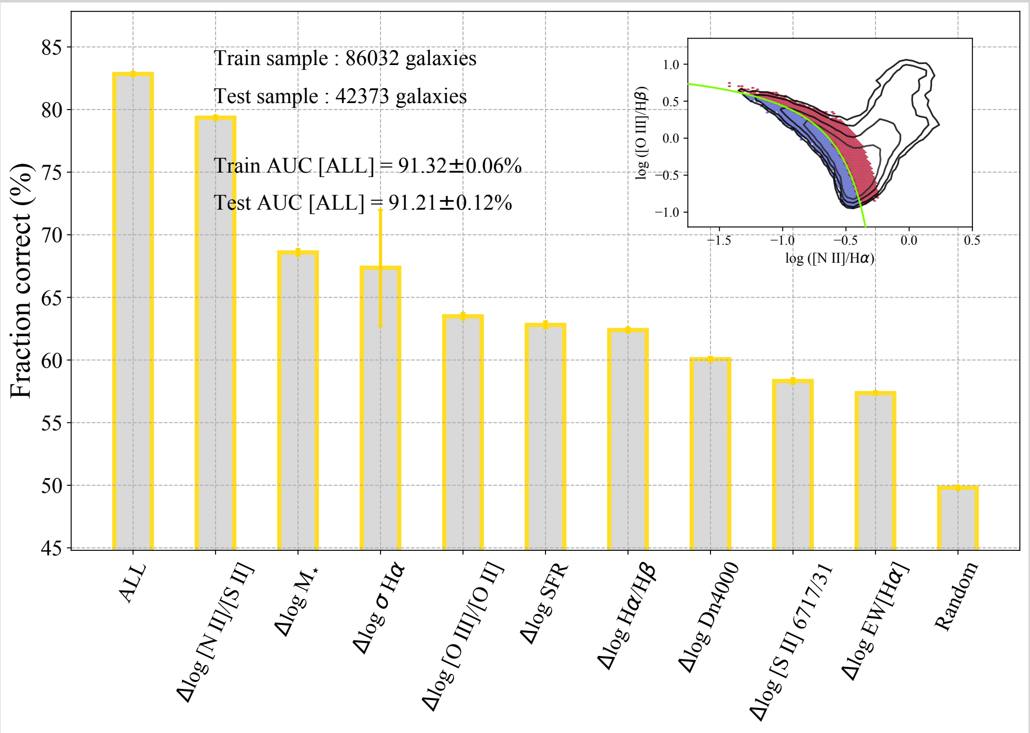
Random Forest



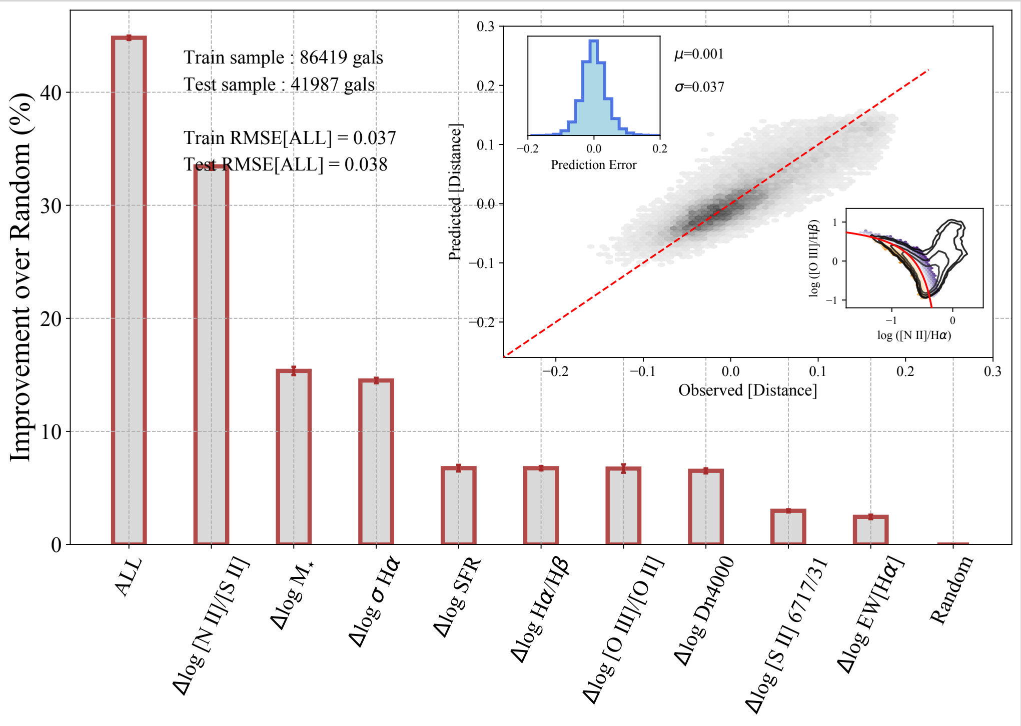
100 independent **trees**

