Colloque du Programme Physique Chimie du Milieu Interstellaire 2024

Complex Organic Molecules detected in hot cores with Machine Learning

Nina Kessler

Advisers : T. Csengeri, S. Bontemps

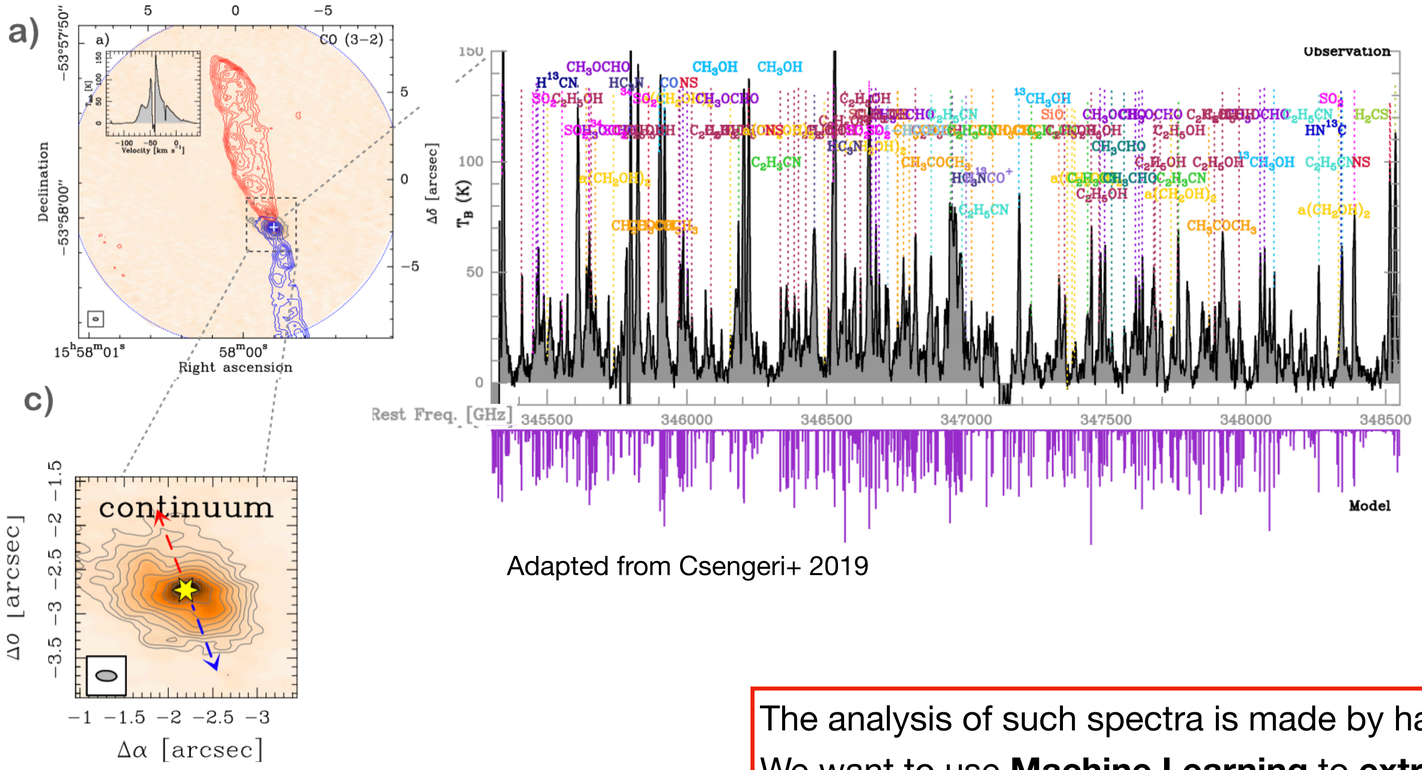
External consultant : D. Cornu (LERMA)

Complex organic molecules (COMs) are detected in chemically **rich sources**

Improvement of interferometers = **larger bandwidth** at **high spectral resolution**.

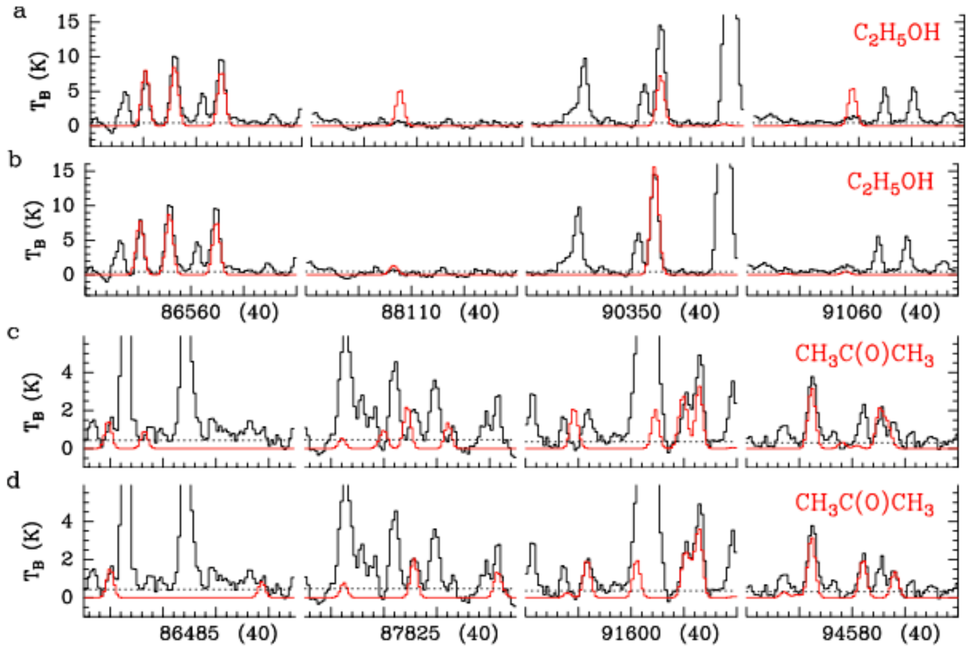
Identification of molecular species in the ISM : now > 300

Spectra of hot cores are **rich in molecular emission**, and present a **high abundance of COMs** (Belloche+ 2014, Csengeri+ 2019).



Adapted from Csengeri+ 2019

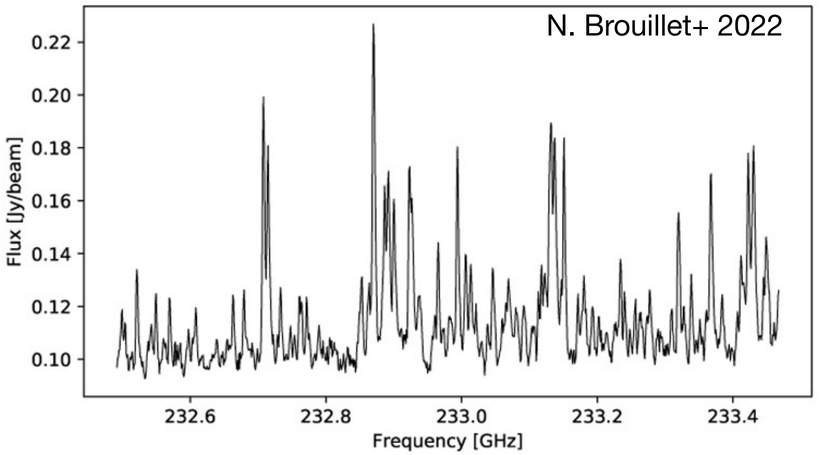
COMs → LTE models to fit the emission lines



