IRAM, K. Zacher

Estimating the H2 column density using combined molecular line intensities in the Orion B cloud

Moritz Itzerott (Institute for Physics and Astronomy, University of Potsdam, DE)
Da Eun Kang (Heidelberg University, DE)
Ashley Lieber (Nearby Galaxies Lab, University of Kansas, USA)
Parit Mehta (Institute for Astrophysics, University of Cologne, DE)
Ekaterina Mikheeva (Moscow State University, Astro Space Center of Lebedev Physics Institute, RU)
Léontine Ségal (Institut de Radio Astronomie Millimetrique (IRAM), IM2NP, FR)

Antoine Zakardjian & **Jérôme Pety**

Science Question

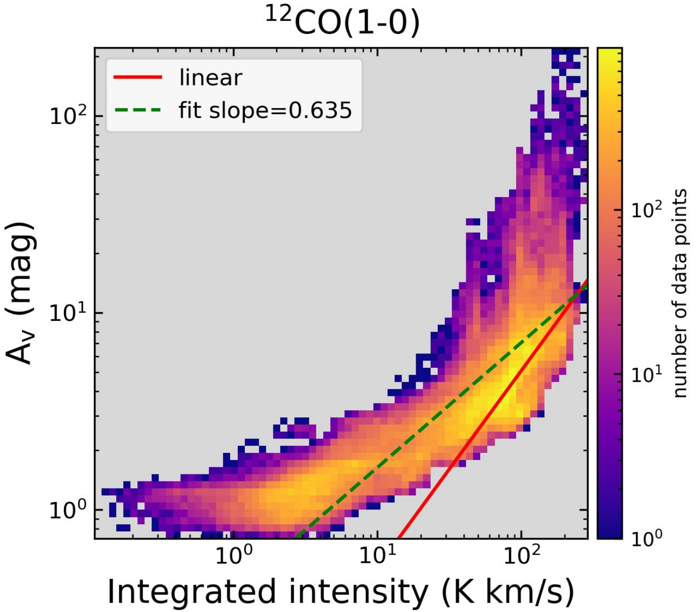
Why do we study Giant Molecular Clouds (GMCs) ?
To investigate the process of star formation in our galaxy and elsewhere

How?

- Cold H2 is invisible → **dust** and/or **molecular lines** (CO, HCO+, … isotopologues)
    - **dust** : optically thin, but for **SED** need **FIR** (low angular resolution), lacks velocity information
    - **rotational lines** : velocity information, achieve high angular resolution from ground (e.g., X_{CO})

Can we go beyond X_{CO} and estimate N_{H2} with more lines and machine learning techniques?