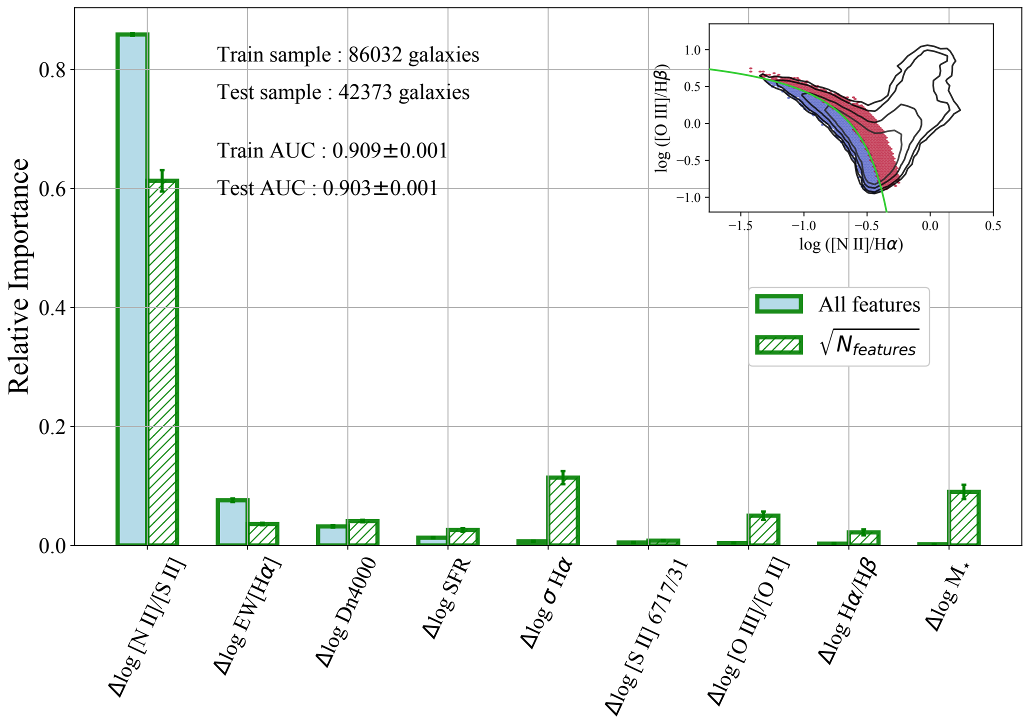
Max depth : None ; **Min sample leaf** : 500 (class); 100 (regression)
Metrics : GINI (class); MSE (regression)