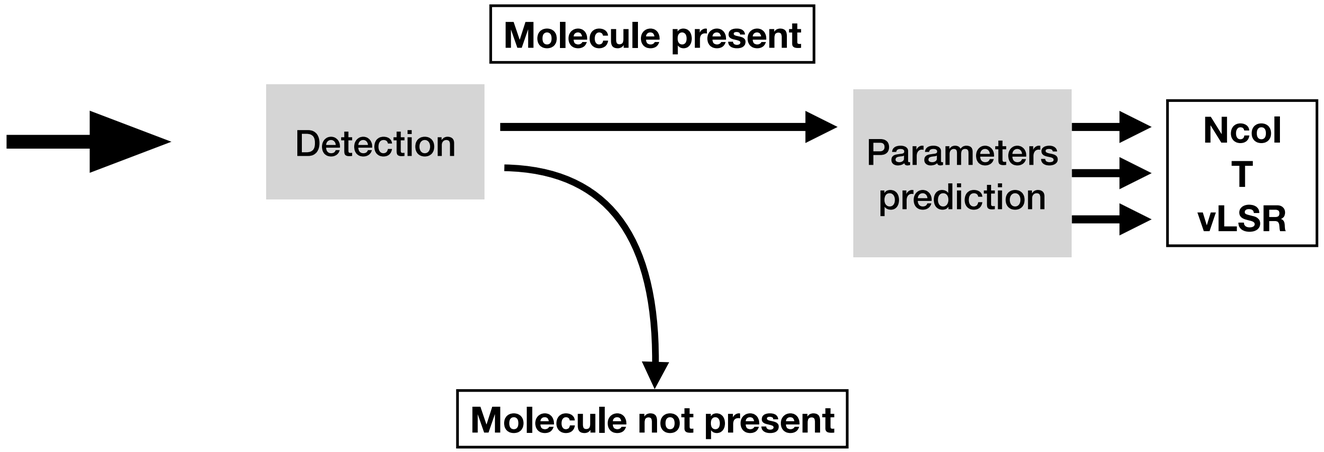
Synthetic (red) and observed (black) spectra of ethanol (a, b) and acetone (c, d) Jorgensen+(2020).

The analysis of such spectra is made by hand which is not optimal.
We want to use **Machine Learning** to **extract** and **process** spectroscopic **data**.

Goal : automatically **extract information** from a spectrum



Input : Data (usually few Tb) with 1D spectra of $10^4 - 10^6$ channels



1. **Representative** training data set
2. Suitable **Artificial Neural Network**
3. Evaluation of the **model's performance**

Classification :

- Detection one by one
- **Multi-label classification**

Regression :

- Regression one by one
- Multioutput regression
- Multibranch regression

There is currently **not enough fully analyzed (sub)millimeter spectra** to built a reliable **training dataset**; the **spectral emission** of hot cores must be **modeled**.

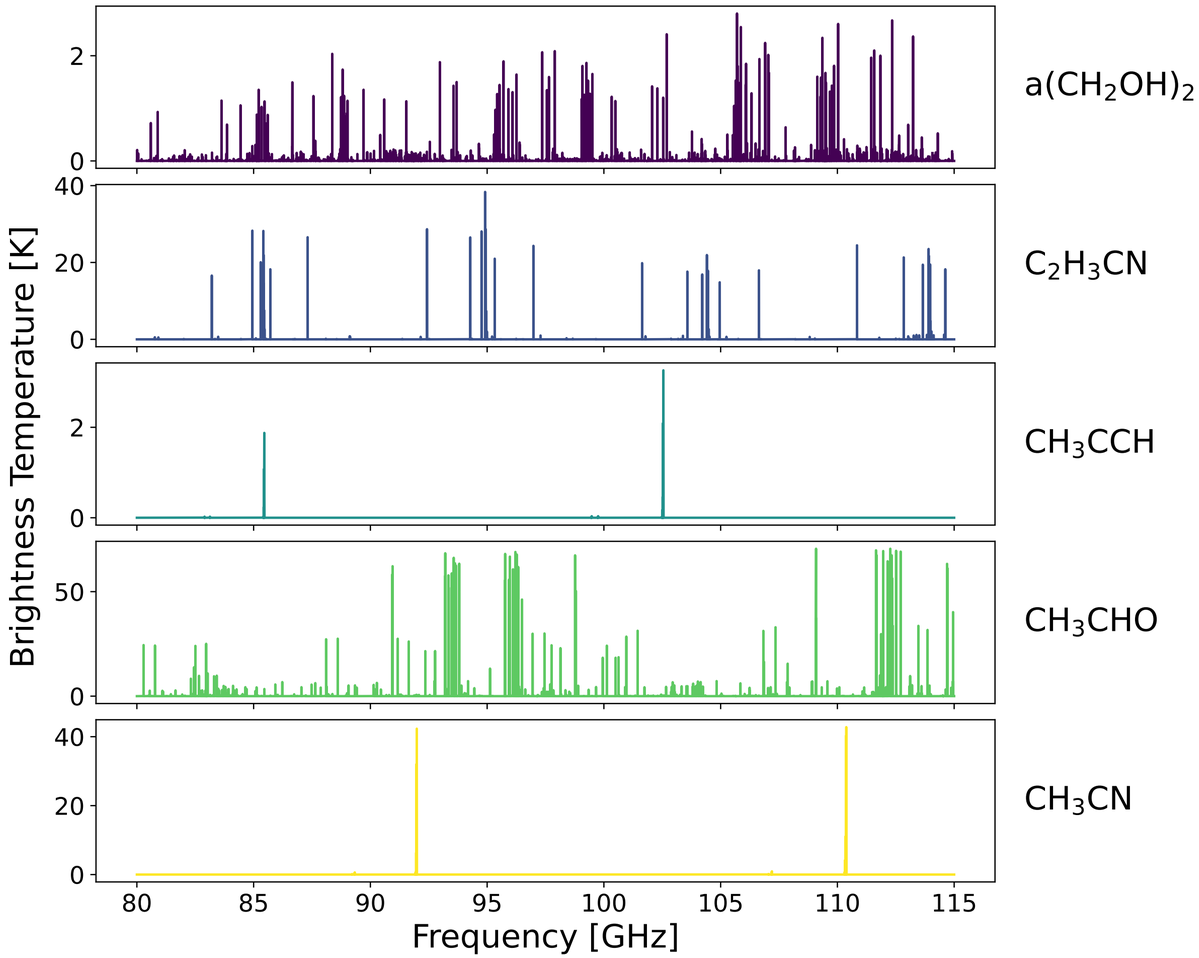
→ Collection of **100k LTE models** from **20 molecules** (+some isotopologues) on **a 80-115 GHz bandwidth** (1MHz resolution).

Browsing the space of parameters defined by :

Parameter	Range
Beam size [arcsec]	3
Column density [cm^{-2}]	10^{12} - 10^{19}
Line width FWHM [$km.s^{-1}$]	1 - 12
Rotational temperature [K]	30 - 300
Source size [arcsec]	0.15 - 15

+ A model used as **a mask to hide the transitions** potentially impacted by **local physics** with the species **:**