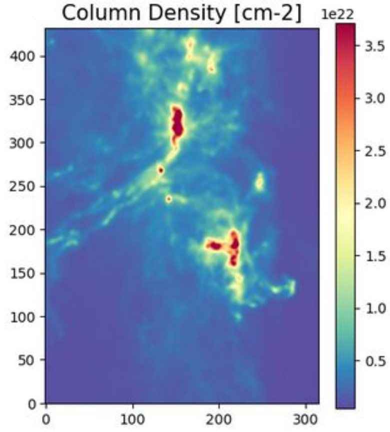
IRAM 30m Telescope , K. Zacher

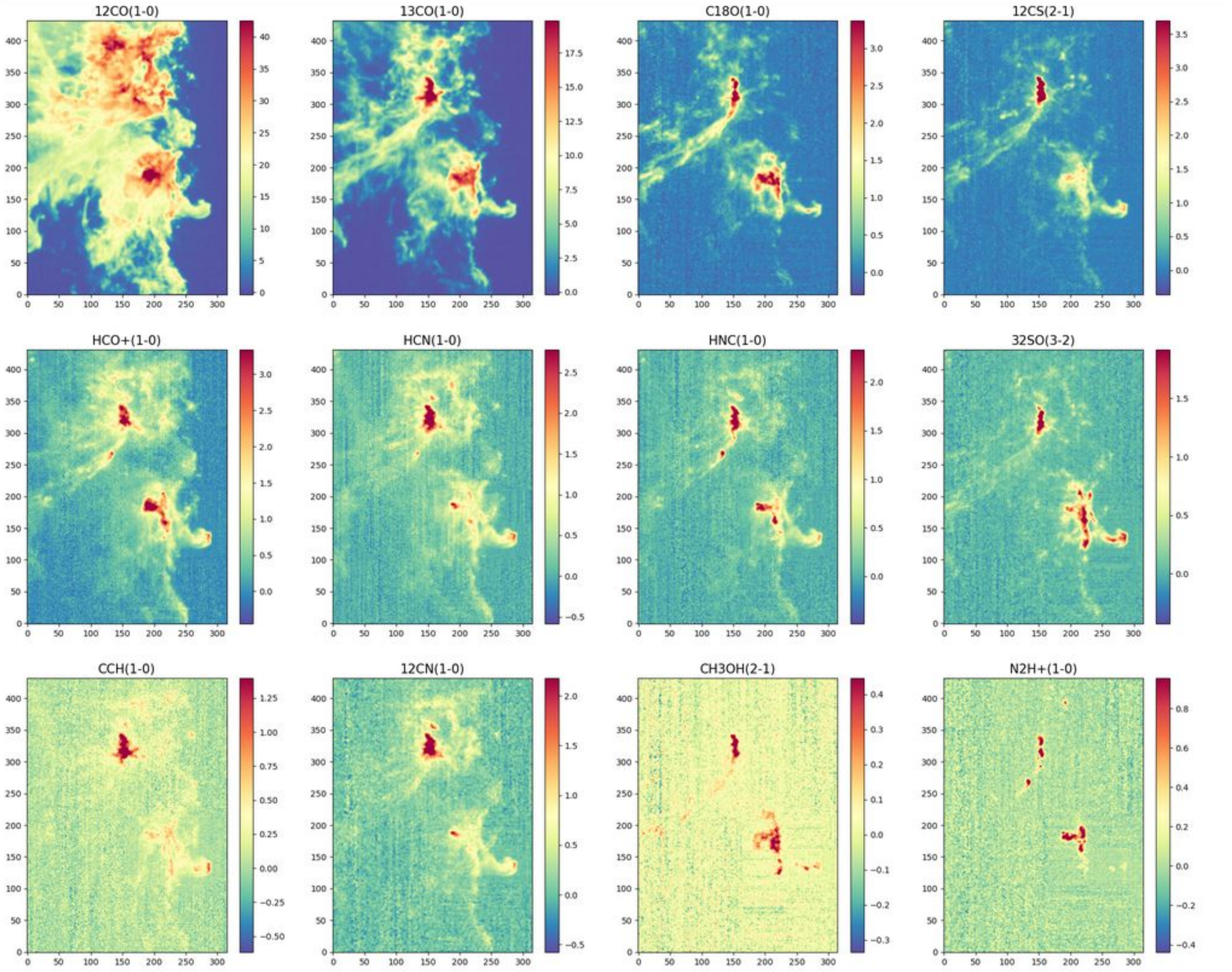


The ORION-B Dataset

IRAM 30-m Telescope

- Resolution 0.07 pc/px
- Image Size is 5x7 pc
- 12 Millimeter Rotational Transition Lines





Correlation between different line intensities

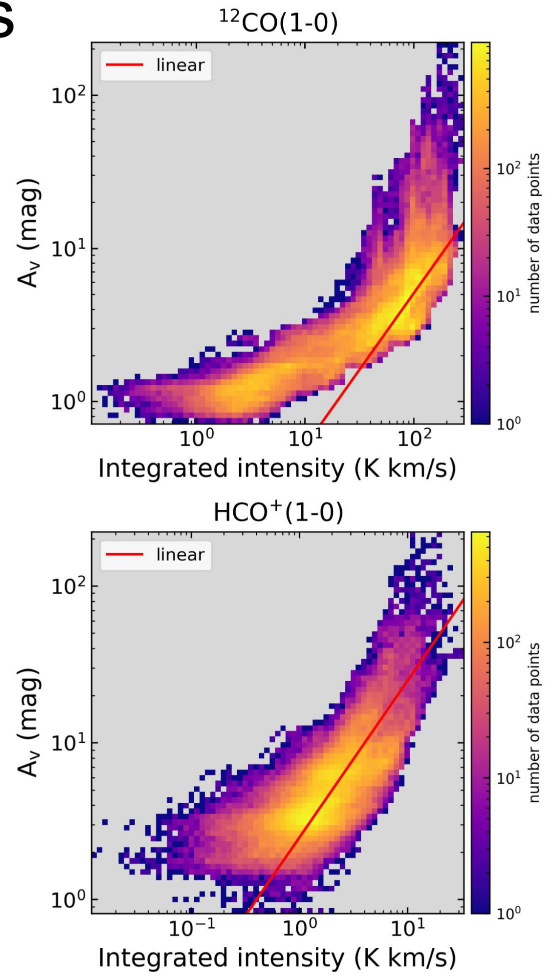
Not only CO molecular-line intensity but also other molecular-line intensities have relations to N(H_2)

