Dimensionality Reduction and Clustering of the PHANGS dataset

Dalya Baron
Carnegie Observatories

Vatican Observatory Summer School on Big Data and Machine Learning 2023 (VOSS-2023)

PHANGS: Physics at High Angular resolution in Nearby GalaxieS

High resolution observations of nearby galaxies with several telescopes, covering a large fraction of the electromagnetic spectrum:

- 90 galaxies @ ALMA: CO emission spectra.
- 38 galaxies @ HST in 5 bands: UV to optical photometry.
- 19 galaxies @ MUSE: optical spectroscopy.
- 19 galaxies @ JWST in 8 bands: near and mid-infrared photometry.

Between 1,000,000 - 10,000,000 pixels per galaxy!

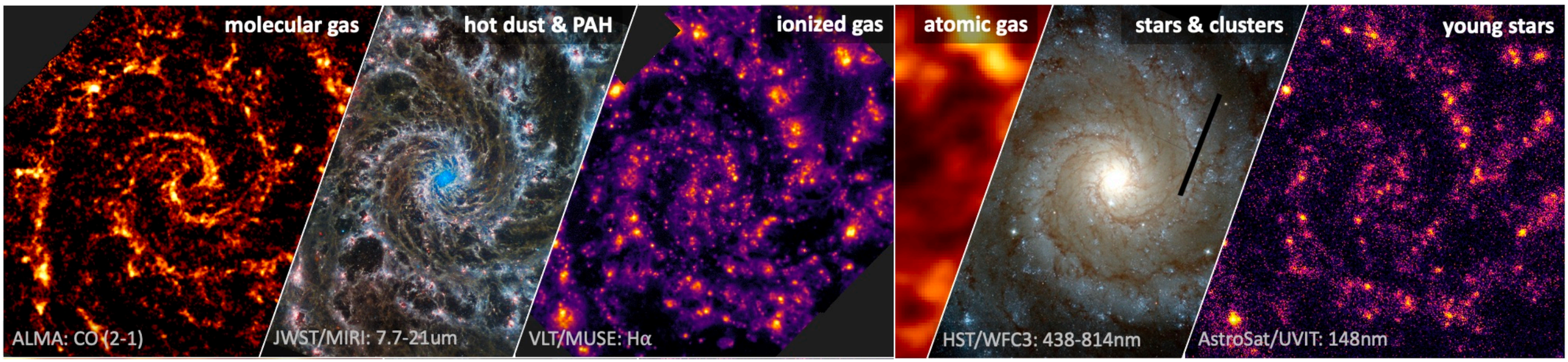
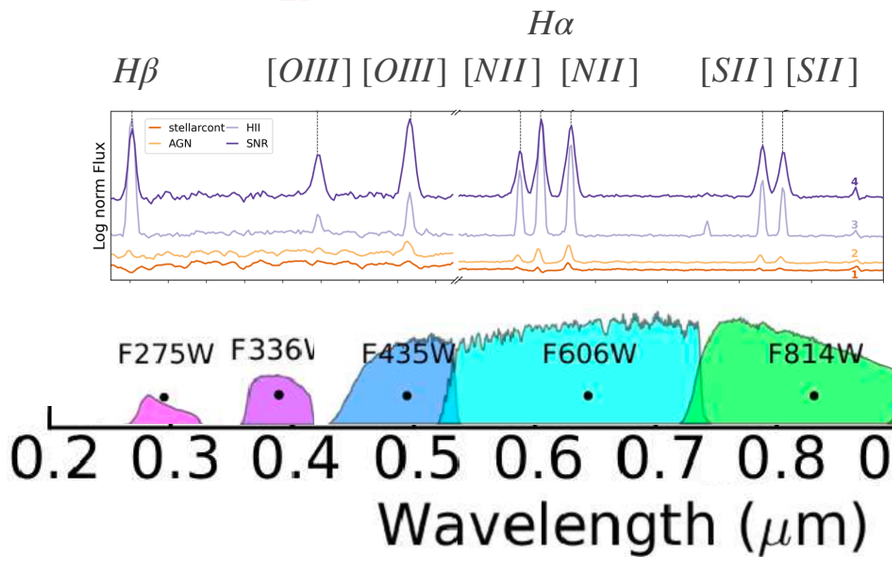
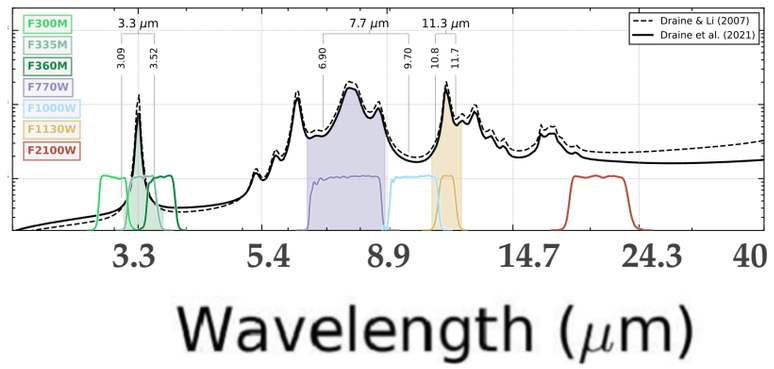


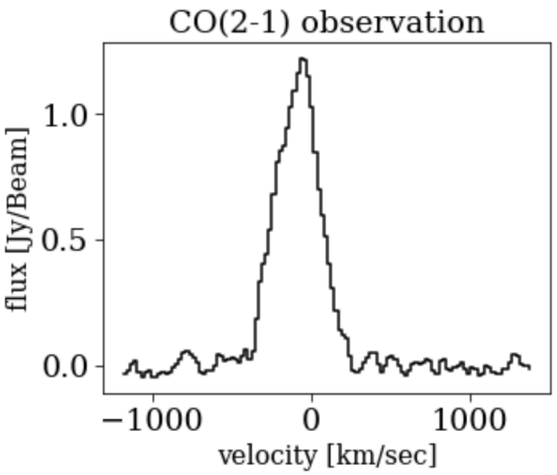
Image by the PHANGS collaboration

By J. Sun

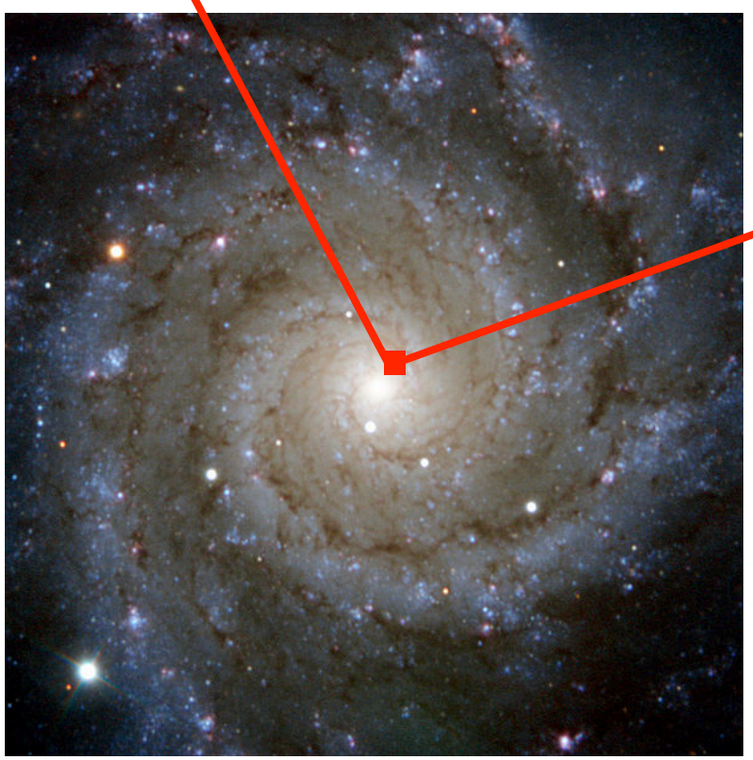
Complex dataset with heterogeneous features







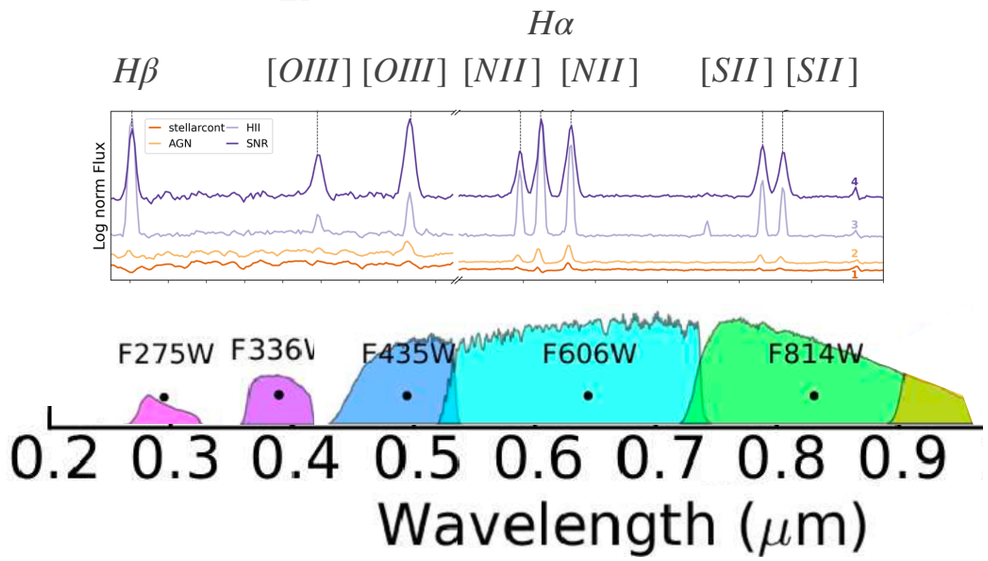
UV | optical | near-infrared | mid-infrared | mm