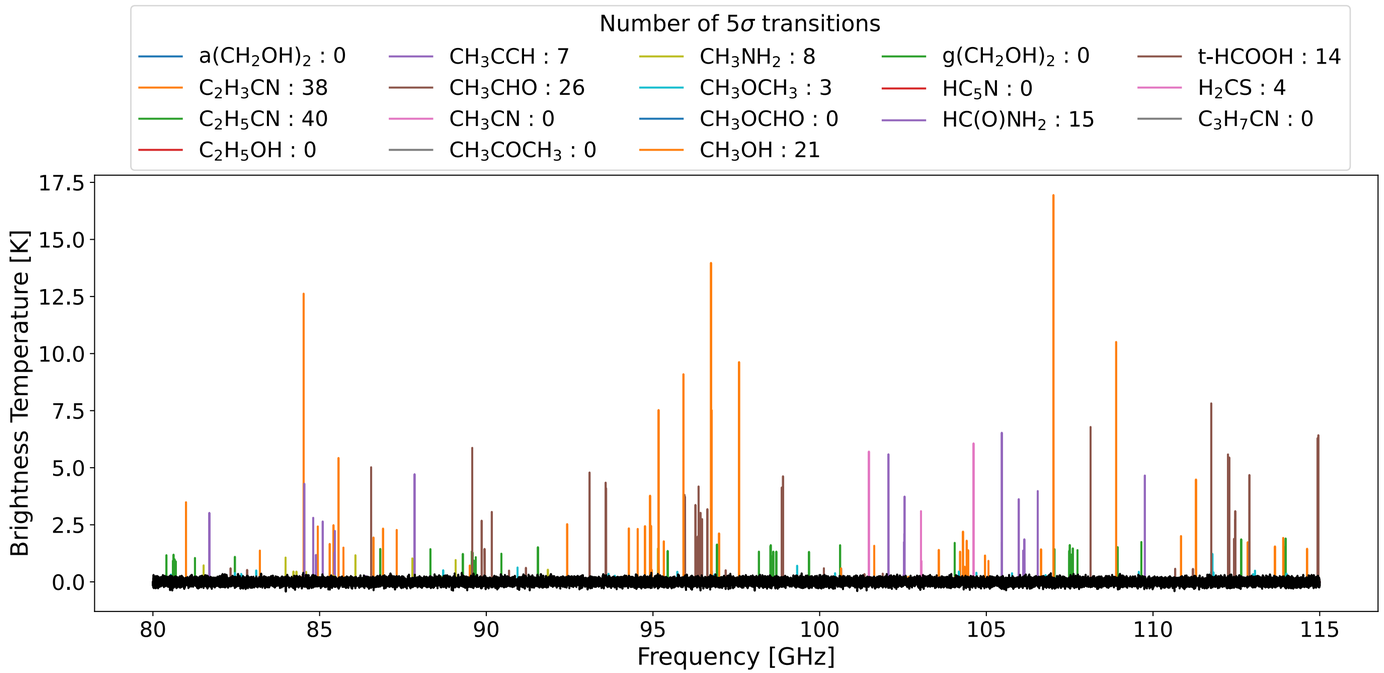
Carbon chains	O-bearing	N-bearing	S-bearing
CCH (6)	$CO^{\dagger}$ (3)	CN (9)	CS (1)
	HCO^{+} (4)	HCN (1)	HCS^{+} (1)
	H_2CO (21)	HNC (1)	H_2S (4)
	H_2CCO (40)	HNCO (50)	OCS (5)
	SiO (1)	N_2H^{+} (1)	SO (6)
		NH_2OH (18)	SO_2 (110)

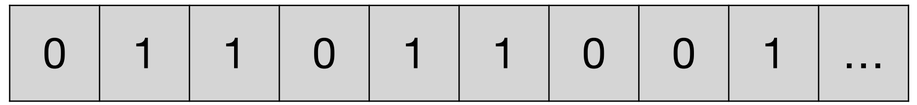


The **training dataset** needs to be **representative** of the conditions we observe in star forming regions.

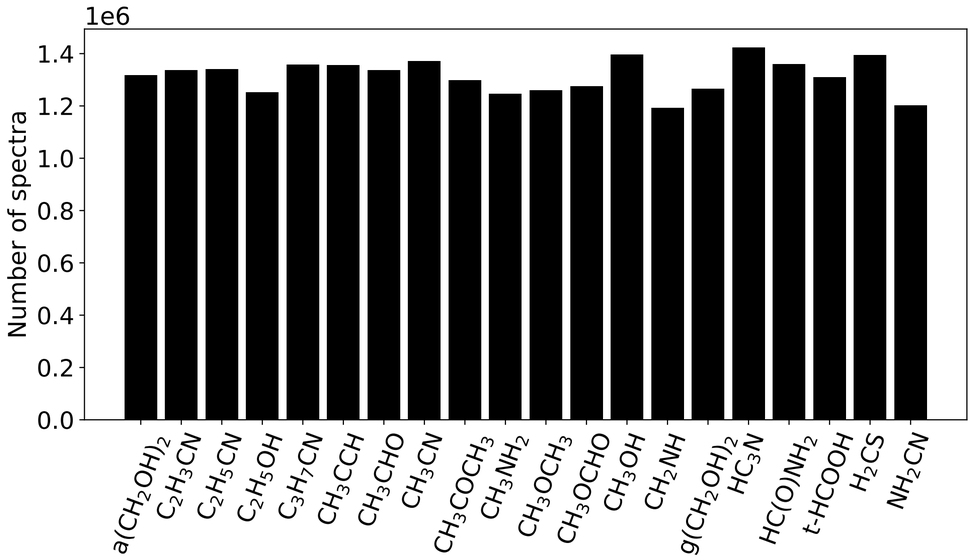
Training set :

- Representative complexity
- Random combination of spectra
- Noise (affects the detectability)
- Artefacts

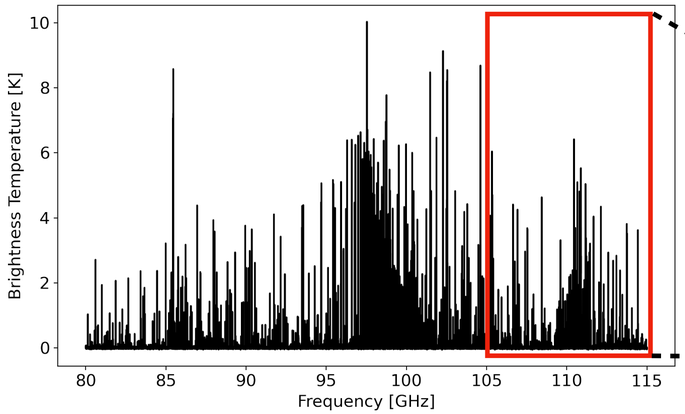




The goal is to have a homogeneous representativity of all the species within the training dataset.

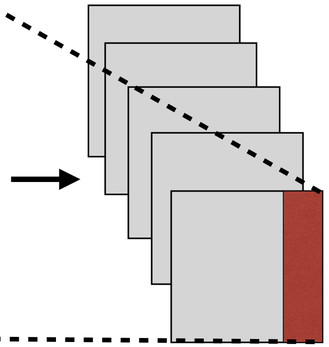


Convolutional Neural Networks are able to detect and identify molecules within a spectrum**.**



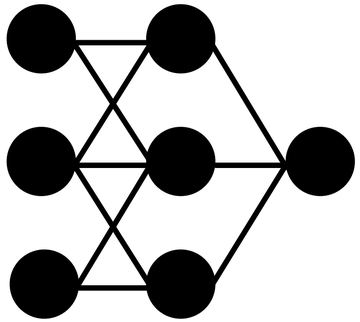
Input spectra

Features extraction



Convolutional layers

Classification



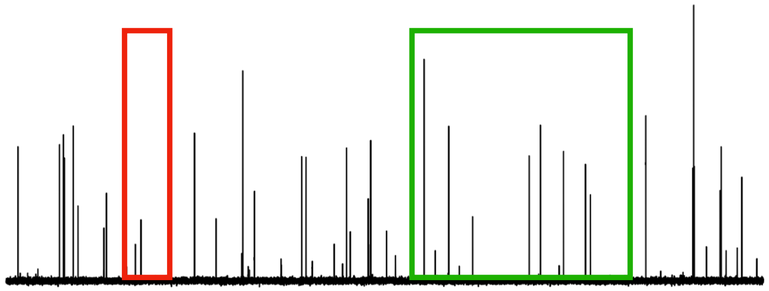
Dense layers

Model Score

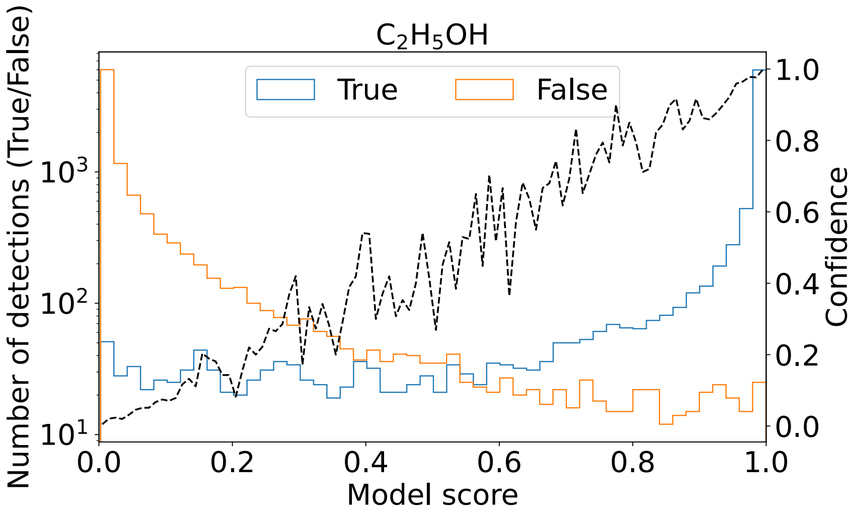
aEG : (5.95 ± 2.77) %
C_2H_3CN : (96.77 ± 2.83) %
C_2H_5CN : (98.06 ± 1.89) %
C_2H_5OH : (23.17 ± 11.11) %
CH_3CCH : (97.19 ± 2.06) %
CH_3CHO : (72.35 ± 13.74) %
CH_3CN : (89.64 ± 6.24) %
CH_3COCH_3 : (19.9 ± 8.88) %
…