We want to explore

- if different lines correlate to each other
- if the combination of lines has a connection to N(H_2)

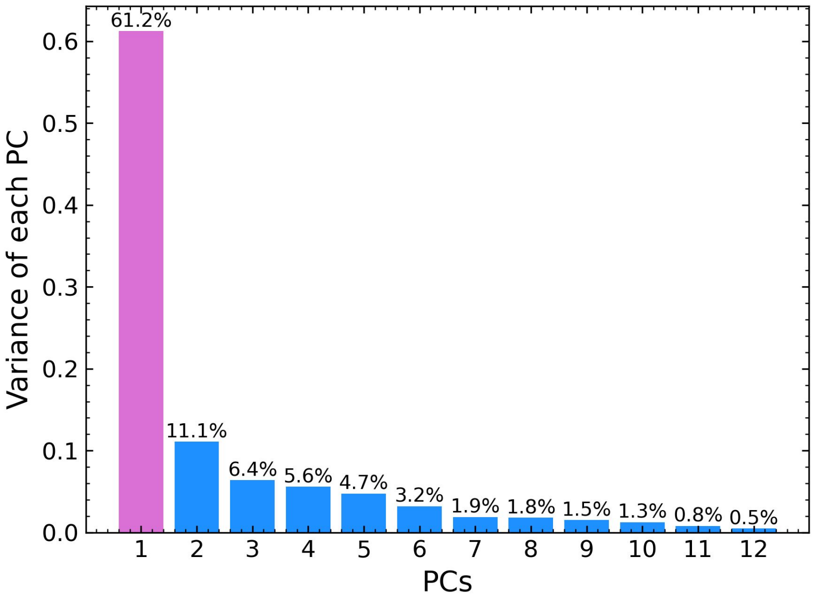
by performing Principal Component Analysis (PCA) on the data of all 12 line intensities

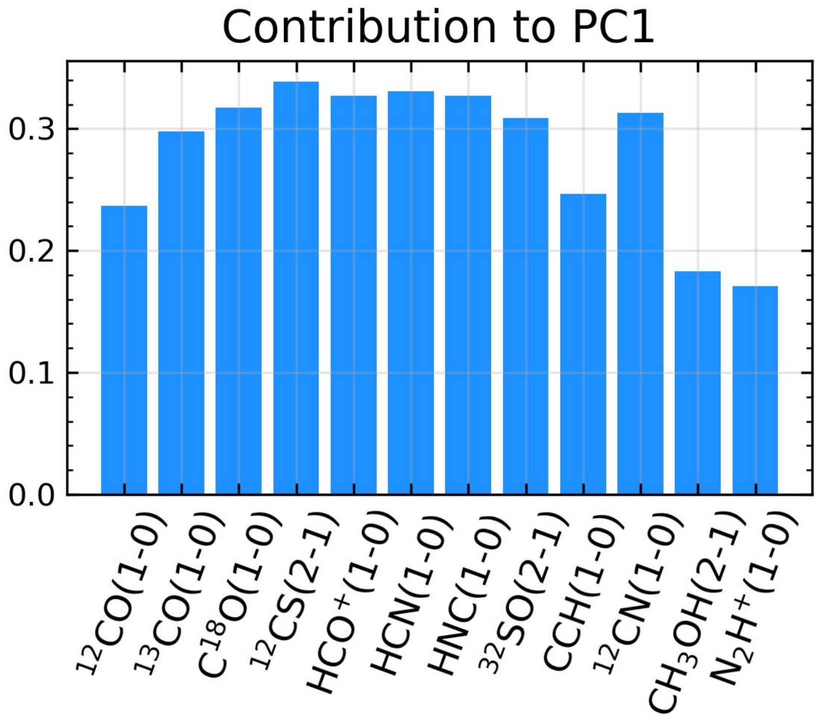
- find linear correlations in the data
- useful for high dimensional data space (e.g., 12 lines)