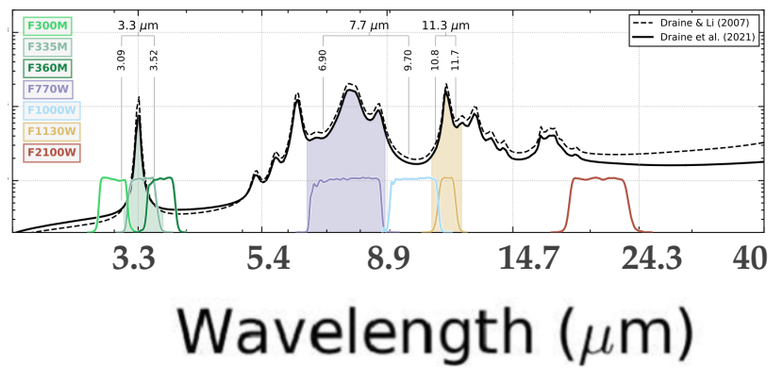


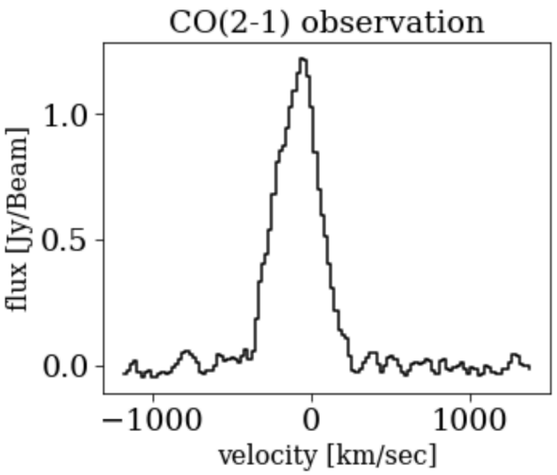
~1000 pixels

~1000 pixels

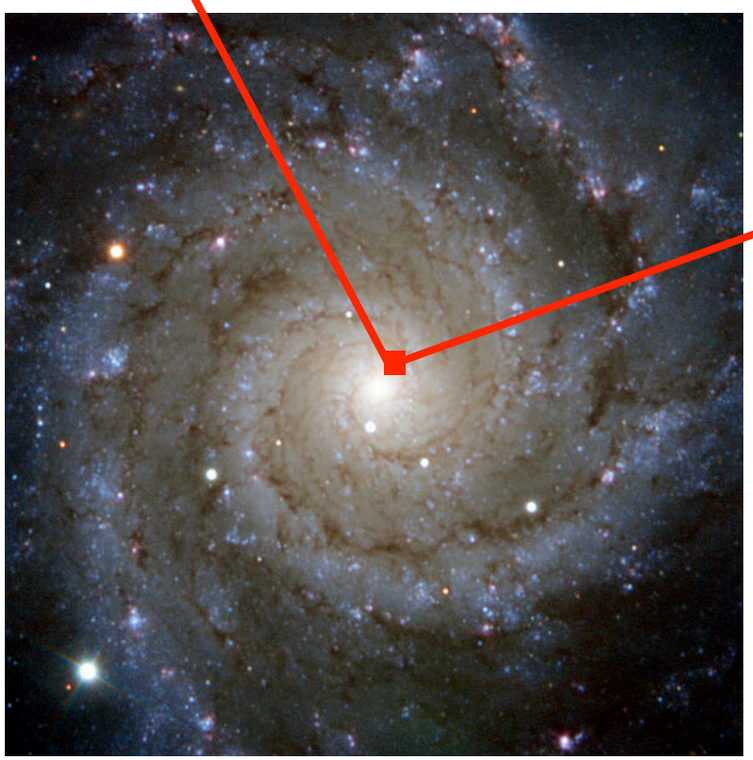
Complex dataset with heterogeneous features







UV | optical | near-infrared | mid-infrared | mm



~1000 pixels

~1000 pixels

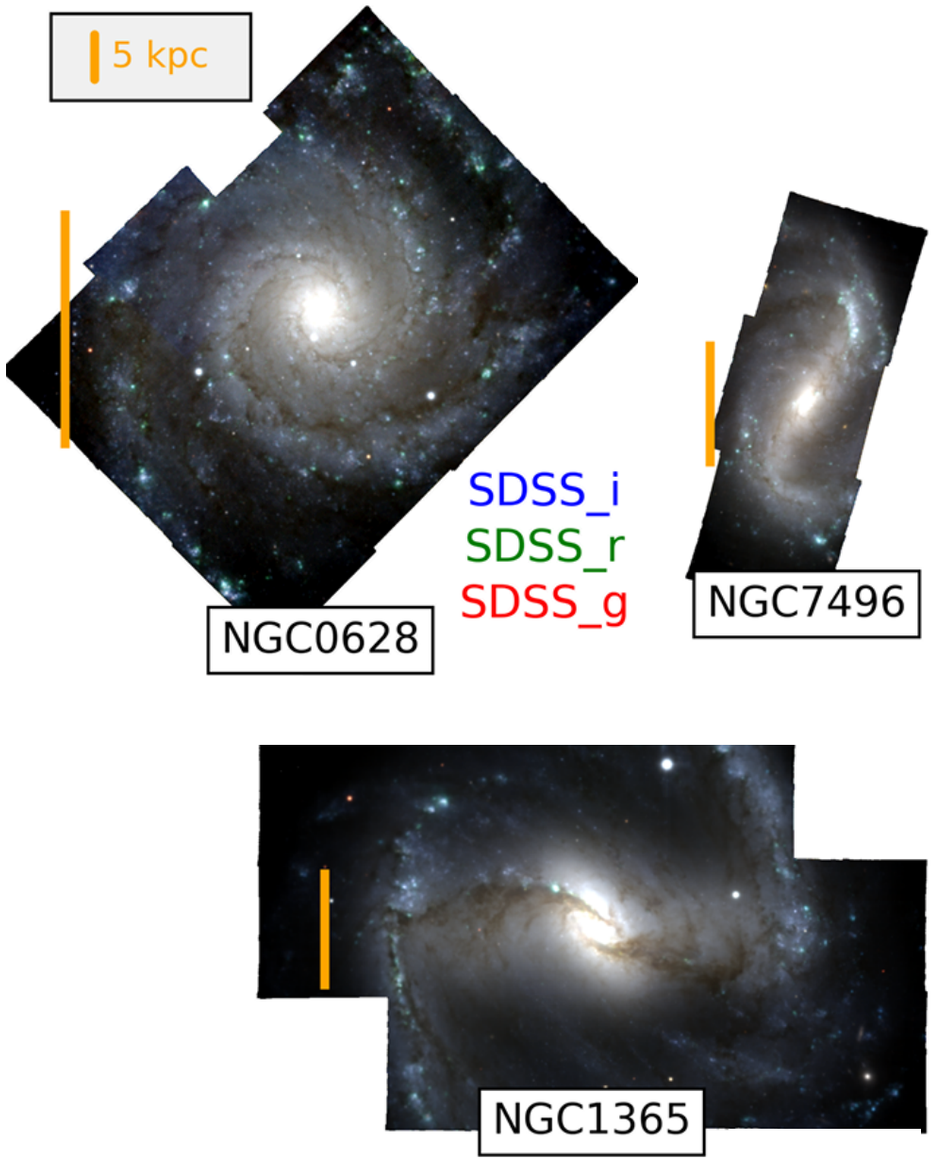
Different resolutions
Different PSFs
Different noise characteristics

Why did I choose this dataset?

- Discovery potential:
  - New observations at a previously unexplored range tend to lead to new discoveries.
  - JWST is exciting.
- Interesting algorithmic questions:
  - How to combine such a heterogeneous data? Can I combine the spectroscopy directly with photometry?
  - How to take into account upper limits and Nan values?

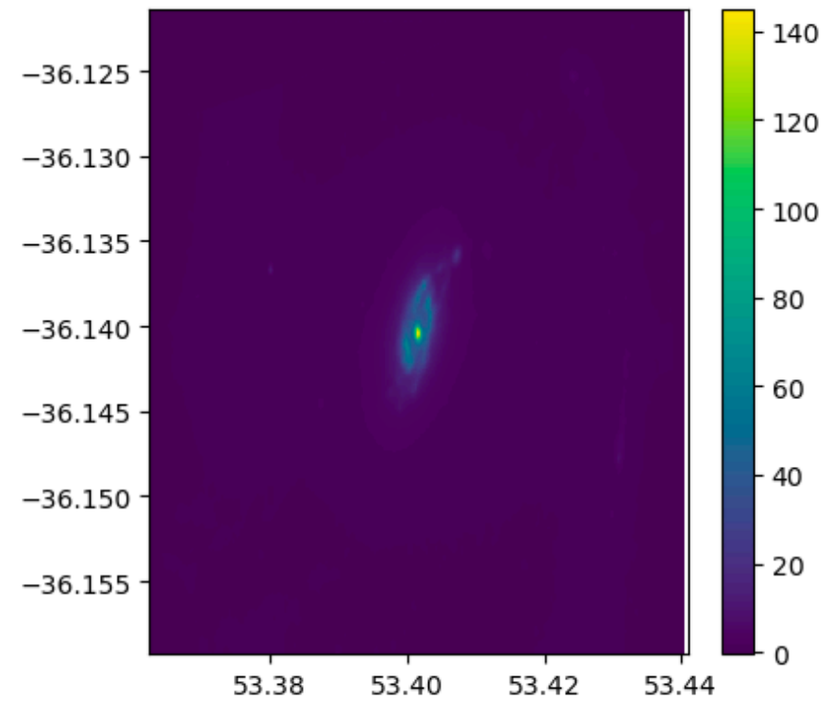
Data selection and pre-processing

- Selected the three galaxies that were already observed with JWST: data is already reduced, vetted, and publicly-available.
- For simplicity, decided to work with a list of derived features rather than the raw data.
- Before deriving features, all the maps in the different wavelengths were convolved to a common resolution of 150 pc.