Small receptive field, e.g. focus on the shape of lines

Large receptive field, e.g. focus on the distance between 2 lines

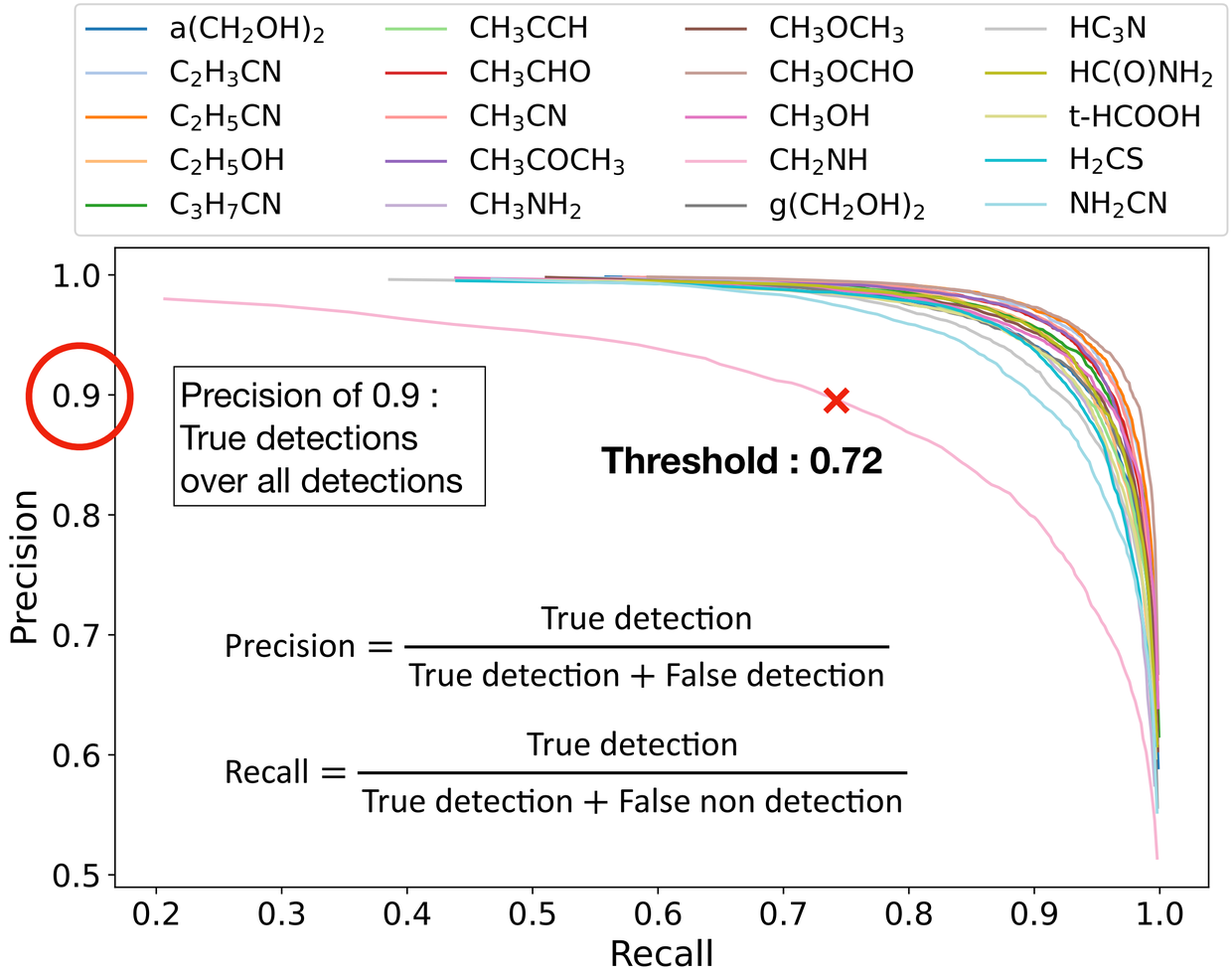


35k channels

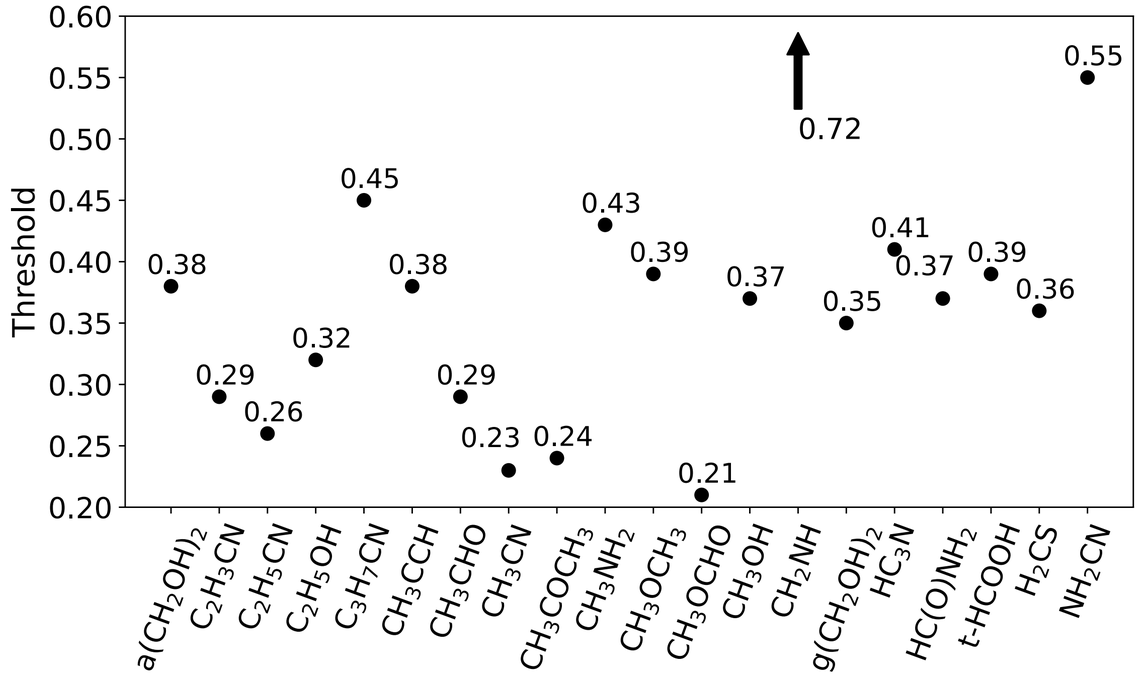


The **calibration** of the model score is based on the **statical answer** of the model on a dataset.

The Receiver Operating Characteristic (ROC) curves unveil the **performance** of a classifier model depending on the **precision and recall at varying confidence threshold**.