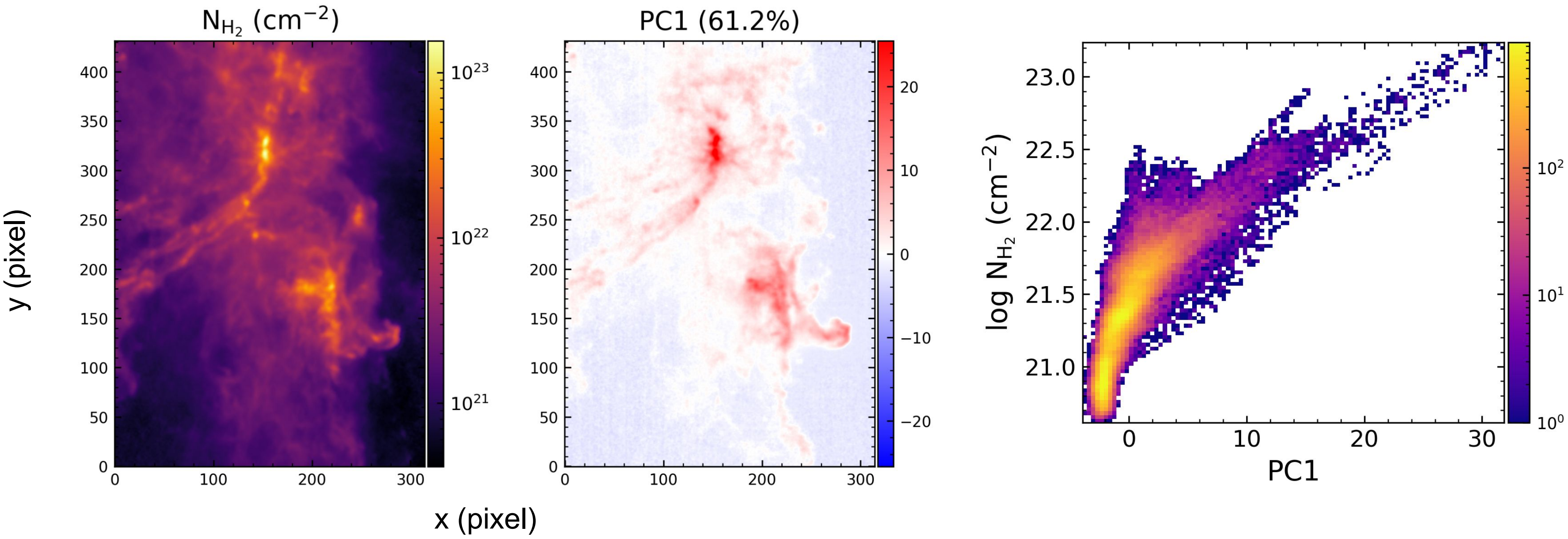
Which component is the most important?

PCA finds intrinsic characteristics from the data (PC-space) and shows which characteristics are more useful or less useful to explain the distribution of data.