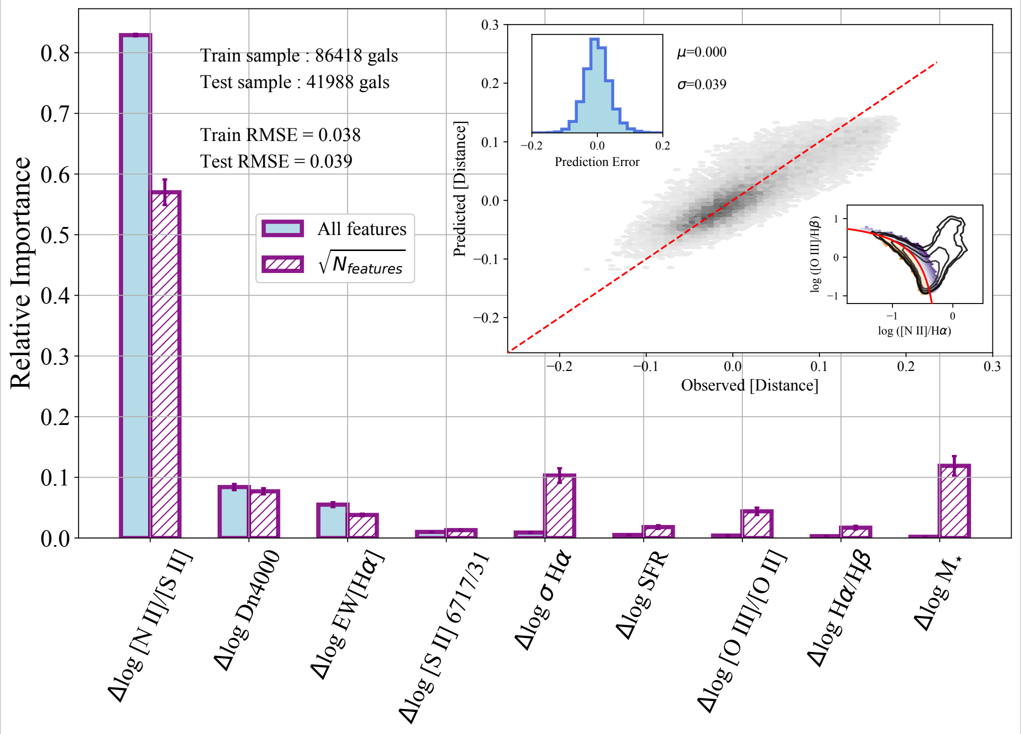
Classification



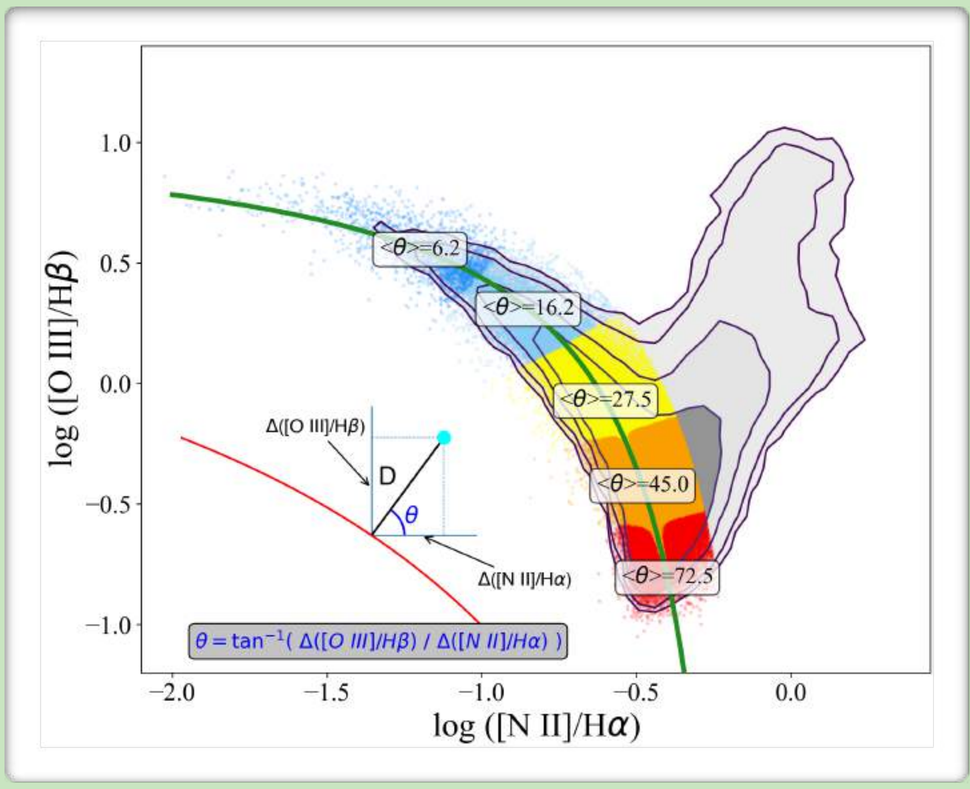
Regression



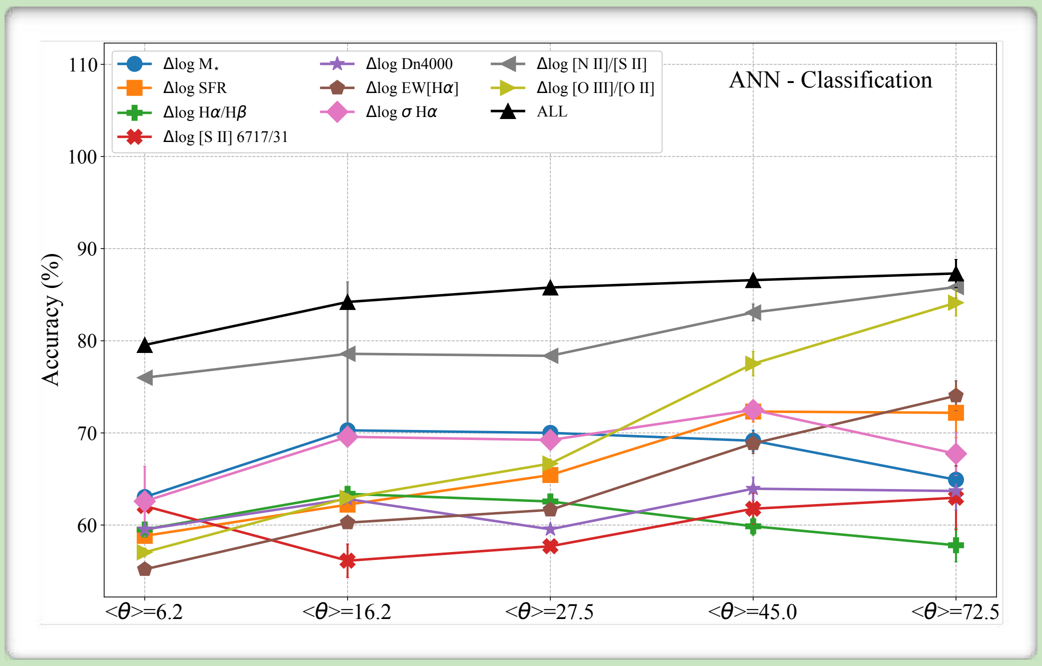


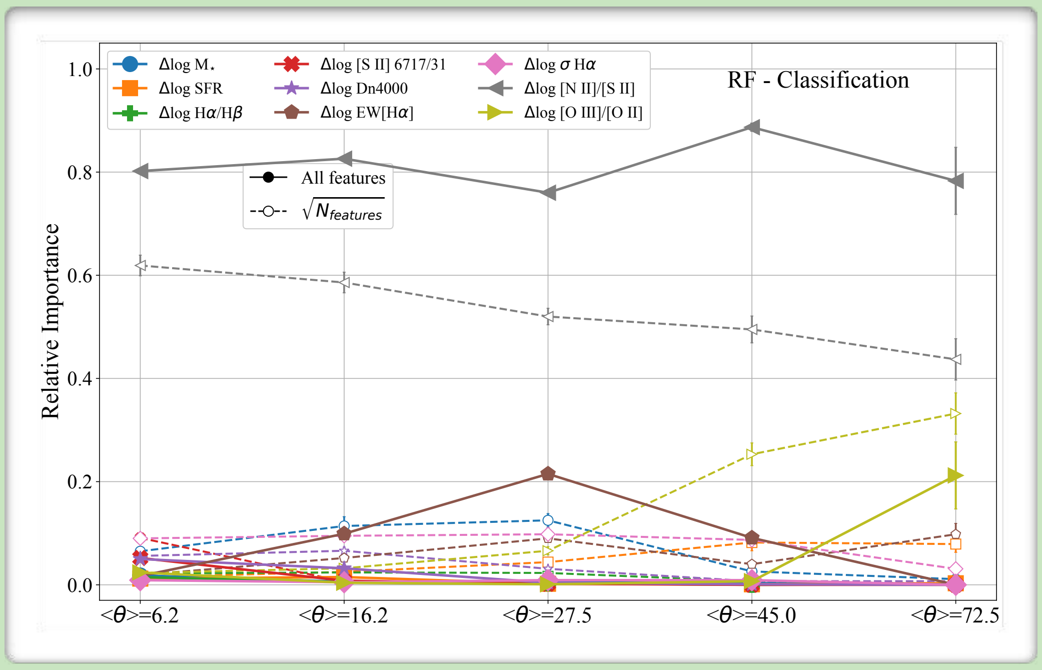


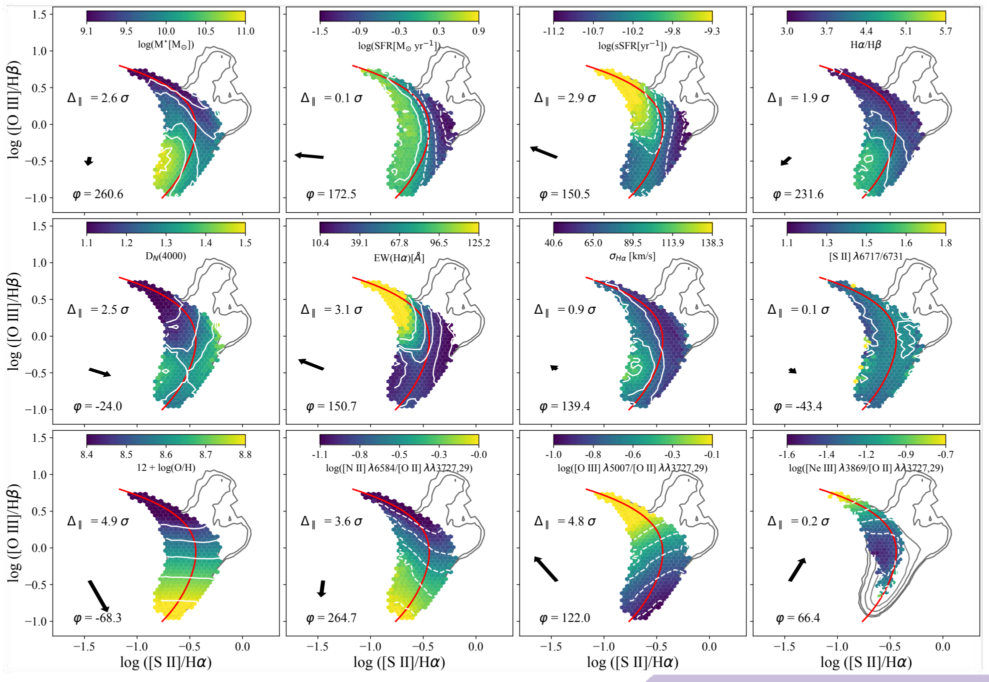
- Train - Test sample : 67% - 33% galaxies
- 30 independent runs
- ALL-parameter run achieves >80% accuracy (class) and > 45%IoR (regression)
- Δ log([N II]/[S II]) (i.e., deviations in N/O abundance) ranked as the most important parameter