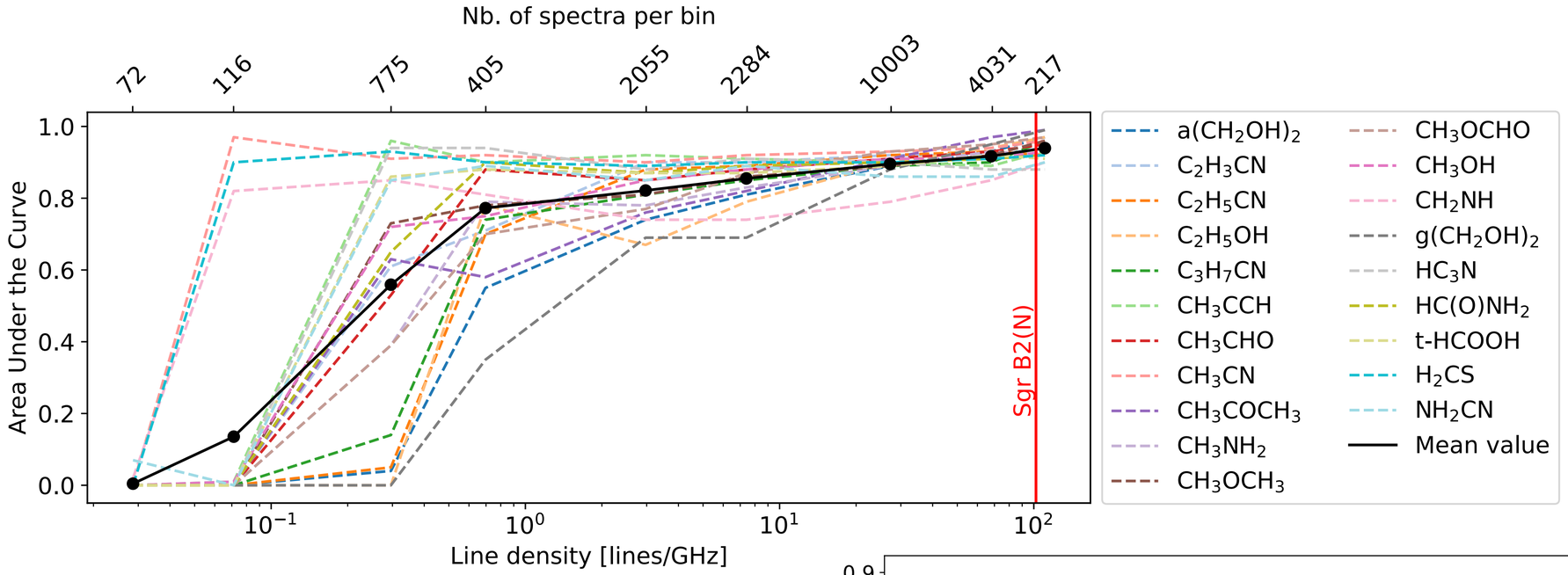


They allow to determine a **model score threshold** depending on a desired precision **to consider a molecule as detected.**



This threshold is arbitrary. It is model and molecule dependent !

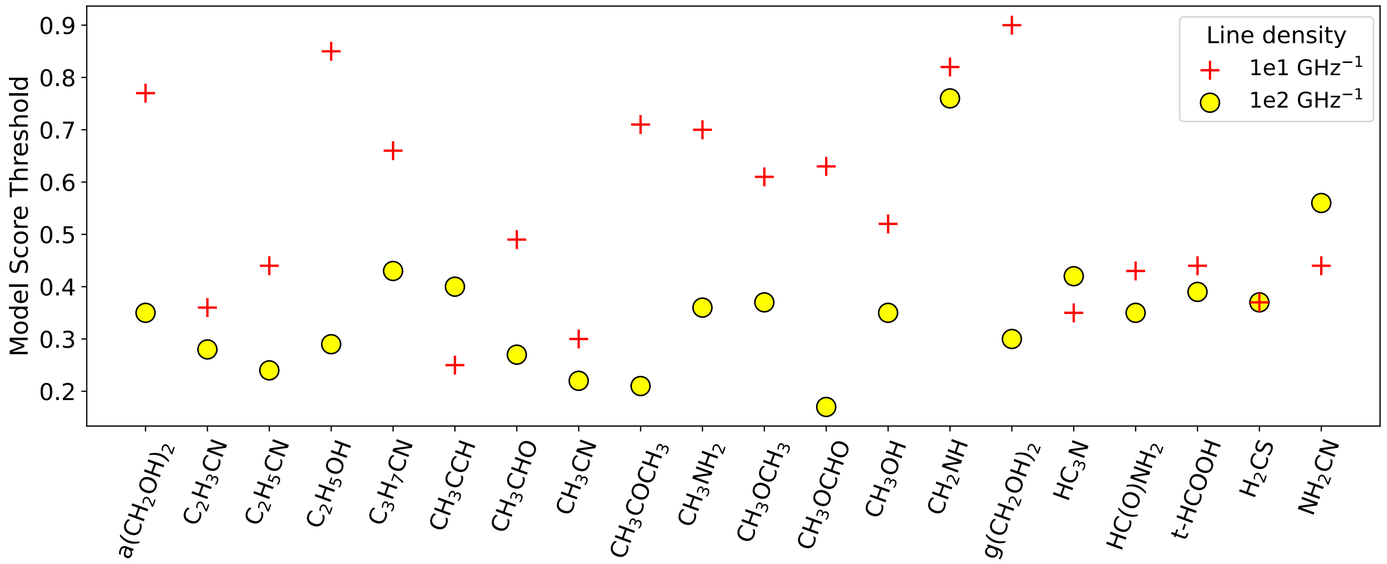
The model presents **different performance** in function of **spectral line density**.



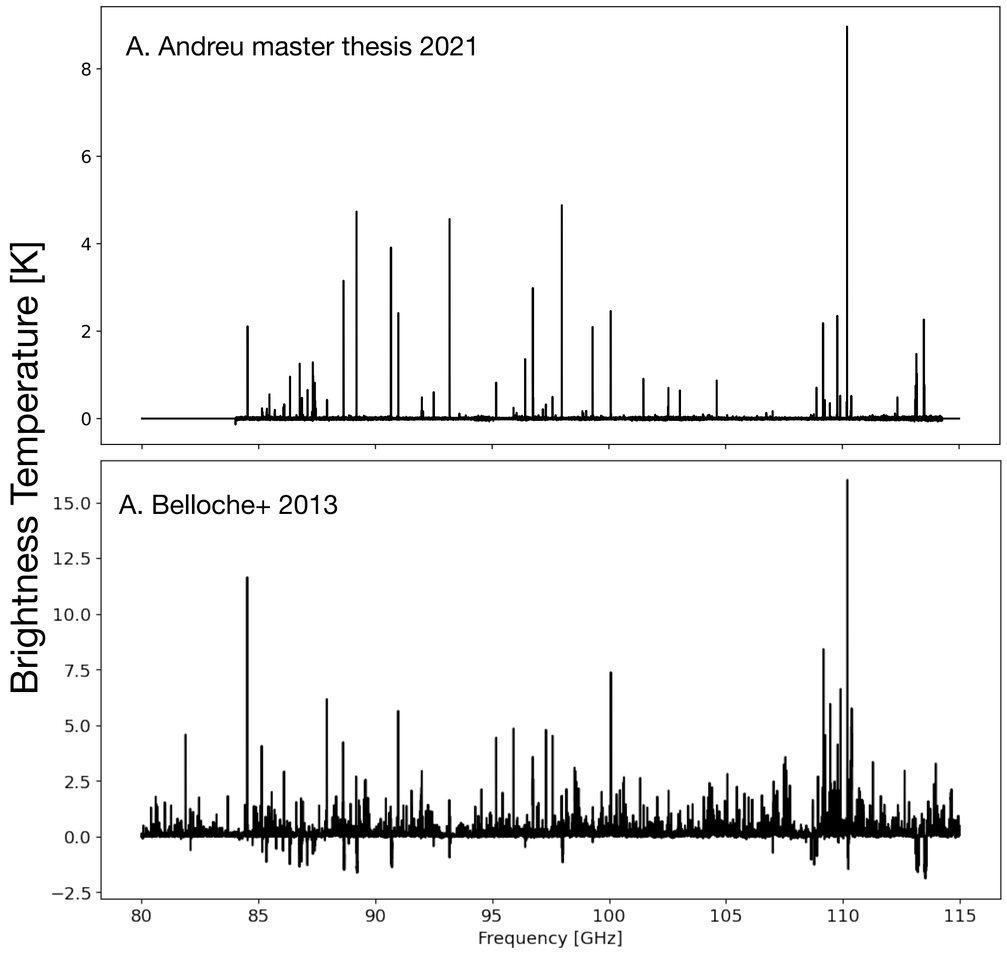
The model presents a better performance at **low line densities** for small molecules, and at **large line densities** for complex molecules.