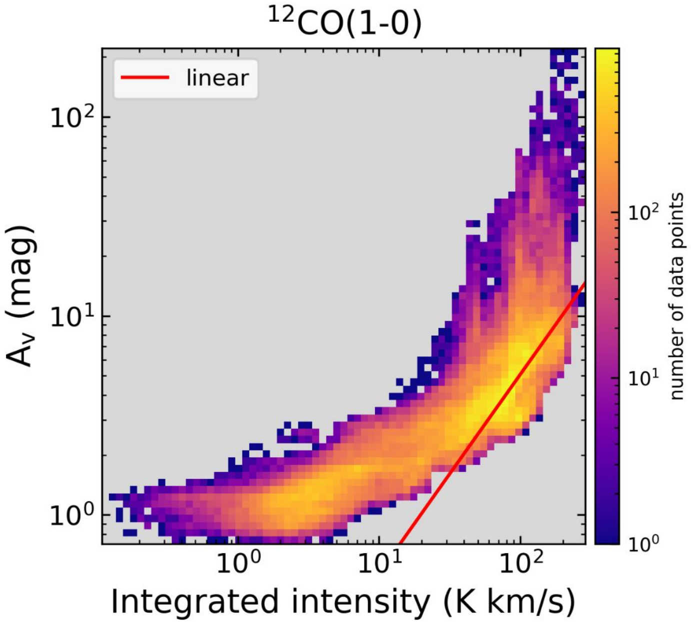


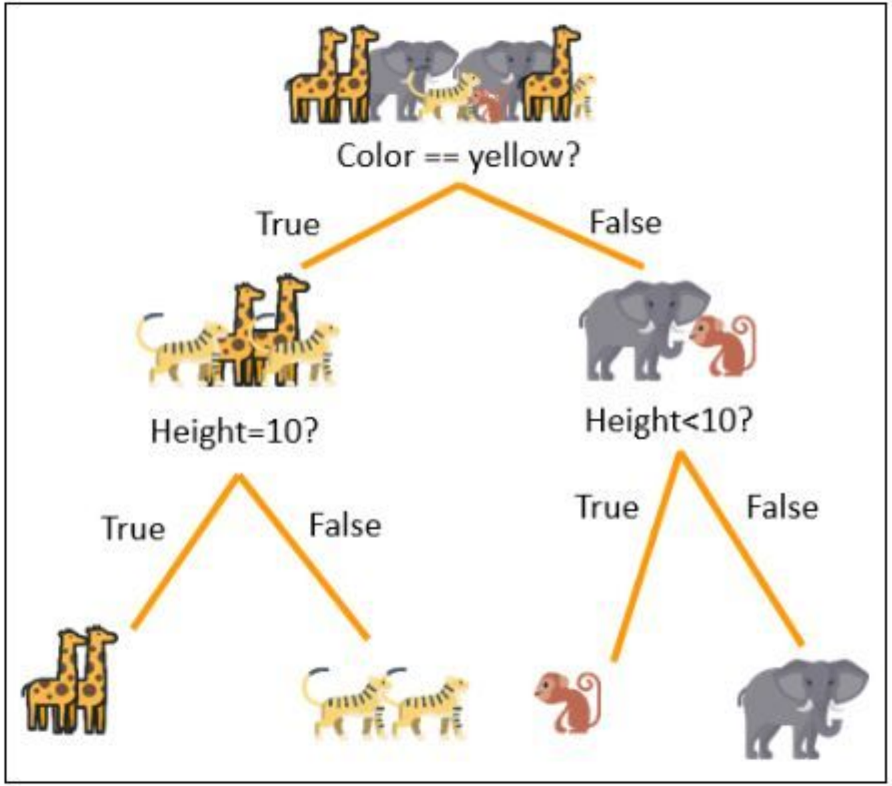
Relation between PC1 and $N(H_2)$



- All line intensities are related to $N(H_2)$
- Not a perfect linear correlation.
- We need a non-linear function to link the line intensities and $N(H_2)$

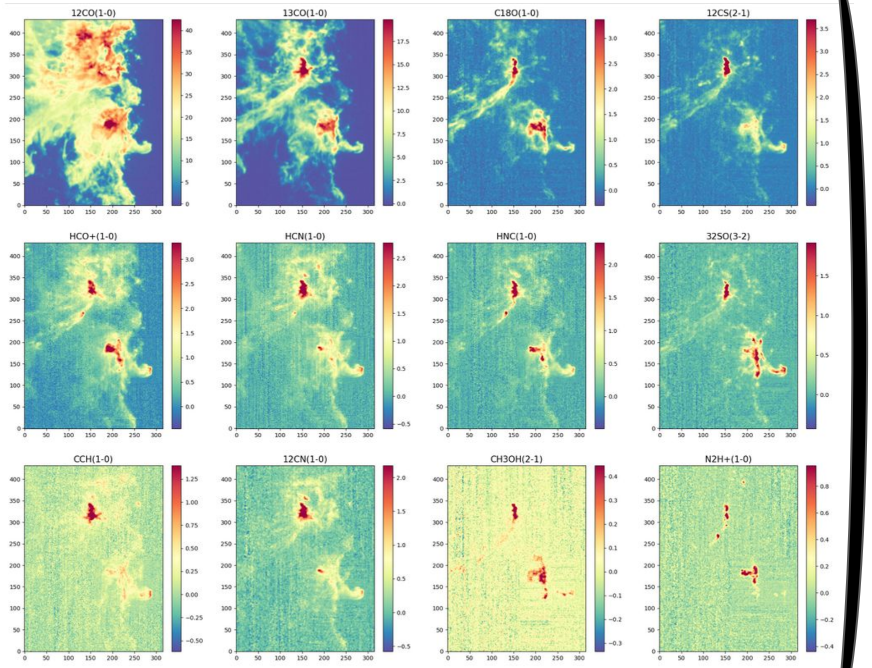
How do we “learn” $N(H_2)$ from line intensities?