Before cleaning: inspecting the maps

Same data, three different scales:




```
plt.pcolormesh(RA, DEC, JWST_image)
plt.colorbar()
```



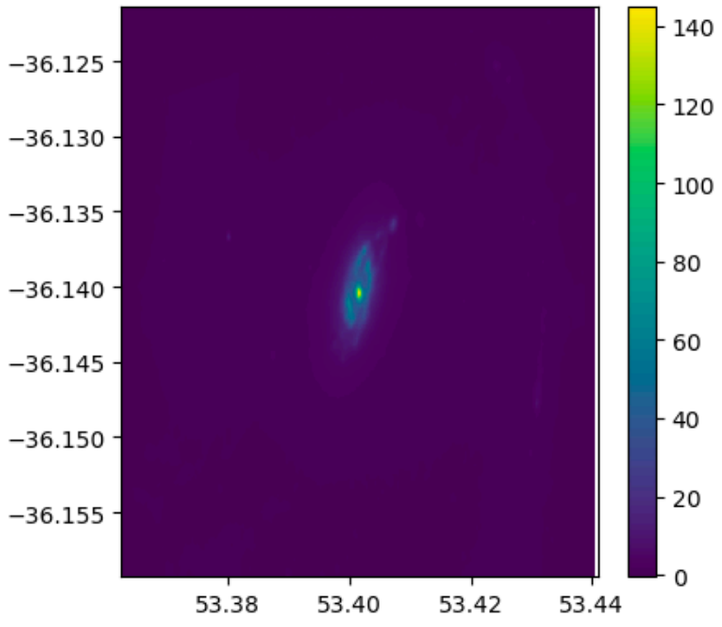
```
plt.pcolormesh(RA, DEC, JWST_image,
               norm=LogNorm())
plt.colorbar()
```

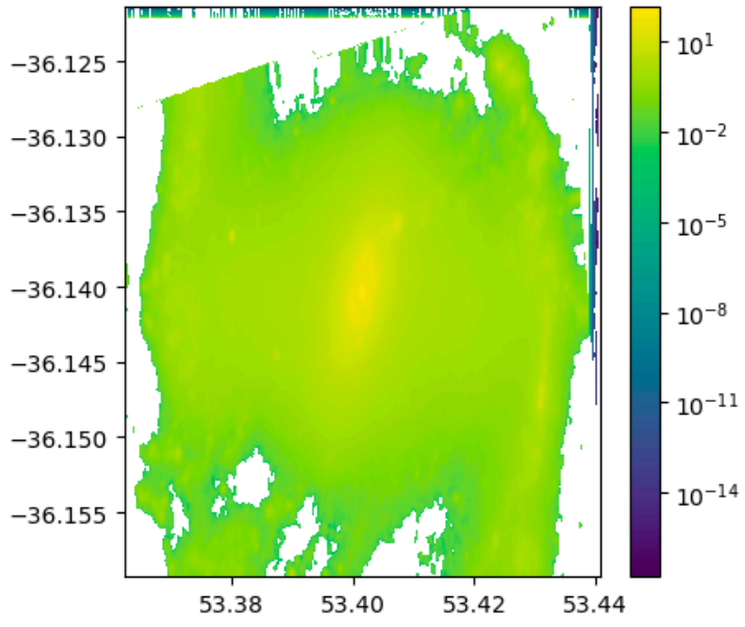


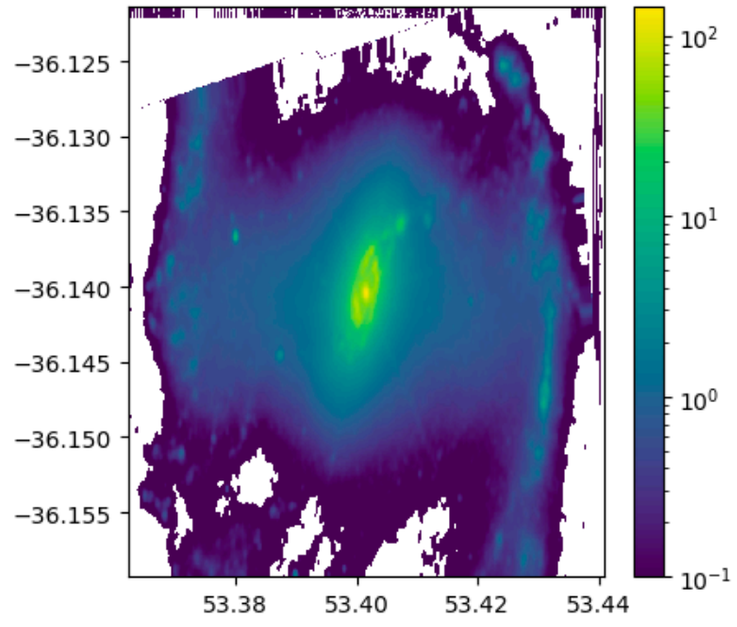
```
plt.pcolormesh(RA, DEC, JWST_image,
               norm=LogNorm(vmin=0.1))
plt.colorbar()
```

Before cleaning: inspecting the maps

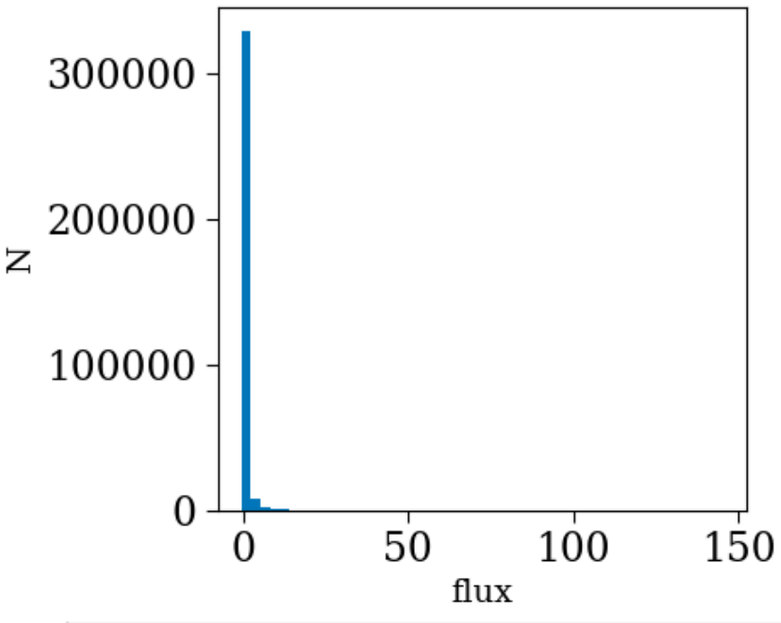
Same data, three different scales:



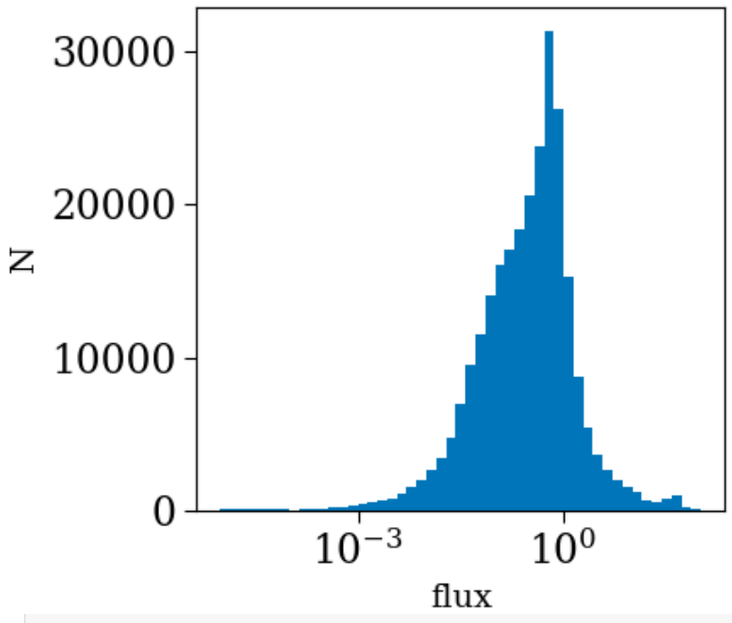




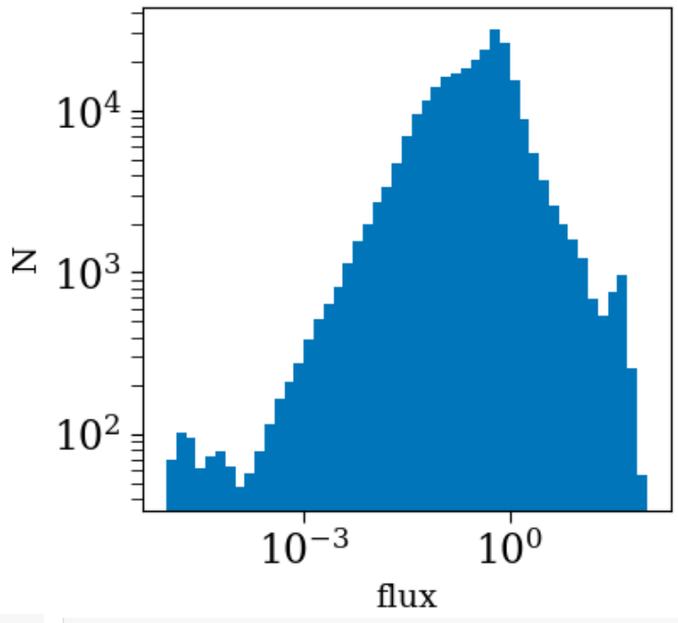
Histograms:




```
plt.hist(JWST_image.flatten(),
         bins=50)
```



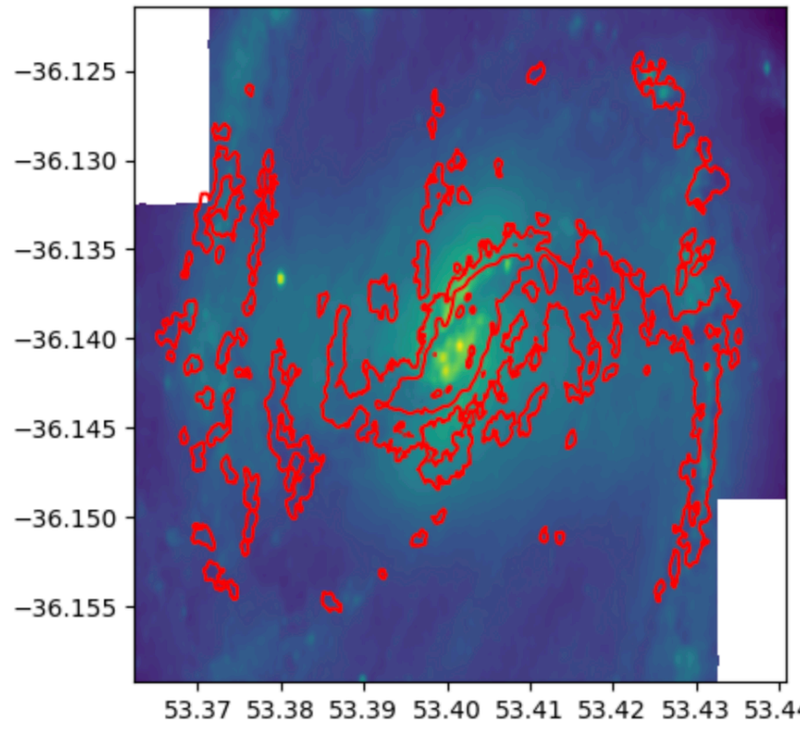

```
plt.hist(JWST_image.flatten(),
         bins=np.logspace(-5, 2, 50))
plt.xscale("log")
```




```
plt.hist(JWST_image.flatten(),
         bins=np.logspace(-5, 2, 50))
plt.xscale("log")
plt.yscale("log")
```