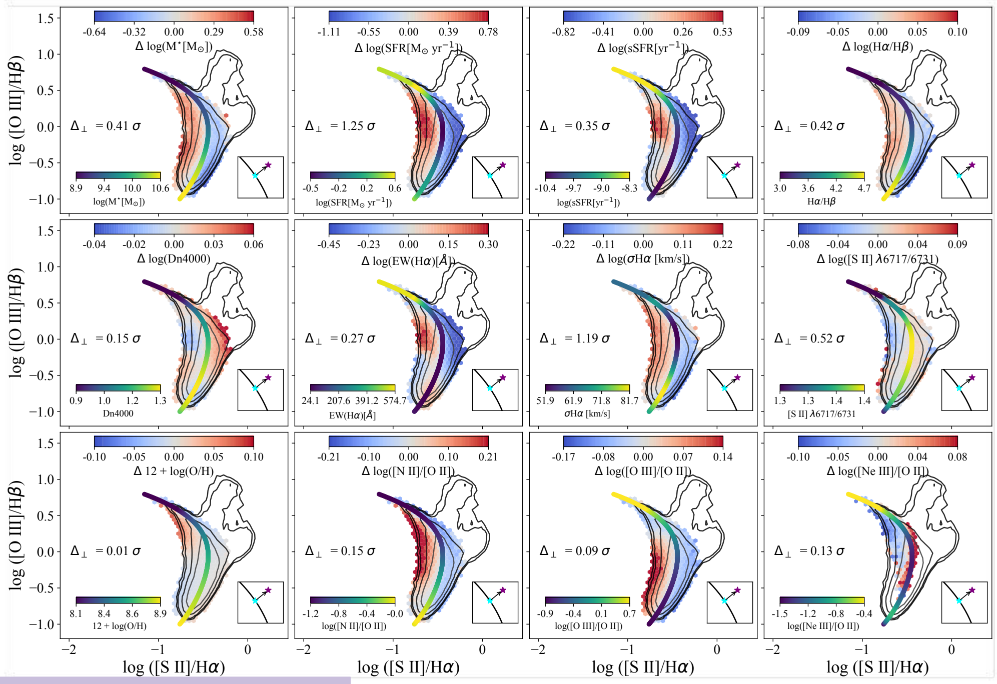


- Inclination ϑ of the orthogonal offset-vector changes along the sequence, affecting the strength of its [N II]/Hα- and [O III]/Hβ- components
- We assess the performance of the parameters by dividing the diagram in five sectors of different <ϑ>