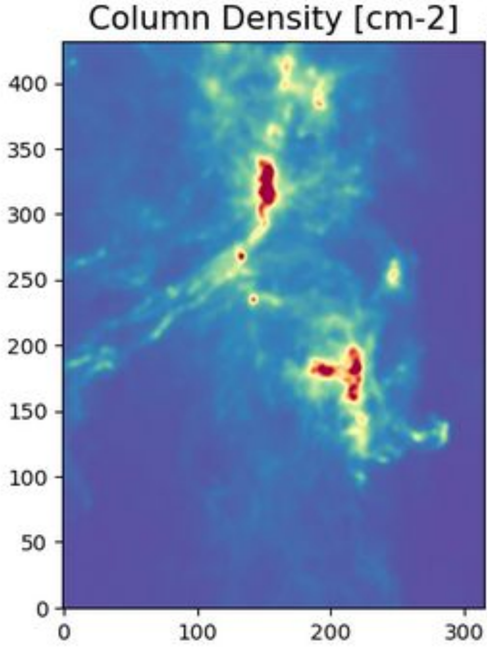


How do we “learn” $N(H_2)$ from line intensities?

f

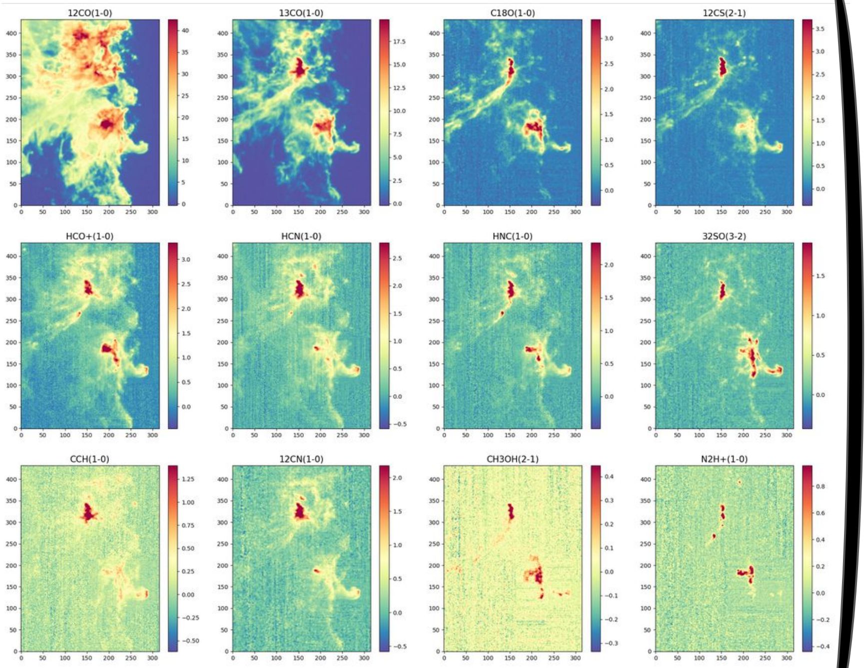


=

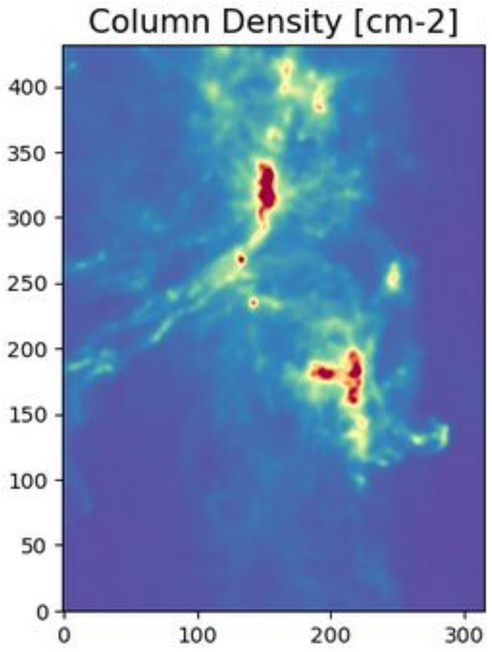


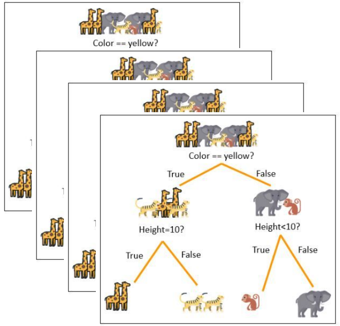
How do we "learn" $N(H_2)$ from line intensities?

f (



) =





Use Random Forests!