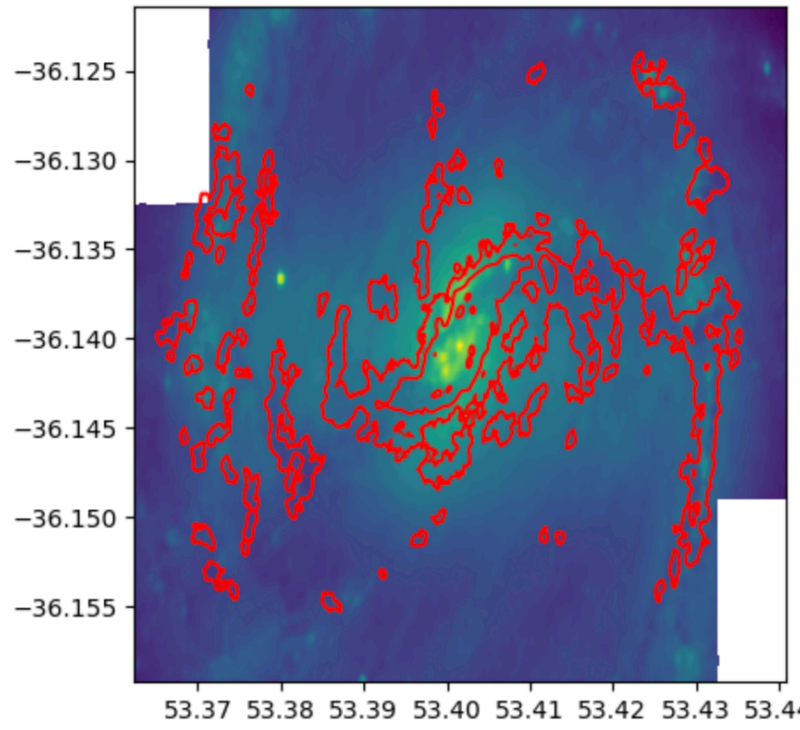
Selected features

1. From ALMA data: CO flux and width whenever detected. Pixels where CO is not detected were discarded.



Selected features

1. From ALMA data: CO flux and width whenever detected. Pixels where CO is not detected were discarded.

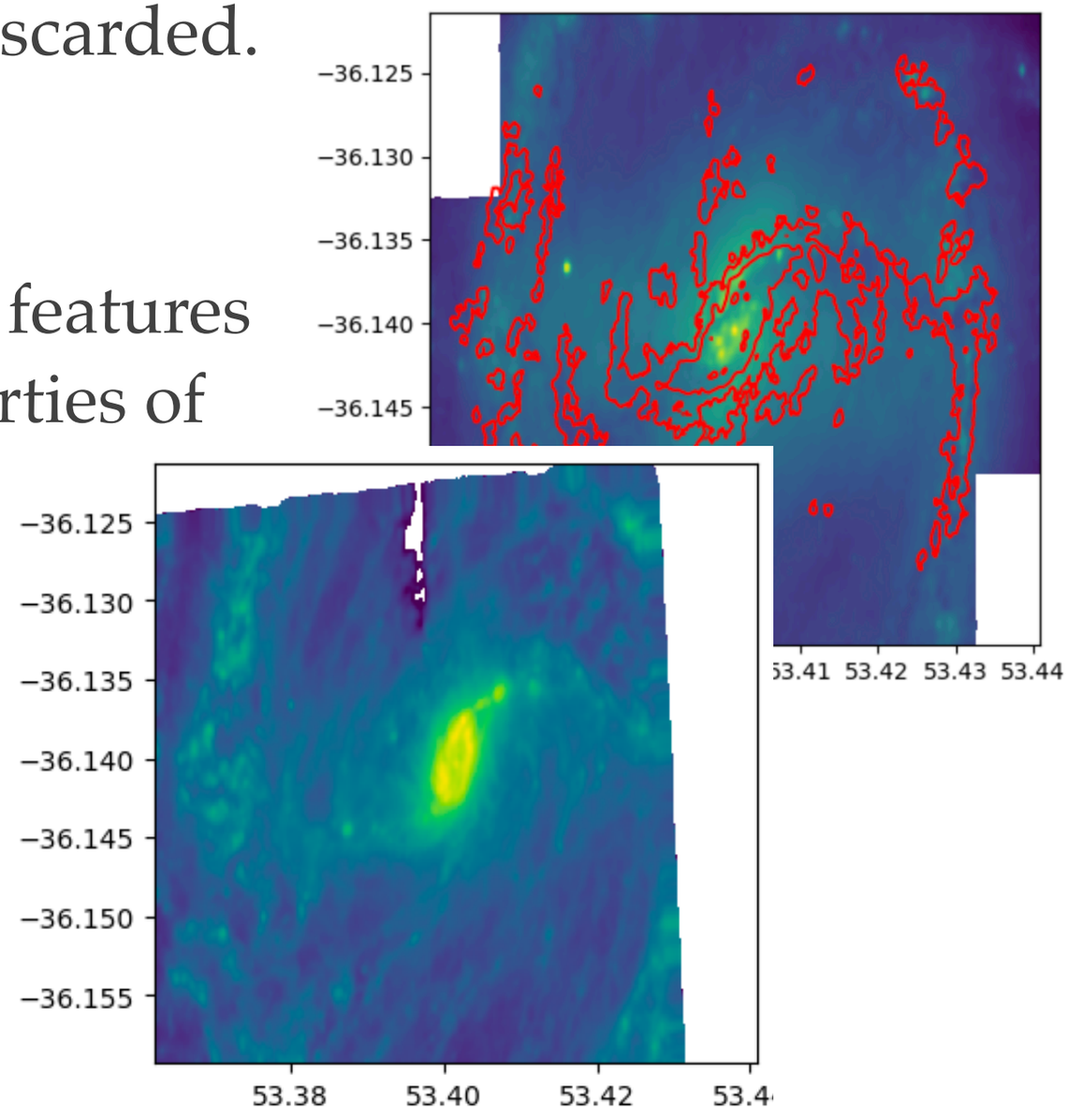


2. From MUSE data: a set of derived features that correspond to different properties of the stars and gas. I attempted to be as close as possible to the physics. For example:

- Derive the dust-corrected $H\alpha$ flux.
- Use line ratios of interest (e.g., $[OIII]/H\beta$) rather the emission line fluxes.

Selected features

1. From ALMA data: CO flux and width whenever detected. Pixels where CO is not detected were discarded.



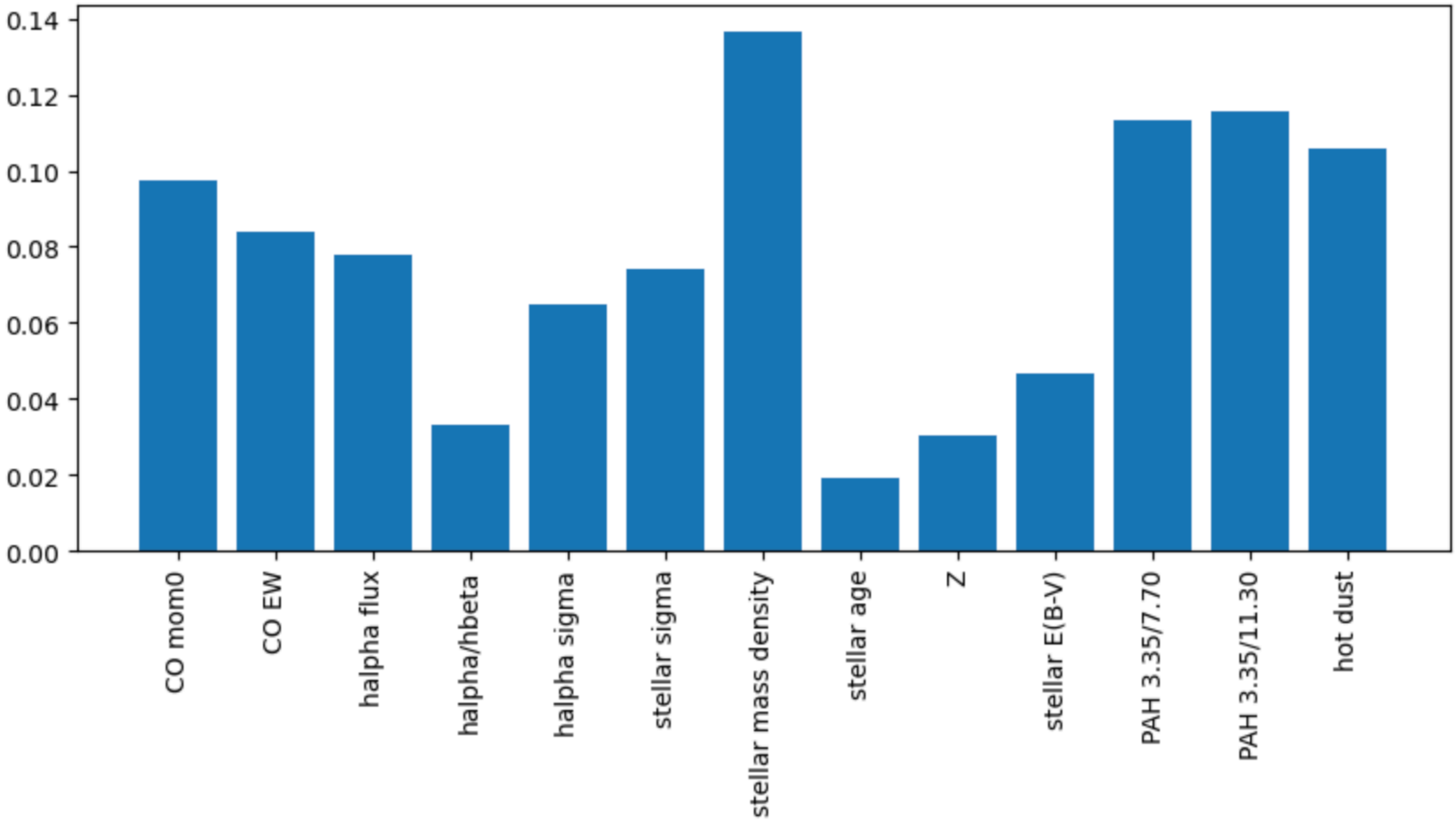
2. From MUSE data: a set of derived features that correspond to different properties of the stars and gas. I attempted to be as possible to the physics. For exa

   - Derive the dust-corrected *Hα* flu
   - Use line ratios of interest (e.g., [O rather the emission line fluxes.