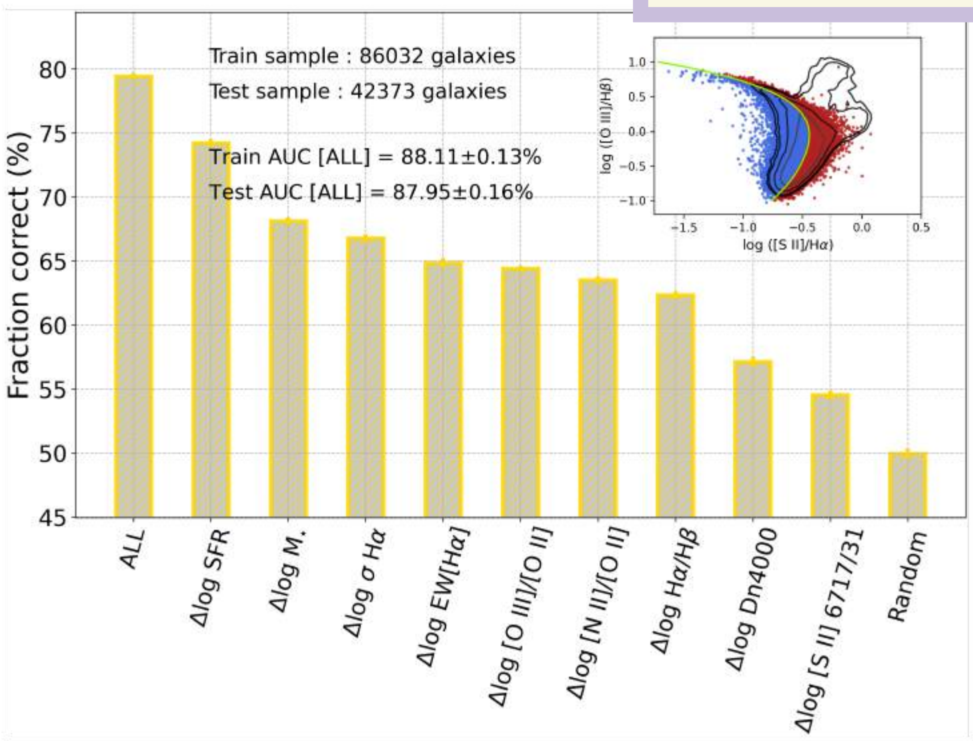


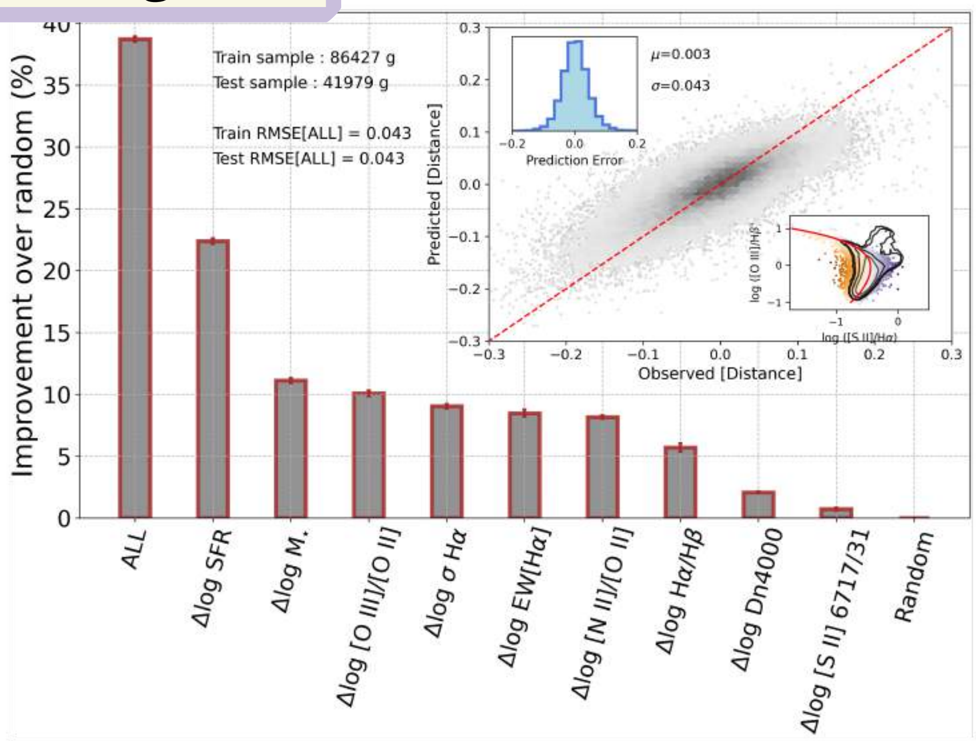


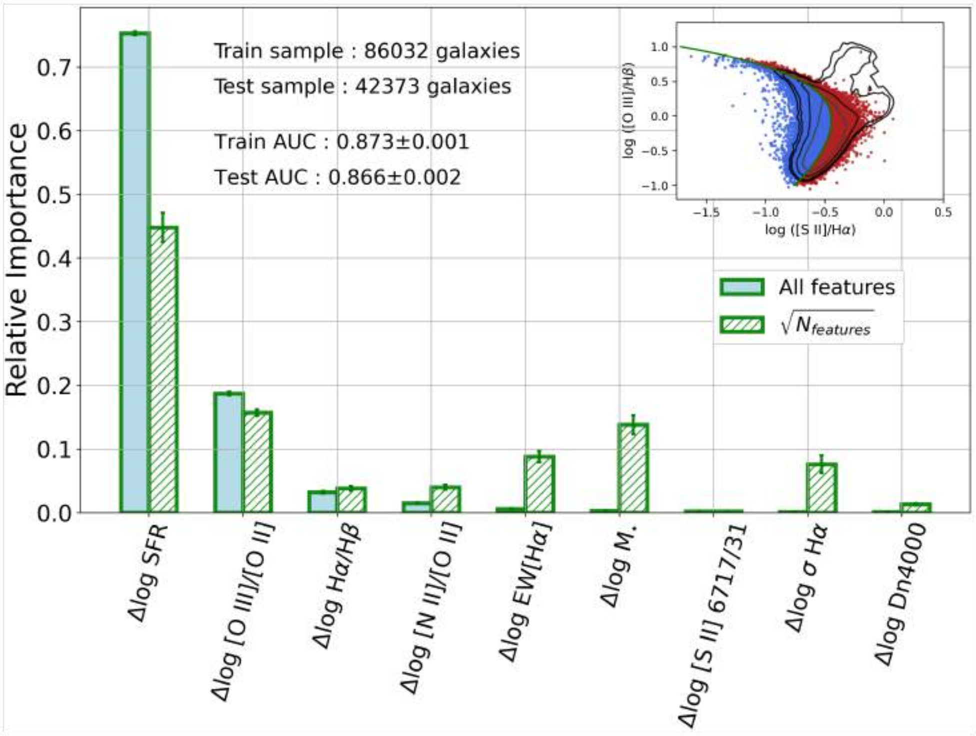


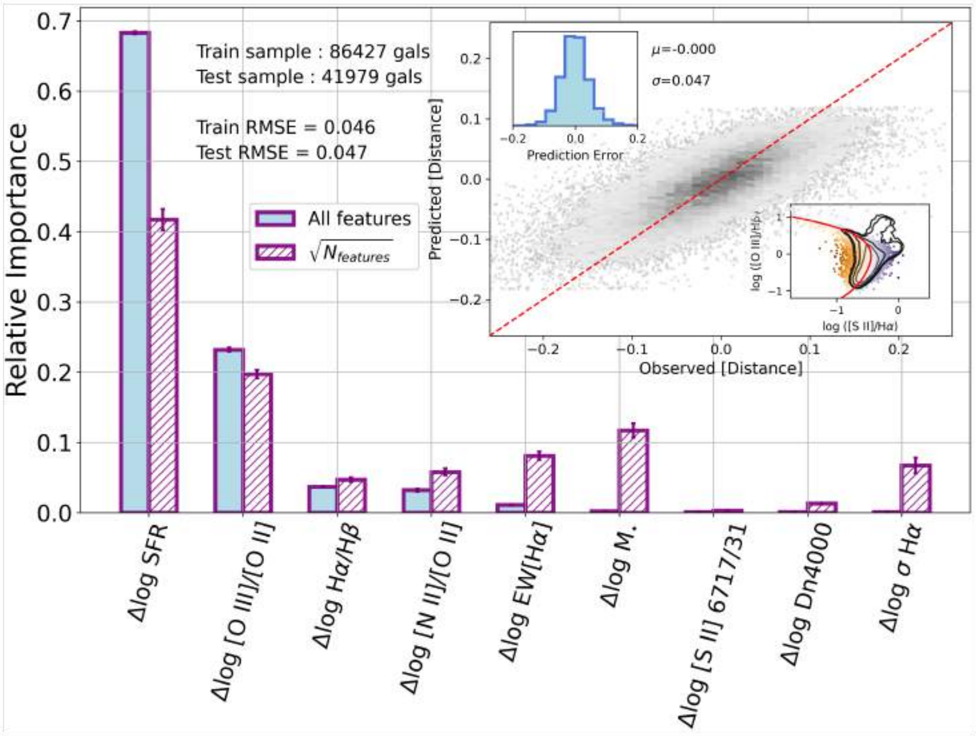


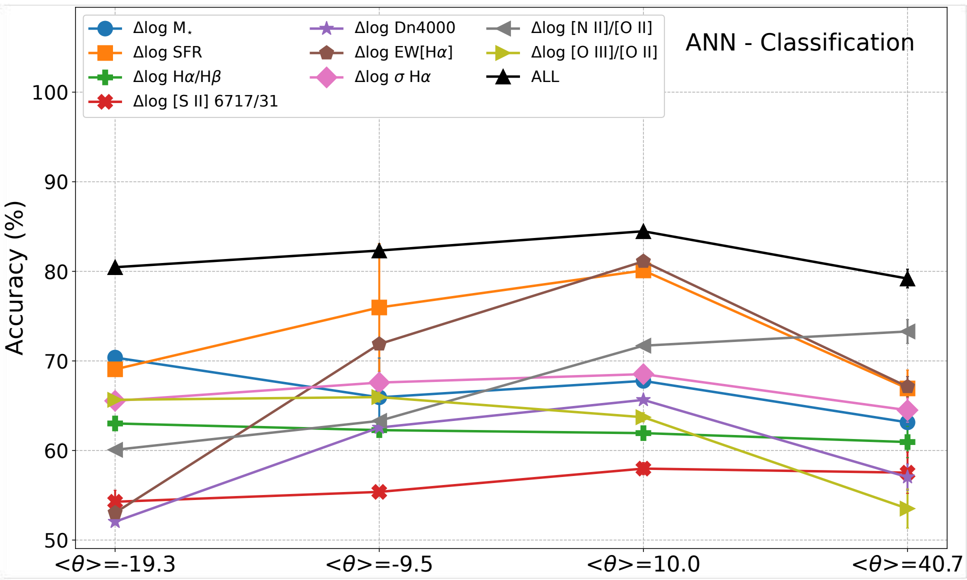
The [S II]-BPT diagram

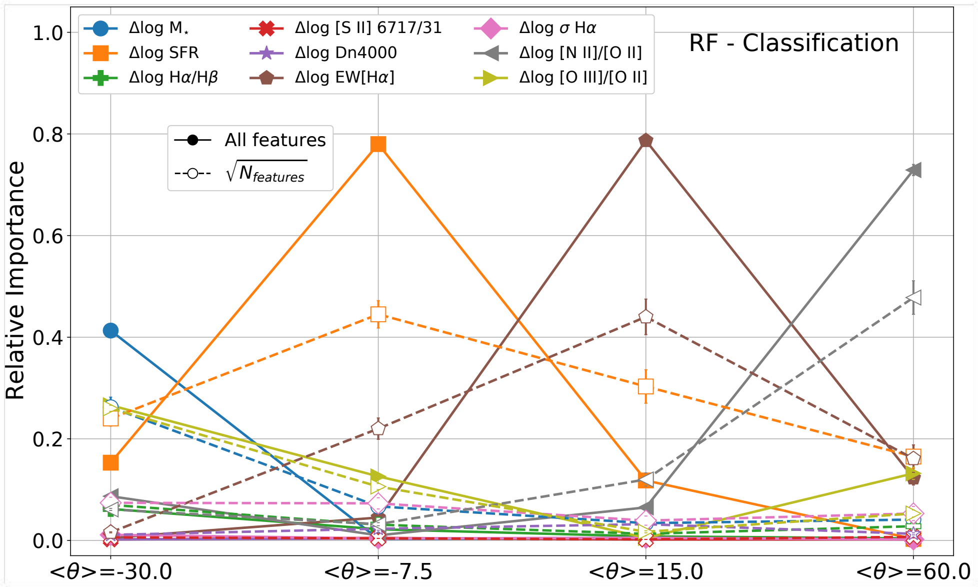

Train sample : 86032 galaxies
Test sample : 42373 galaxies
Train AUC [ALL] = 88.11±0.13%
Test AUC [ALL] = 87.95±0.16%
Fraction correct (%)
ALL
Δlog SFR
Δlog M.
Δlog σ Hα
Δlog EW[Hα]
Δlog [O III]/[O II]
Δlog [N II]/[O II]
Δlog Hα/Hβ
Δlog Dn4000
Δlog [S II] 6717/31
Random
log ([O III]/Hβ)
log ([S II]/Hα)


Train sample : 86427 g
Test sample : 41979 g
Train RMSE[ALL] = 0.043
Test RMSE[ALL] = 0.043
μ=0.003