- ★ Collection of decision trees
- ★ Good for non-linearly correlated data,
- ★ No need to normalize data, remove blank or missing values etc.
- ★ Randomizes the input sample and averages the result

How is the dataset prepared?

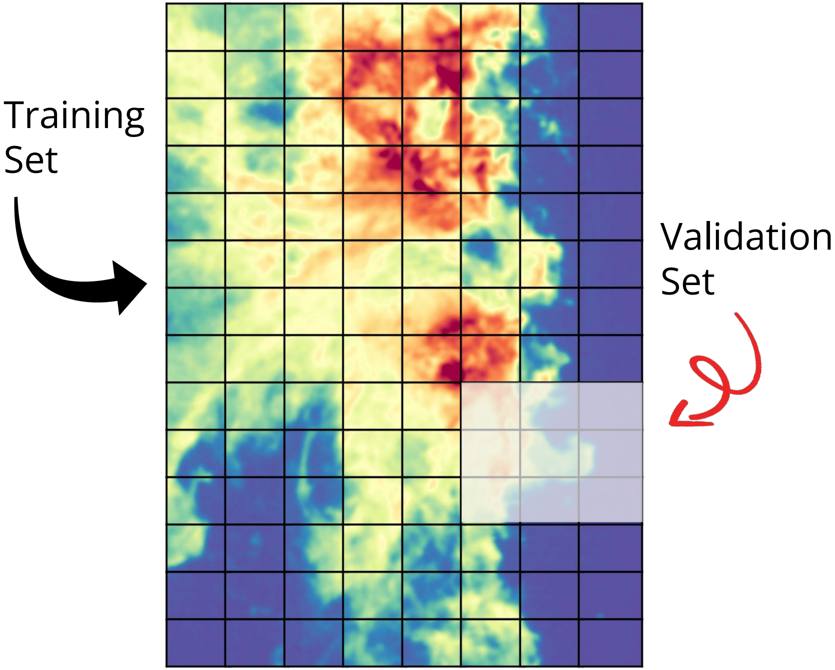
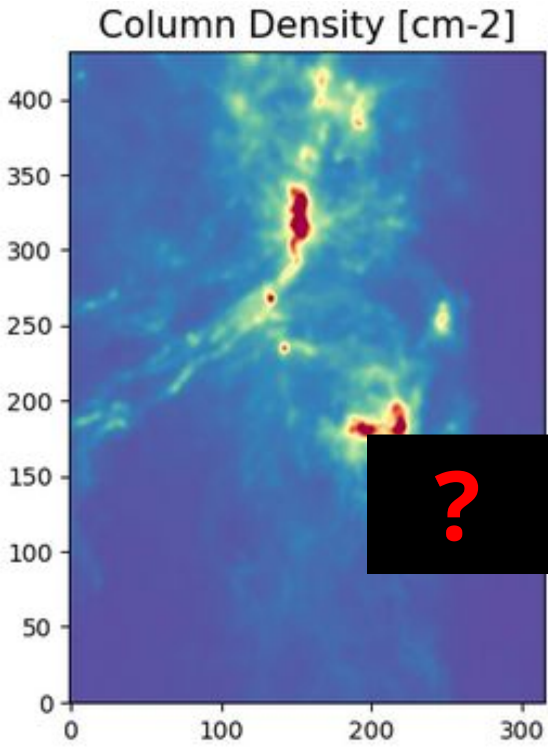