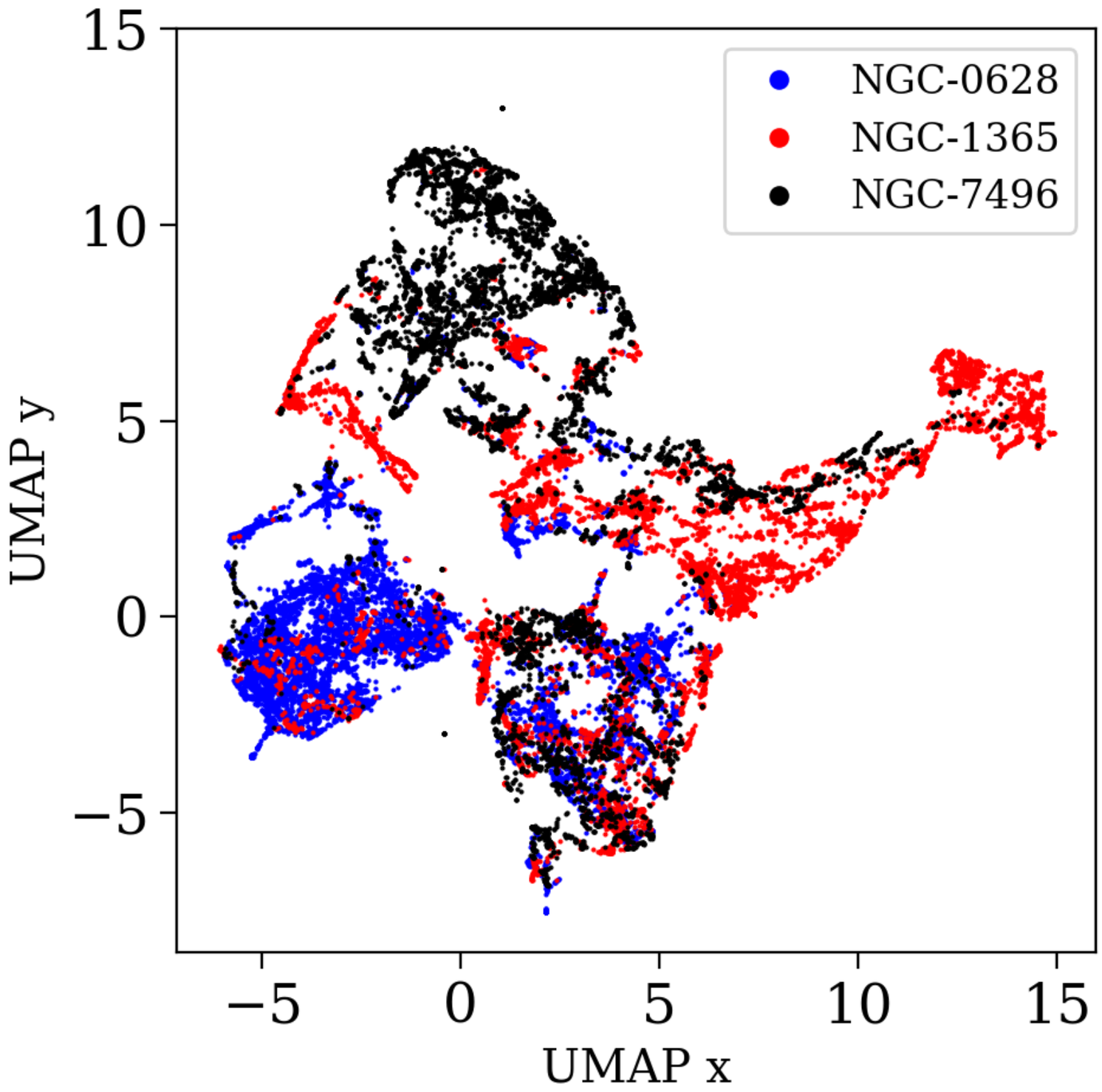
3. From JWST data: trial and error. Started with using the flux in the different bands, then learned about what these bands trace and move to band ratios.

Exploration of feature importance metrics

I spent more than a month going back and forth between different sets of features. I decided to select a set of features that are actually highly-correlated with the hope of finding some interesting relations.

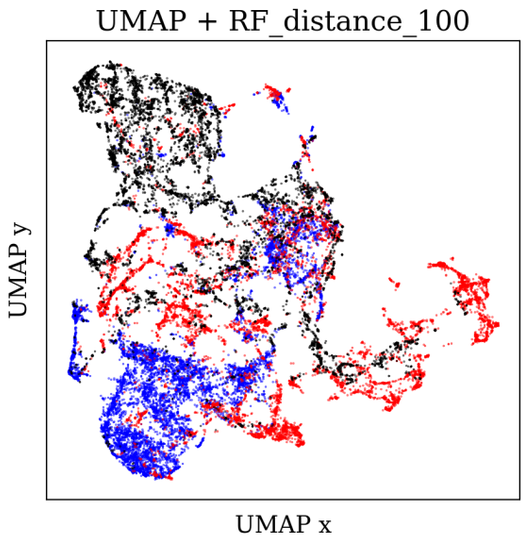


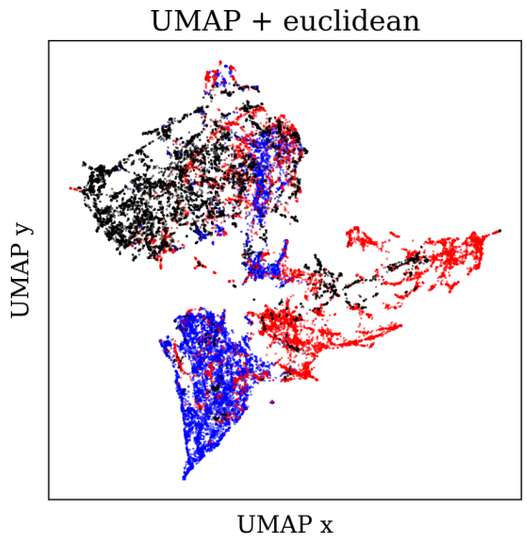
Adopted UMAP for clustering

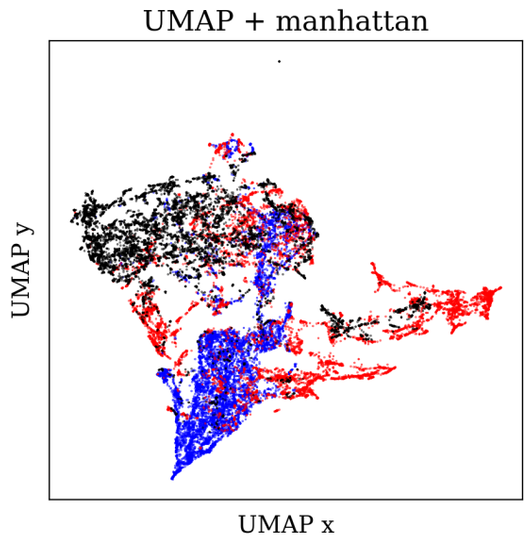


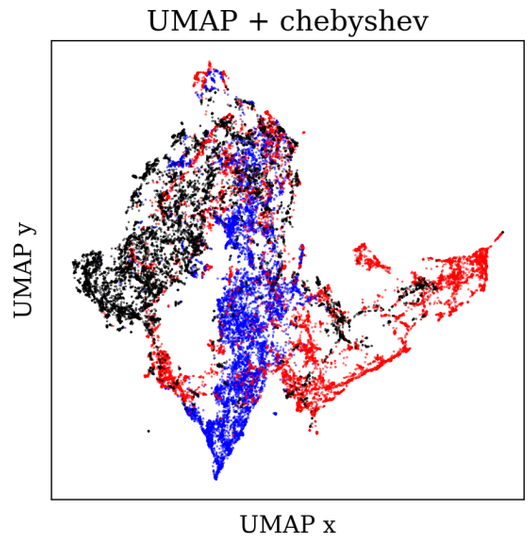
- Adopted distance measure: unsupervised Random Forest (see Baron & Poznanski 2017 for details).
- Adopted UMAP hyper-parameters:
  - min_dist = 0
  - n_neighbors = 25