σ=0.043
Prediction Error
Predicted [Distance]
Observed [Distance]
Improvement over random (%)
ALL
Δlog SFR
Δlog M.
Δlog [O III]/[O II]
Δlog σ Hα
Δlog EW[Hα]
Δlog [N II]/[O II]
Δlog Hα/Hβ
Δlog Dn4000
Δlog [S II] 6717/31
Random
log ([O III]/Hβ)
log ([S II]/Hα)


Train sample : 86032 galaxies
Test sample : 42373 galaxies
Train AUC : 0.873±0.001
Test AUC : 0.866±0.002
Relative Importance
All features
$\sqrt{N_{features}}$
Δlog SFR
Δlog [O III]/[O II]
Δlog Hα/Hβ
Δlog [N II]/[O II]
Δlog EW[Hα]
Δlog M.
Δlog [S II] 6717/31
Δlog σ Hα
Δlog Dn4000
log ([O III]/Hβ)
log ([S II]/Hα)


Train sample : 86427 gals
Test sample : 41979 gals
Train RMSE = 0.046
Test RMSE = 0.047
μ=-0.000
σ=0.047
Prediction Error
Predicted [Distance]
Observed [Distance]
Relative Importance
All features
$\sqrt{N_{features}}$
Δlog SFR
Δlog [O III]/[O II]
Δlog Hα/Hβ
Δlog [N II]/[O II]
Δlog EW[Hα]
Δlog M.
Δlog [S II] 6717/31
Δlog Dn4000
Δlog σ Hα
log ([O III]/Hβ)
log ([S II]/Hα)


ANN - Classification
Δlog M.
Δlog SFR
Δlog Hα/Hβ
Δlog [S II] 6717/31
Δlog Dn4000
Δlog EW[Hα]
Δlog σ Hα
Δlog [N II]/[O II]
Δlog [O III]/[O II]
ALL
Accuracy (%)
<θ>=-19.3
<θ>=-9.5
<θ>=10.0
<θ>=40.7


RF - Classification
Δlog M.
Δlog SFR
Δlog Hα/Hβ
Δlog [S II] 6717/31
Δlog Dn4000
Δlog EW[Hα]
Δlog σ Hα
Δlog [N II]/[O II]
Δlog [O III]/[O II]
All features
$\sqrt{N_{features}}$
Relative Importance
<θ>=-30.0
<θ>=-7.5
<θ>=15.0
<θ>=60.0