Illustration: Splitting the

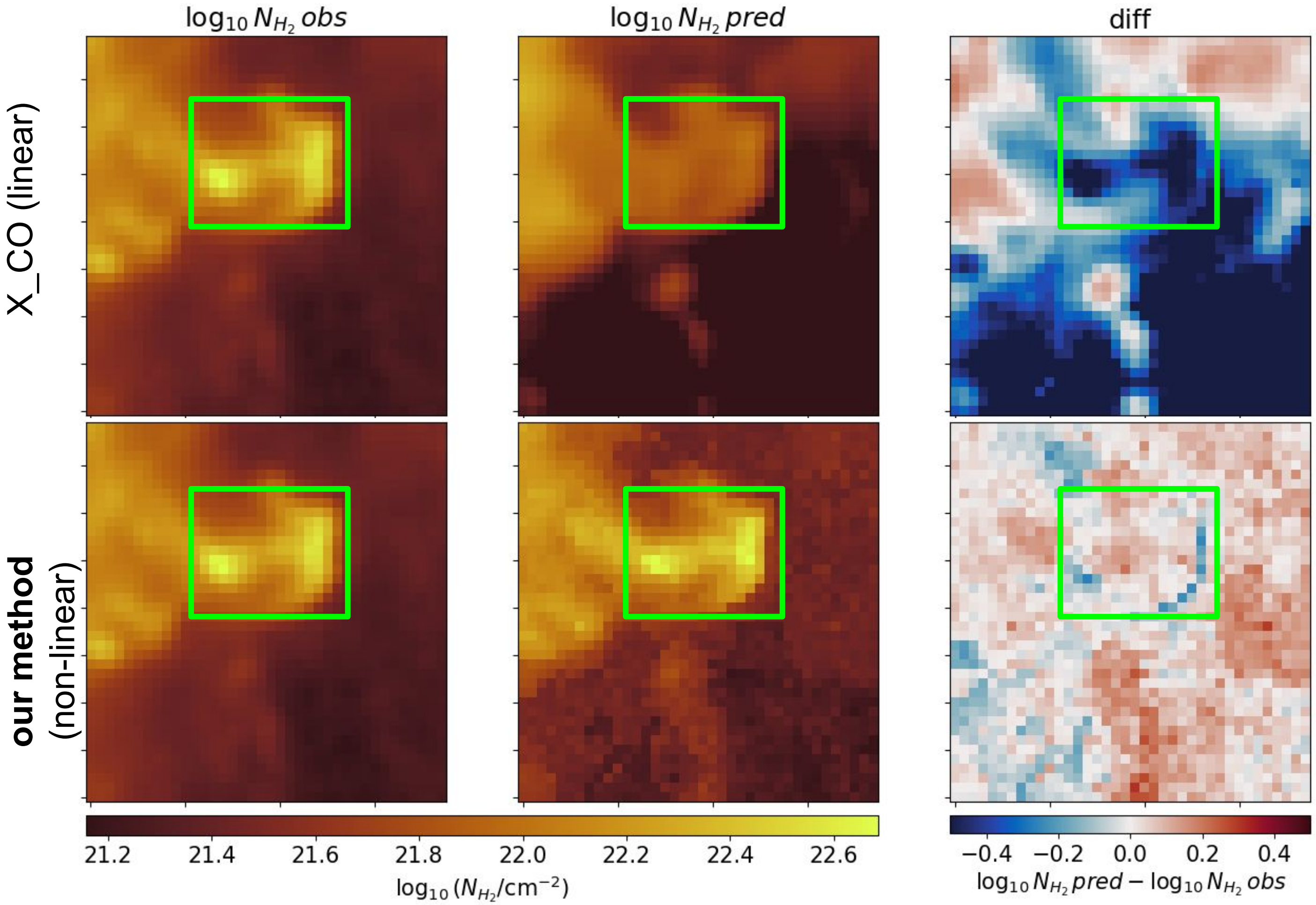
^{12}CO (1-0) dataset



- ★ Split the correlation data into Training and Validation sets
- ★ The Random Forest regression algorithm **predicts the column densities of the validation region**

Validation

By Ken Crawford, CC BY-SA 3.0



Summary and Perspectives 🛤️

What have we done? 😱

★ Investigate the relation between molecular lines and the $N(H_2)$

★ Perform PCA to find correlations between lines

★ Perform RF to link molecular lines to $N(H_2)$ non-linearly

What did we find? 🕵️‍♀️

★ Our RF model works better than simply using X_{CO}

What will we do next? 🚀

★ Test the model robustness on **other data**

- well known GMCs → check on the interpretability of the results

★ Apply the model to **new data**

PCA methodology

★ *unsupervised learning*

★ *transforms the data to a new coordinate system*

★ *the **greatest variance** by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.*