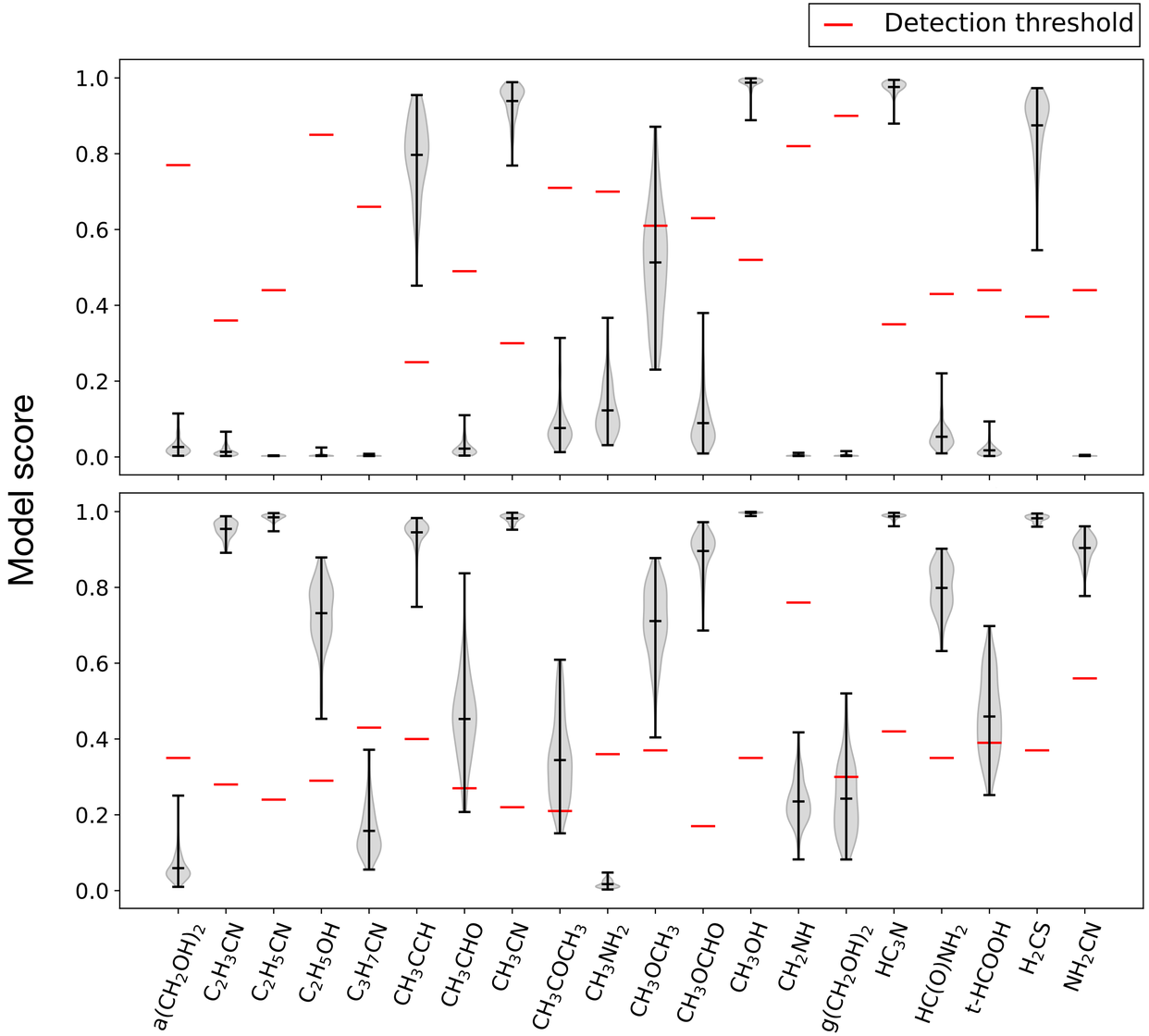
It best performs for spectra having line densities between **10 and 100 lines per GHz**.

Regimes can be identified to **adapt the detection threshold** accordingly.



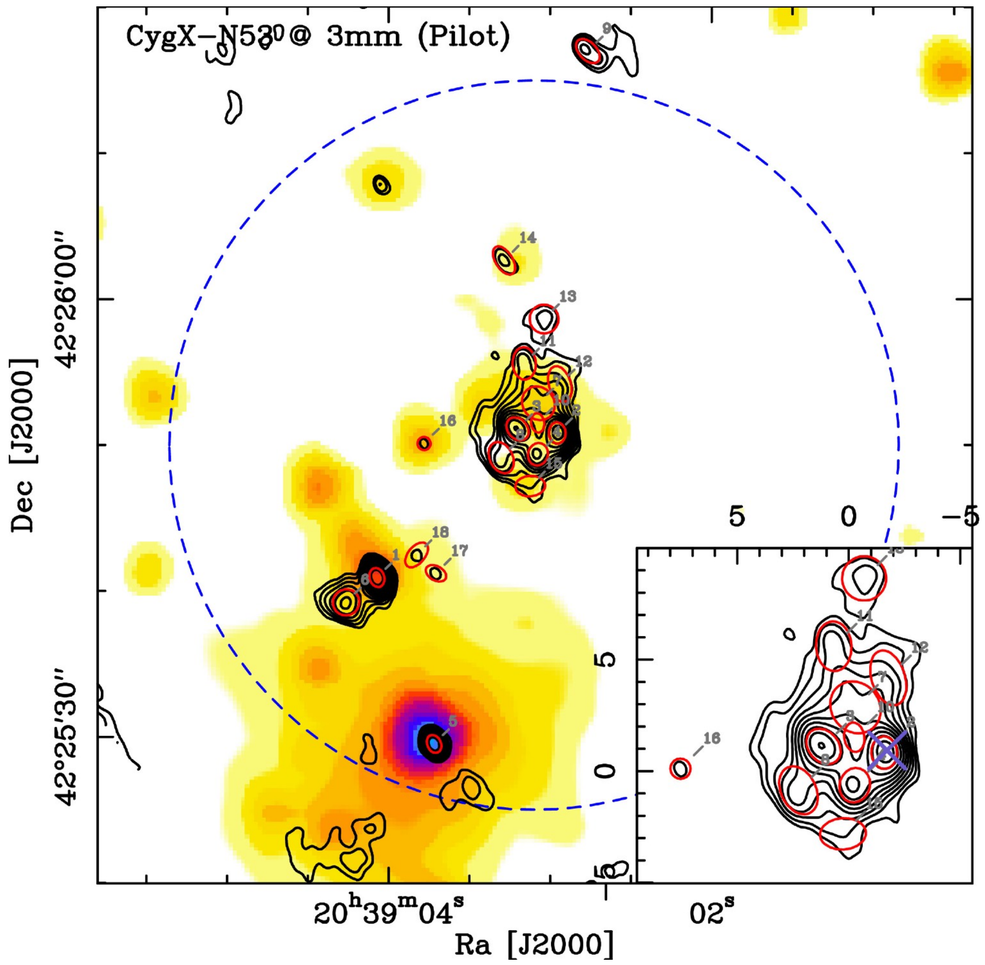
The model is able to retrieve information on the **molecular content** from **millimeter spectra**, here on a spectrum of G320 (ALMA) and of Sgr B2(N) (30m)