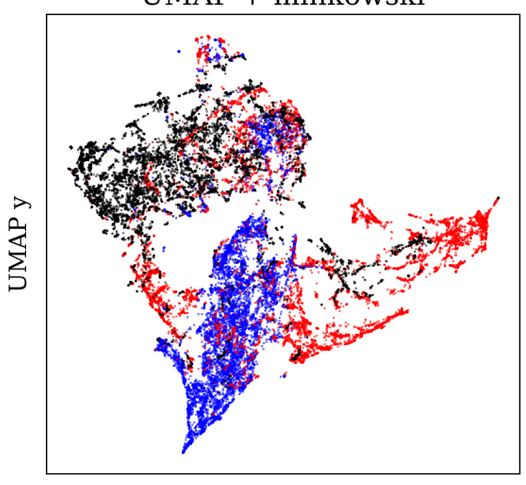
Impact of different distance measures

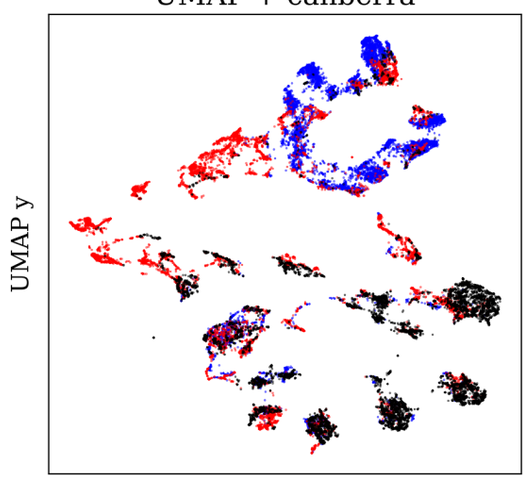


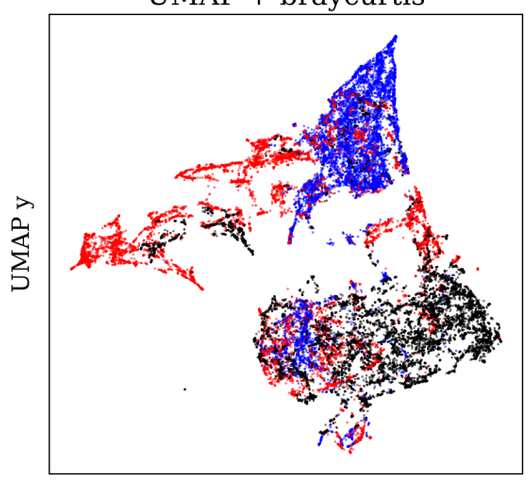


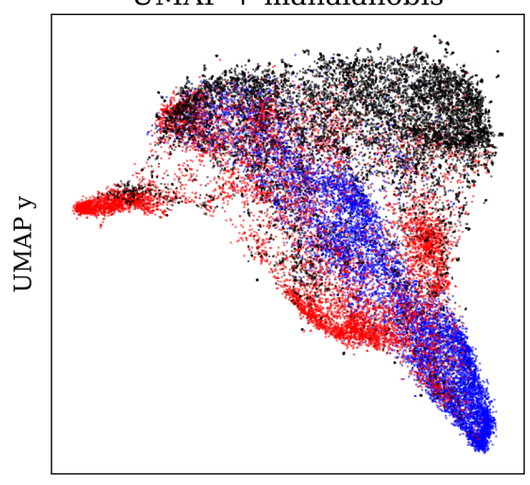


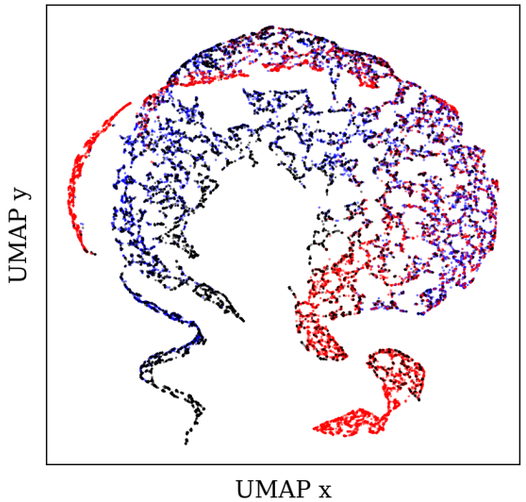


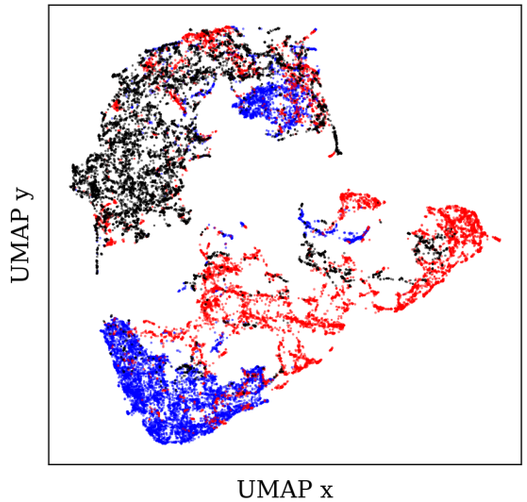


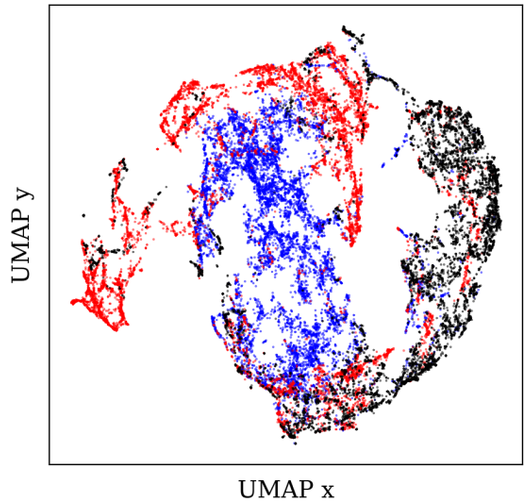








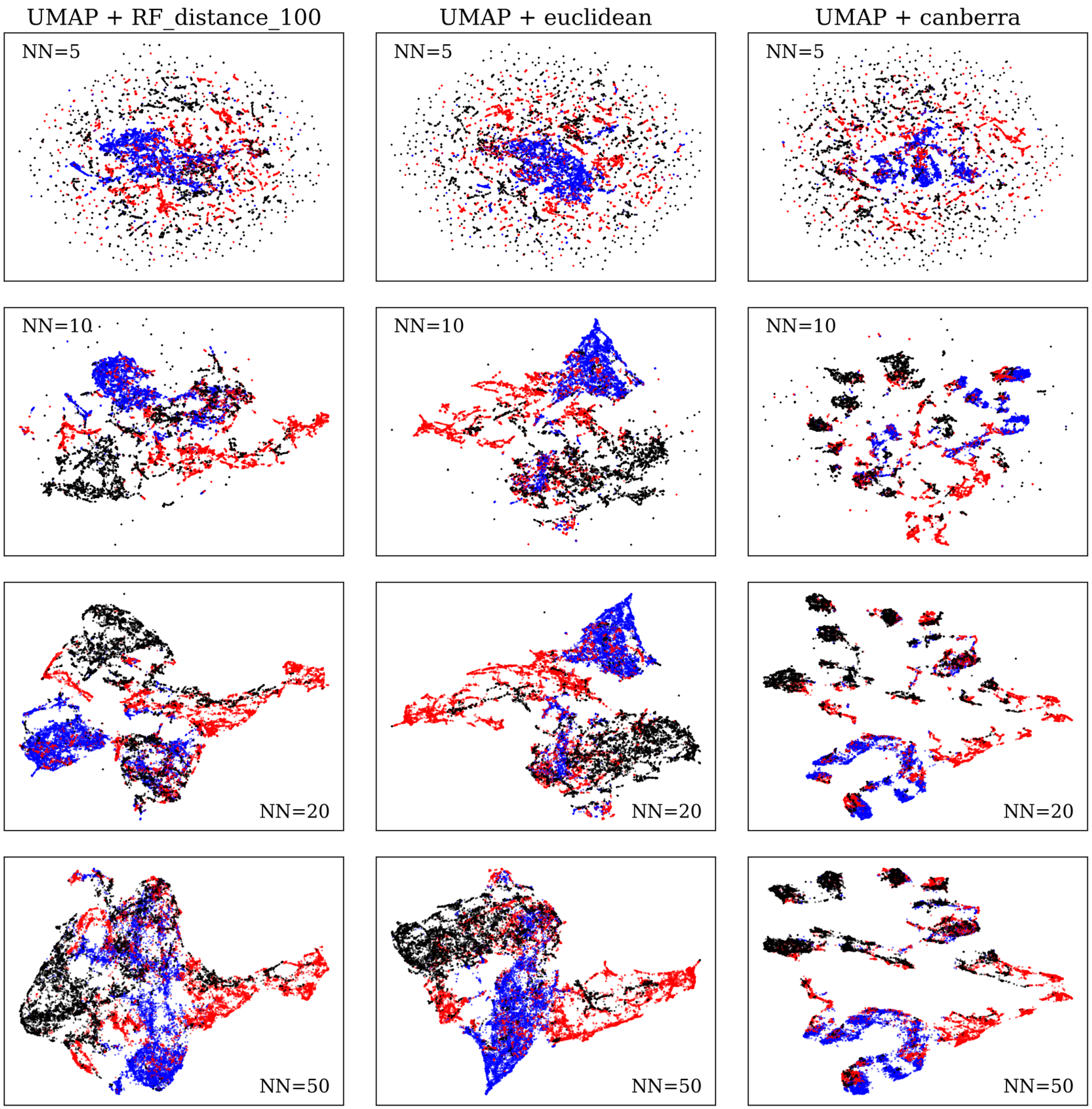




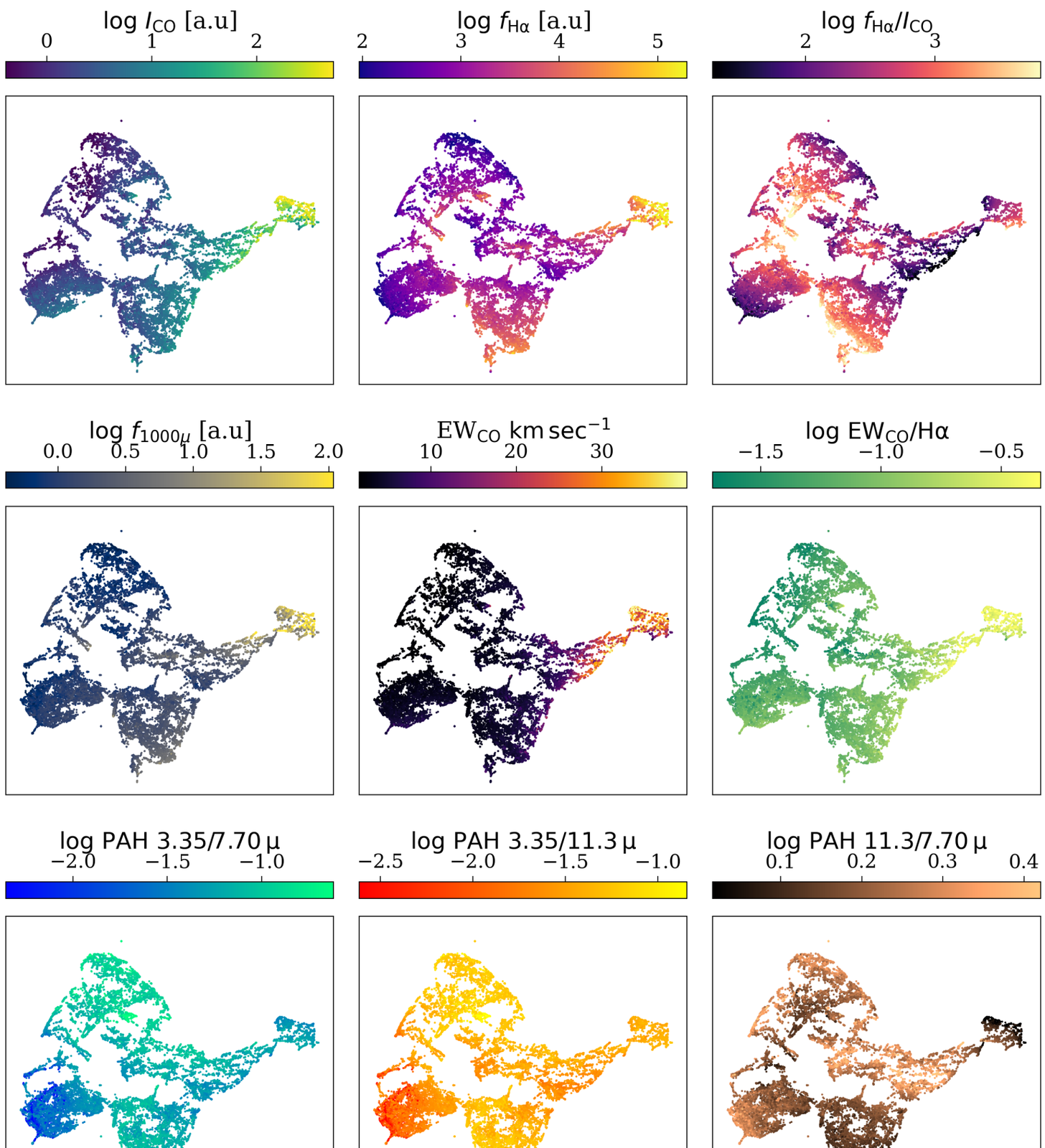
Adopted UMAP hyper-parameters:

- min_dist = 0
- n_neighbors = 25

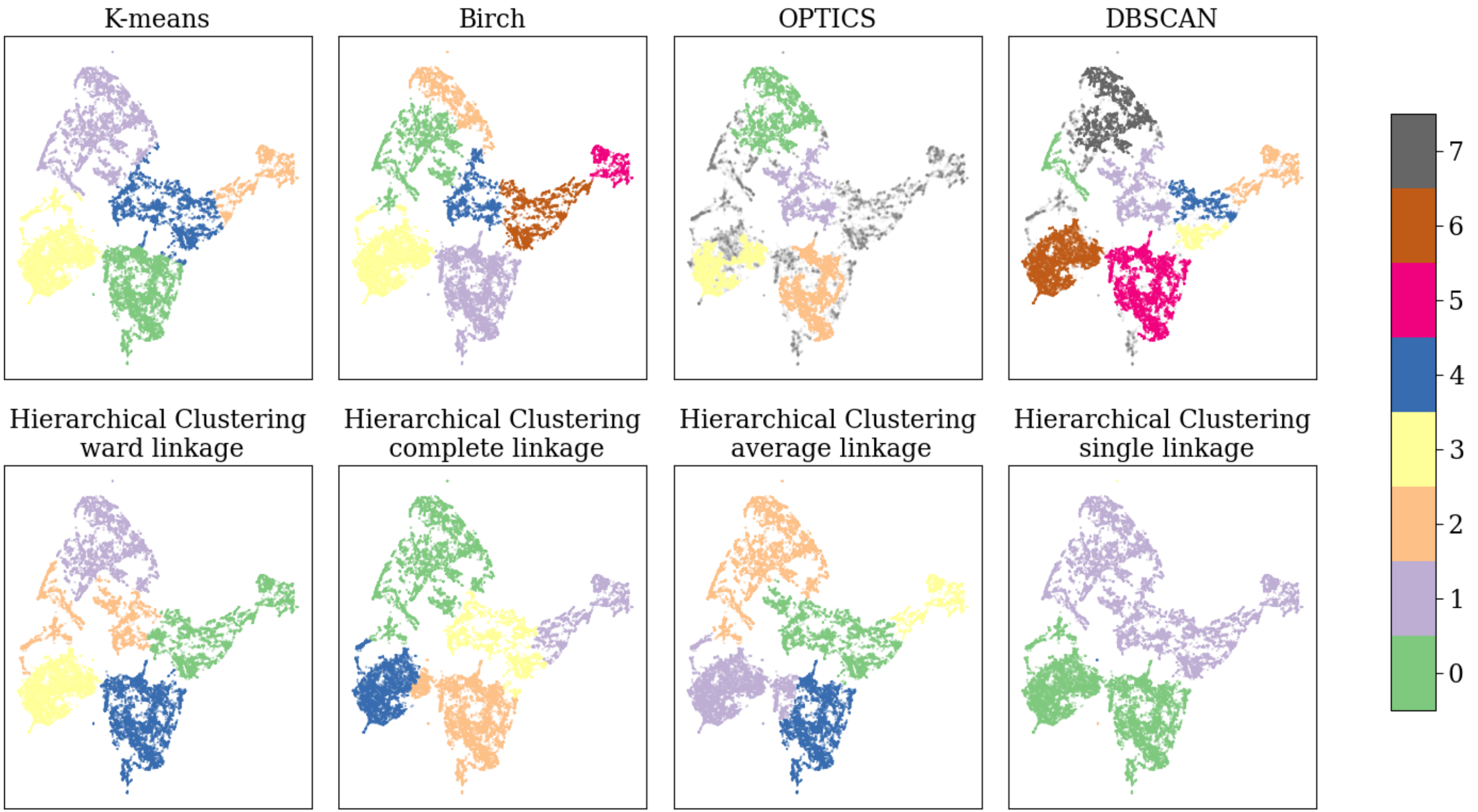
Impact of different UMAP hyper-parameters



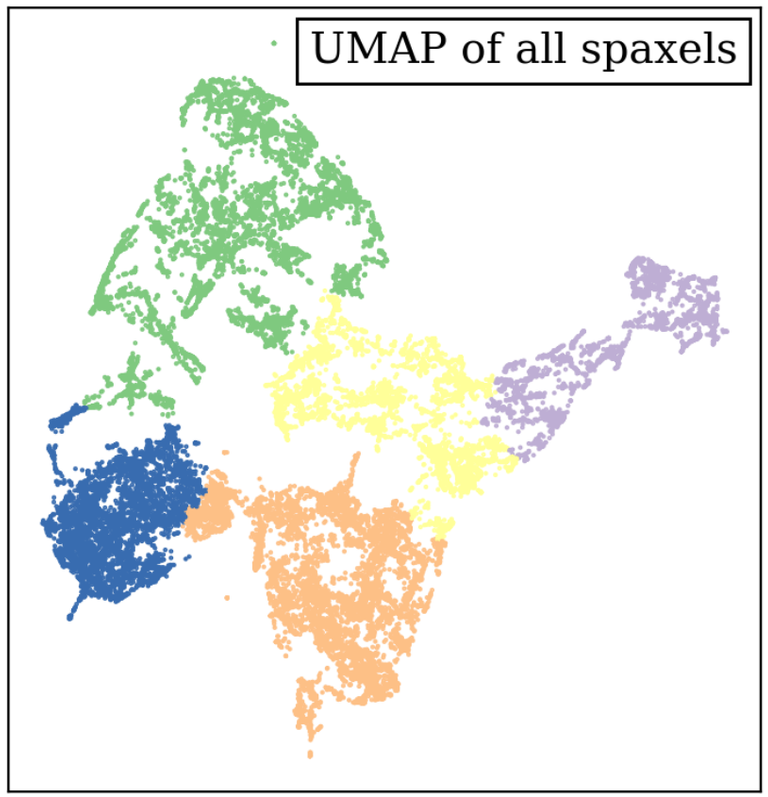
Interpreting the resulting embedding - part 1

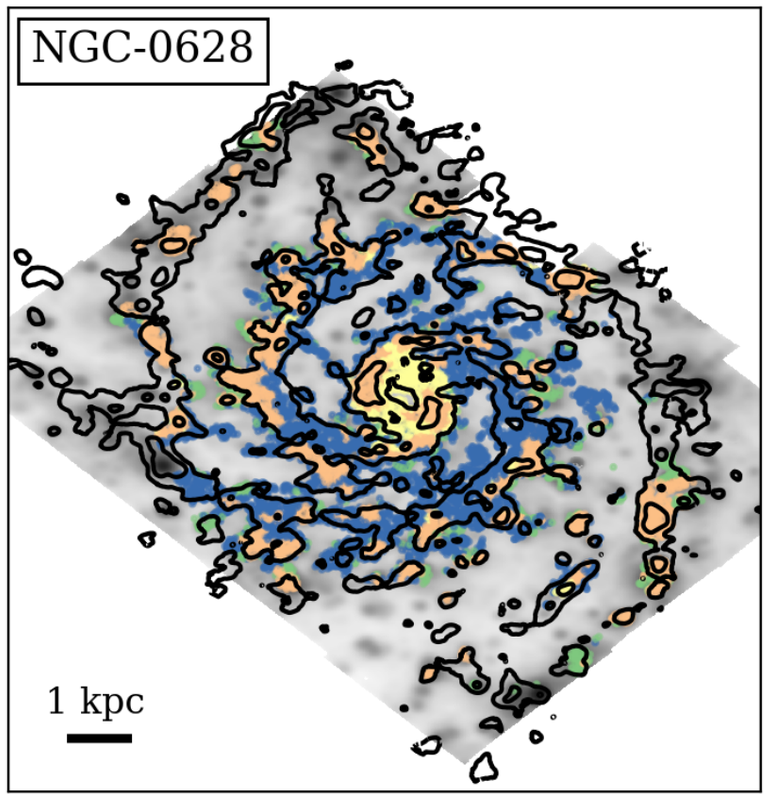


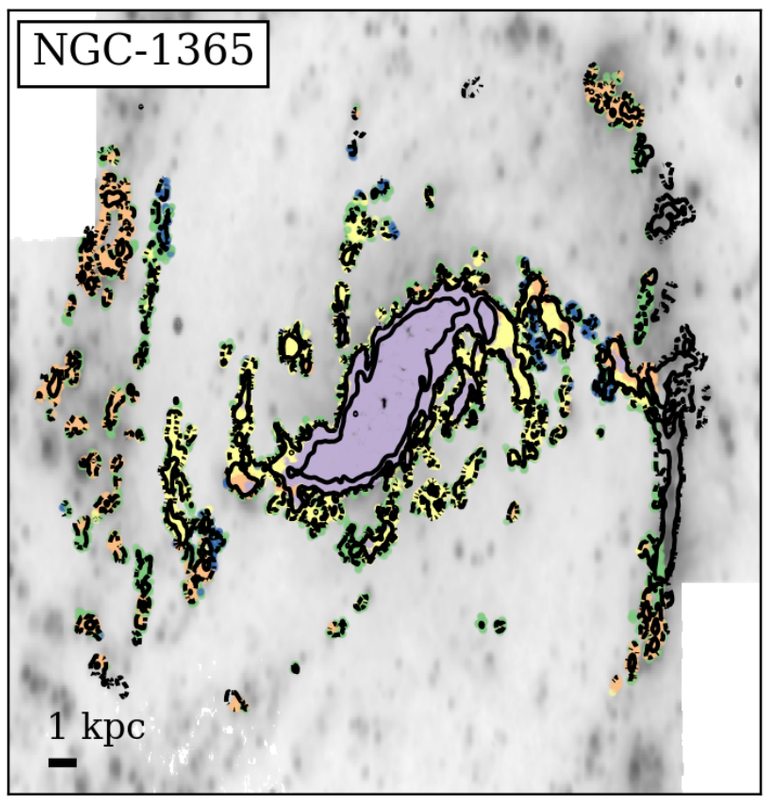
Exploring different clustering algorithms