NASCENT-Stars pilot data : CygX-N53

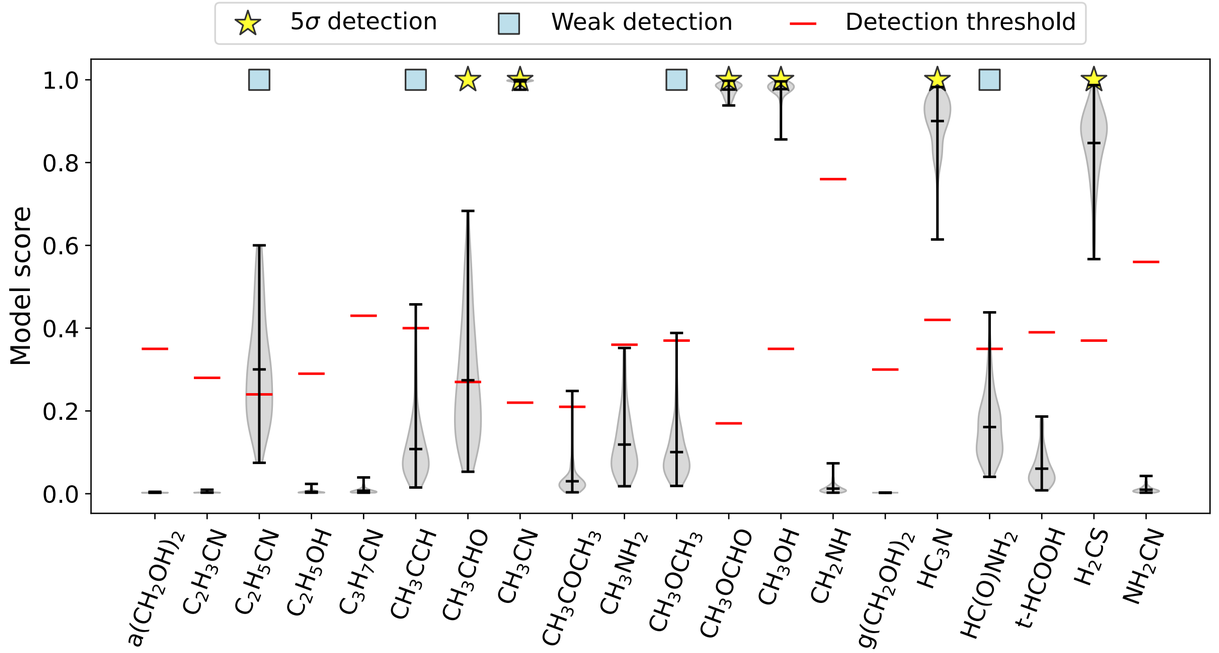
MM1 & MM2 : **high-mass protostars** from the **same gas reservoir** having different evolutionary stages (MM1 clear outflows, MM2 unclear)



Molecular differentiation
(Bouscasse+ 2022, Busch+ 2022)

→ Different heating radius for thermal desorption

→ Non thermal desorption (before gas heating)



First results for N53 MM1 :

- 6 molecules having 5σ transitions are detected by the model.
- C2H5CN is detected even with weak signal.
- 3 molecules having weak signal are not detected.

Kessler et al. in prep.

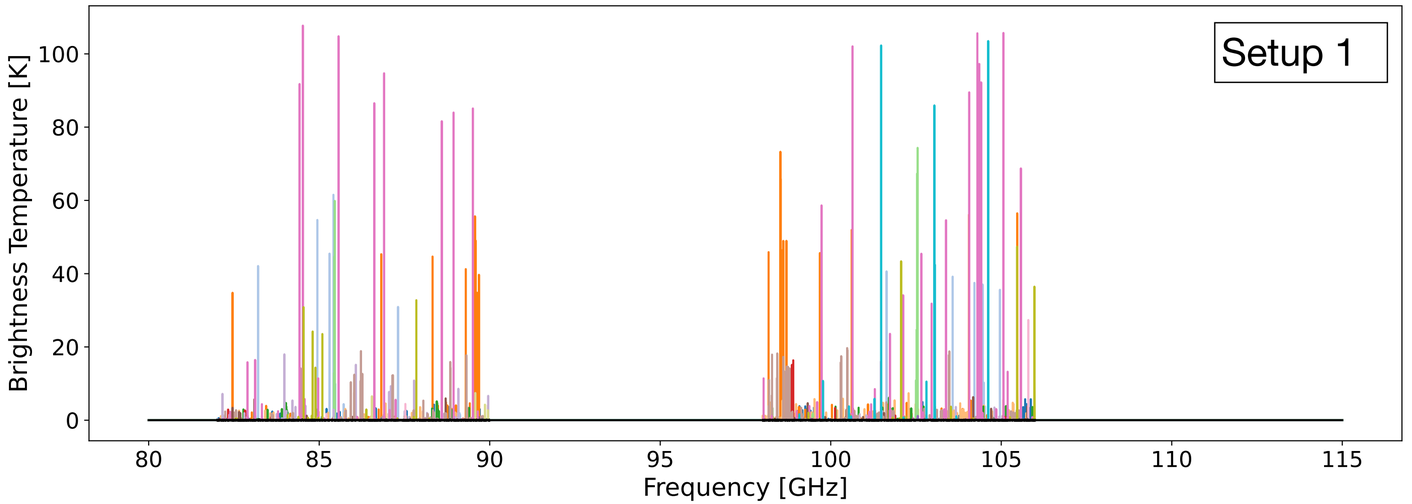
The usage of the model can be adapted to a **specific spectral coverage** through **transfer learning**.

80-115 GHz band

EMIR (30M)

Setup 1 : 82 - 90 & 98 - 106 GHz

Setup 2 : 90 - 98 & 106 - 114 GHz



Molecule	EMIR Setup 1		EMIR Setup 2		EMIR Setup 1 + 2	
	5σ lines	$P(\mathrm{det})[\%]$	5σ lines	$P(\mathrm{det})[\%]$	5σ lines	$P(\mathrm{det})[\%]$
$a(CH_2OH)_2$	155	99.7 ± 9.7	240	99.3 ± 5.0	395	99.8 ± 2.2
C_2H_3CN	49	99.7 ± 3.0	44	97.9 ± 7.2	93	99.8 ± 0.1
C_2H_5CN	81	99.8 ± 1.1	77	99.6 ± 1.0	158	99.8 ± 0.2
C_2H_5OH	109	99.8 ± 2.1	103	99.0 ± 4.1	212	99.8 ± 0.0
CH_3CCH	10	98.2 ± 10.6	0	0.6 ± 0.2	10	99.3 ± 0.6

The **learned information is reused** for a new but **related task**.

Few iterations of complementary training are enough to retrieve the same performance as for a full band.

Kessler et al. in prep.

How to **interpret** the so-called **black box** ?

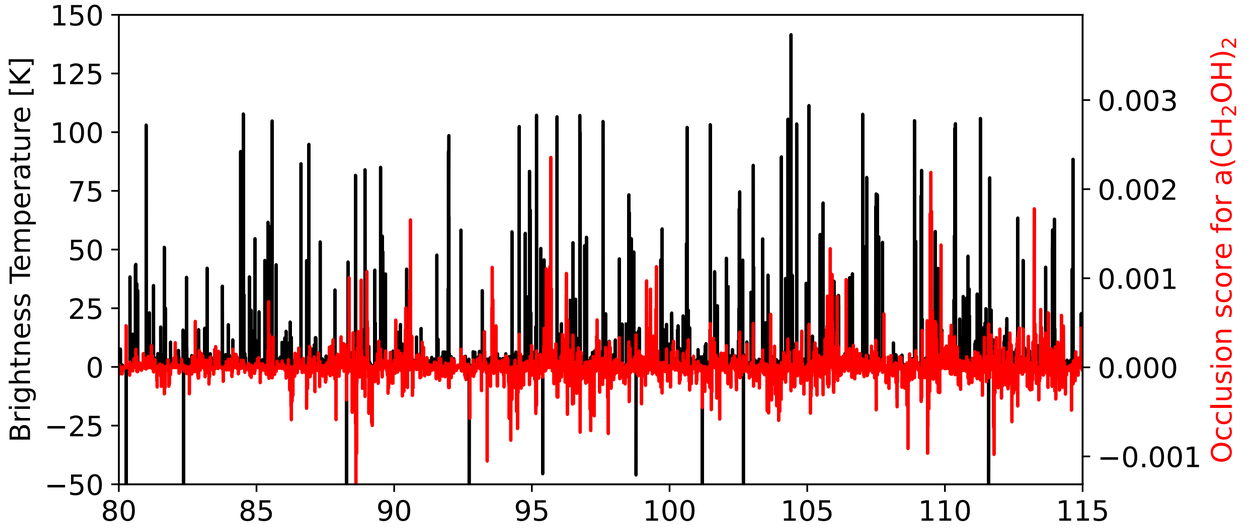
Why ? Understand its **choices**, the **learned features**, the underlying **biases.**