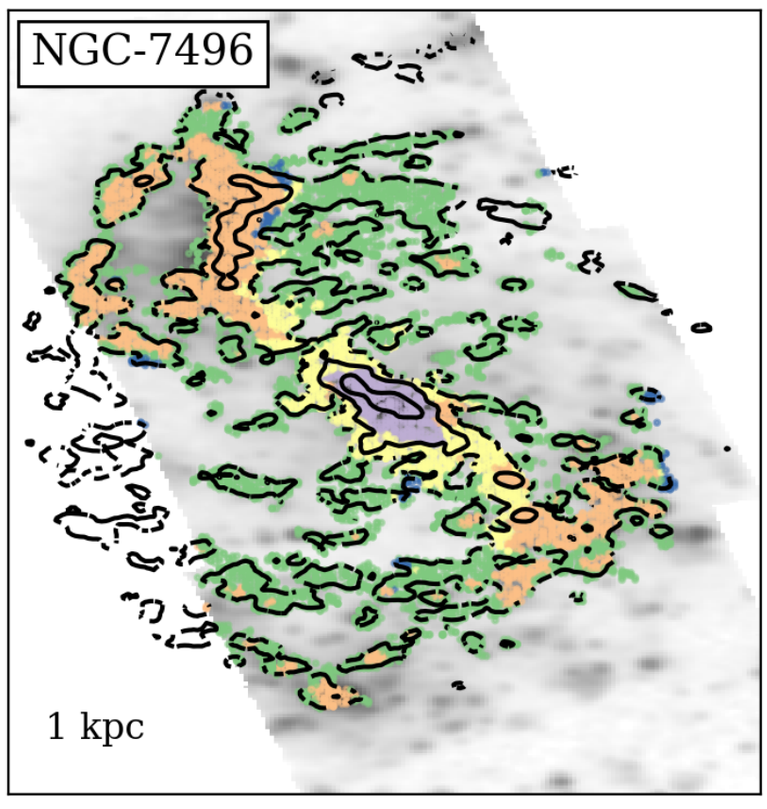


Interpreting the resulting embedding - part 2

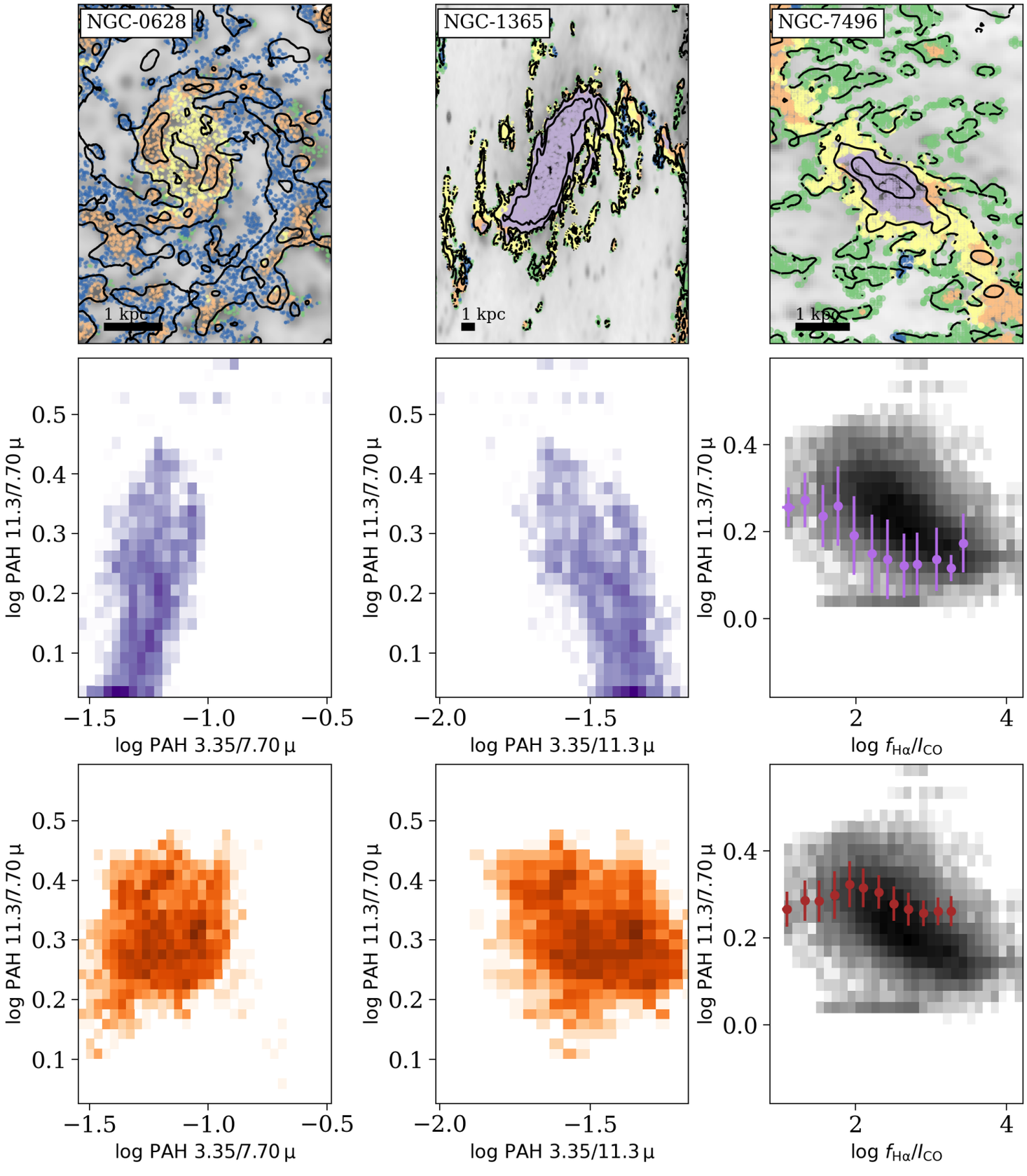




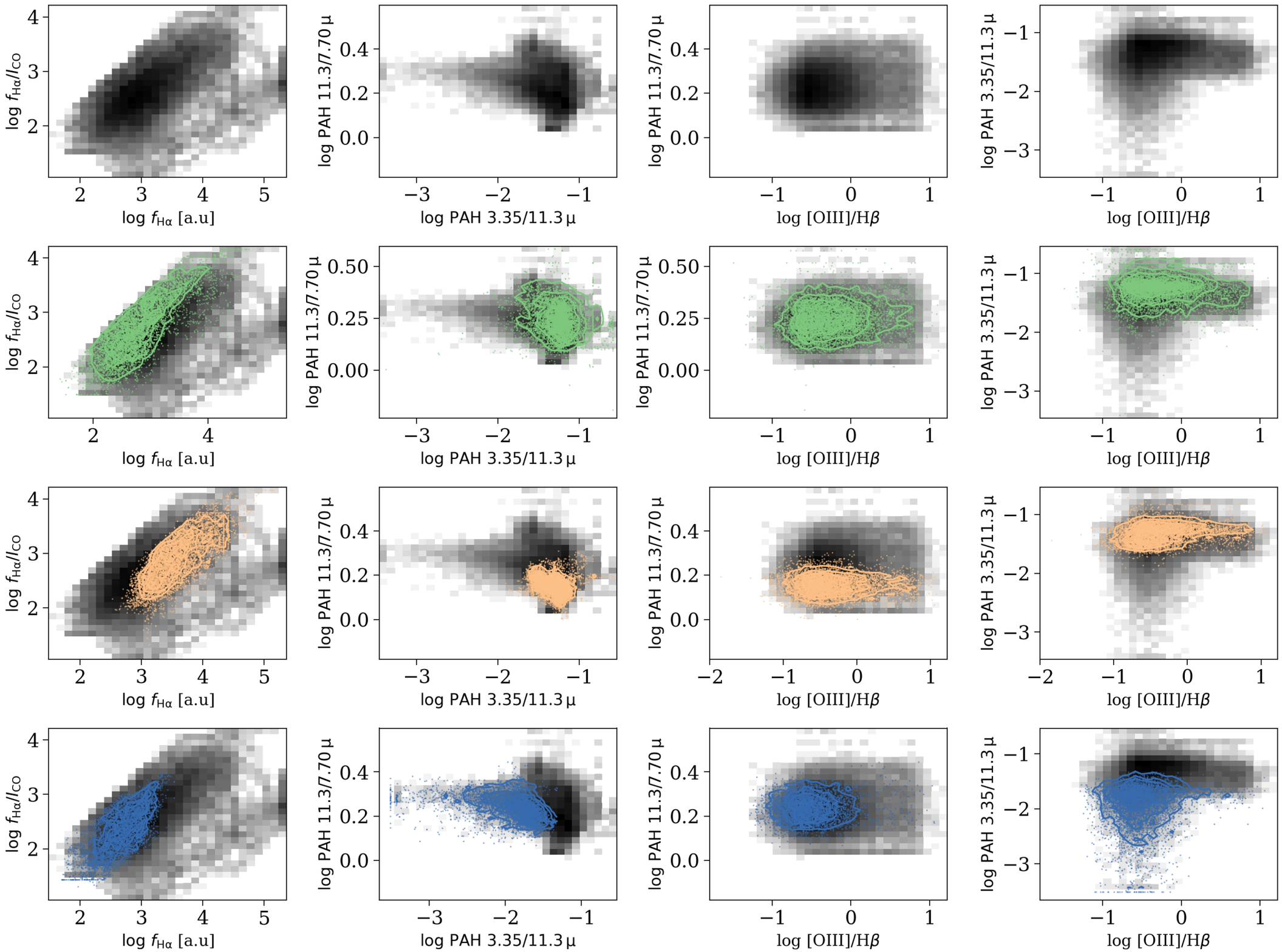




Interpreting the resulting embedding - part 3



Interpreting the resulting embedding - part 3



Next steps

Two main directions to go to from here:

1. Include upper limits and Nan values in the analysis.

2. Slowly transition from derived features to the raw data (e.g., project 12!)