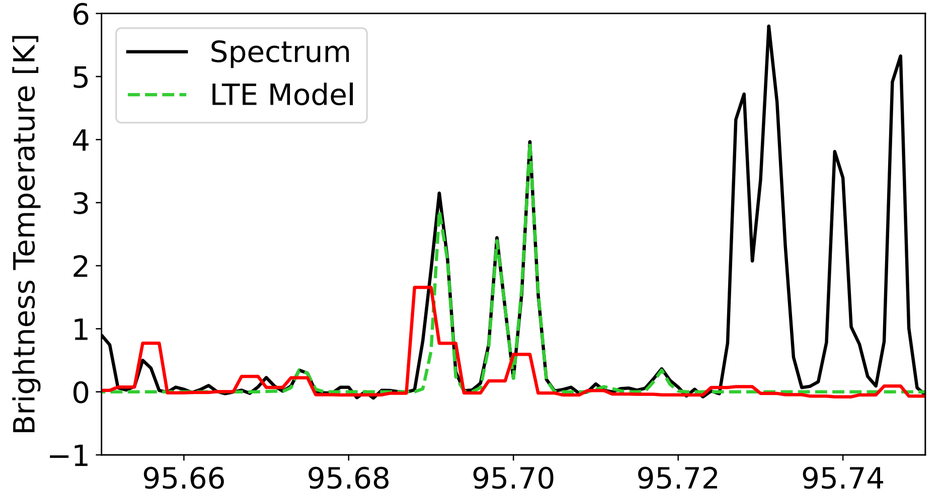
How ? Saliency maps, **occlusion analysis**, LIME, SHAP

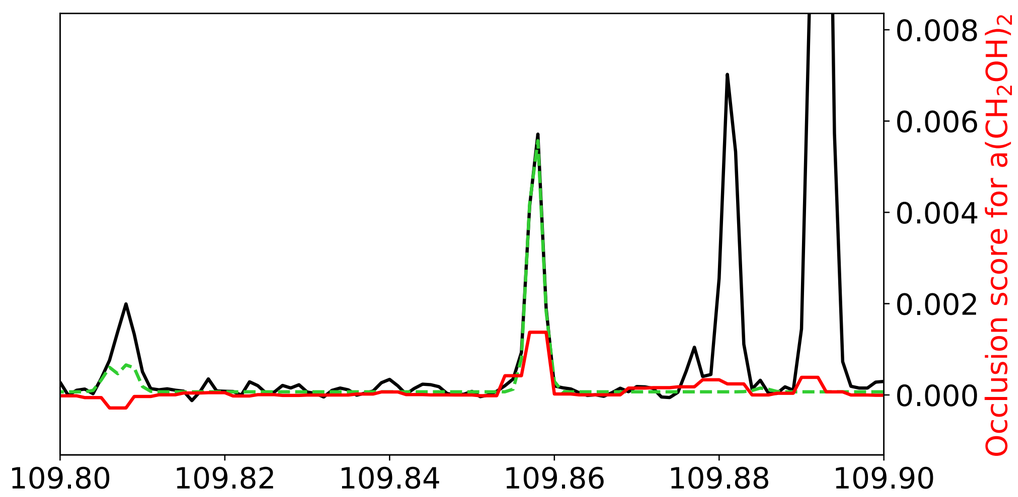
↓

Sliding window that hides lines

We can target the transitions that are important for the model to identify molecules.



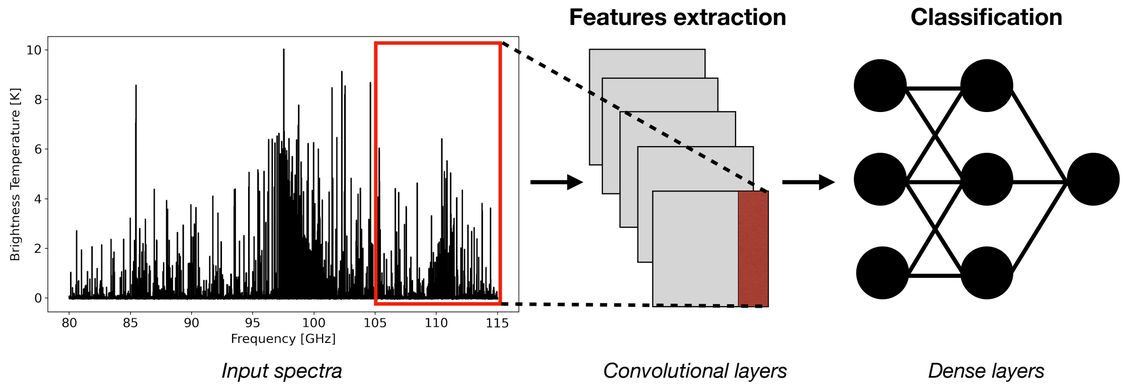




Wrap-up

Artificial Intelligence can improve the analysis of spectroscopic data from star forming regions.

We developed a deep learning method to **detect and identify** the signature of molecular species in **millimeter spectra** (Kessler et al. in prep.).



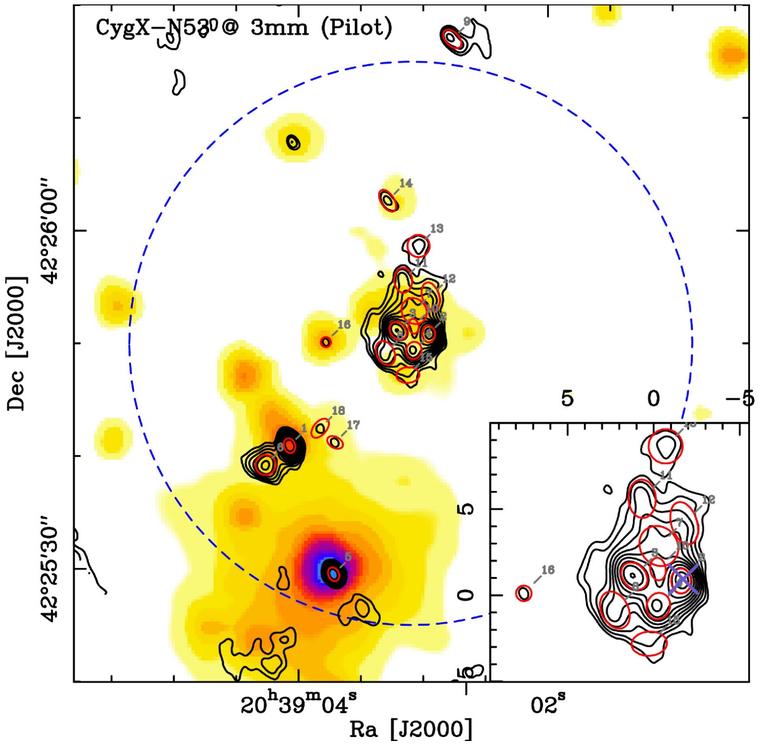
Prospects :

→ Make statistics on data cubes to study the molecular content of Cygnus-X with the **large program NASCENT-STARS**.

→ Interpretation of the model behavior + biases

→ Regression to estimate physical parameters

Email : nina.kessler@u-bordeaux.fr



ALMA (ESO/NAOJ/NRAO)