# Exorcising the Ghost in the Machine

## Synthetic Spectral Data Cubes for Assessing Big Data Algorithms

**Mauricio Araya**, Mauricio Solar,
Diego Mardones and Teodoro Hochfärber

Laboratory of Interdisciplinary Research in Astroengineering
Universidad Técnica Federico Santa María
Chile

# Astronomical Big Data Analysis

# Astronomical Big Data Analysis

Machine Learning is the solution,

# Astronomical Big Data Analysis

Machine Learning is the solution, right?

# Astronomical Big Data Analysis

Machine Learning is the solution,

right?

No free lunch!

- data consuming, or...
- highly dependent of prior knowledge
- verifiable (labelled) real data is scarce
- more advanced $\sim$ more complex
- more flexible $\sim$ more parameters
- data analysis science (study)
- parameter tunning is a nightmare

# Astronomical Big Data Analysis

Machine Learning is the solution,

right?

No free lunch!

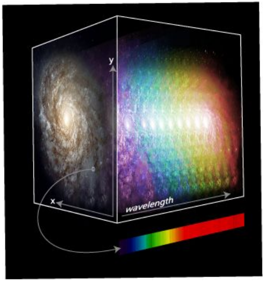
- data consuming, or...
- highly dependent of prior knowledge
- verifiable (labelled) real data is scarce
- more advanced $\sim$ more complex
- more flexible $\sim$ more parameters
- data analysis science (study)
- parameter tunning is a nightmare

We need synthetic data!

# Astronomical SYnthetic Data Observations (ASYDO)

- **Synthetic** ALMA-like data generator
- **Simple/mock** astrophysical models
- **Arbitrary** data cubes observations (FITS)
- Opportunity for Machine Learning
  - Labelled and reliable data
  - Unbounded number of samples
  - Data-driven sensitivity analysis
- Opportunity for information technologies in general
  - Assessment of bulk-data transfer, compression, etc.
  - Assessment of image analysis techniques
  - Assessment of storage systems

# Data Cube Characterization

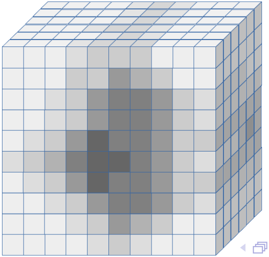


A spectroscopic data cube with calibrated temperatures, with two spatial axes and a frequency axis.
Basic model:

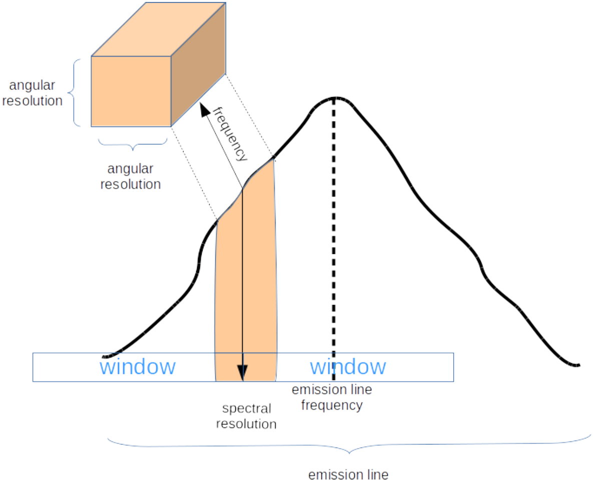
$$C(x,y,f) = \hat{C}(x,y,f) + \mathcal{N}(0,\sigma) \tag{1}$$

What to simulate in $\hat{C}(x,y,f)$?

- Emission lines frequency, energy, etc (DB)
- Local radial velocity gradients
- Red-shift correction
- Broadening functions (frequency distribution model)
- Reasonable spatial distribution models



# Data Cell Characterization

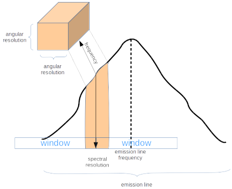


The temperature of a cell is given by the following model:

$$C(x,y,f) = \sum_{l \in \mathcal{L}(x,y,f,l)} \int_{\nu_f - \Delta\nu/2}^{\nu_f + \Delta\nu/2} Br(\nu,l)df + \epsilon \qquad (2)$$

Example

# Data Cell Characterization



The temperature of a cell is given by the following model:

$$C(x,y,f) = \sum_{l \in \mathcal{L}(x,y,f,l)} \int_{\nu_f - \Delta\nu/2}^{\nu_f + \Delta\nu/2} Br(\nu, l) df + \epsilon \qquad (2)$$

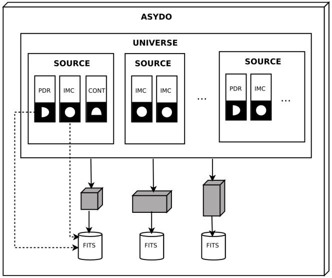
## Example

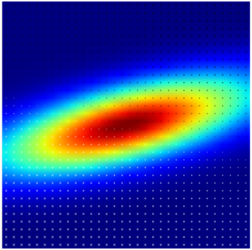
By assuming a Gaussian line profile,

$$C(x,y,f) = \sum_{l \in \mathcal{L}(x,y,f,l)} \int_{\nu_f - \Delta\nu/2}^{\nu_f + \Delta\nu/2} \frac{T_l(x,y)}{S_l(x,y)\sqrt{2\pi}} exp\left(-\frac{(\nu - \nu_l(1 + Z_l(x,y)))^2}{2S_l(x,y)^2}\right) df + \epsilon \quad (3)$$

For each line $l$ we need to generate $T_l$ (temperature), $Z_l$ (redshift) and broadening parameters $\Phi_l$ (i.e. $S$ in example)

# ASYDO Elements





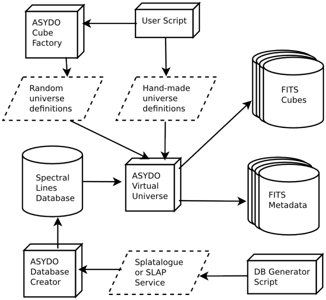
Elements:

- Persistent object called Virtual Universe (VU)
- Sources that belongs to VU ($\alpha_S$,$\delta_S$,$z_S$).
- Sources have several components (structures)
- Components use a models
  - Molecular Clouds, PDR, blackbody, continuum
- A model generates each $T_l$, $Z_l$ and $\phi_l$
- Arbitrary observations
  - Angular position ($\alpha$,$\delta$)
  - Angular resolution $\Delta\theta$ and the Field of View (FOV)
  - Central frequency $\nu$, spectral resolution ($\delta\nu$) and spectral bandwidth (BW)

Modules:

- `asydopy.vu` virtual universe (asydo core)
- `asydopy.db` line database manipulation (SLAP service)
- `asydopy.factory` generate randomized cubes

# ASYDO Elements



ASYDO Cube Factory

User Script

Random universe definitions

Hand-made universe definitions

FITS Cubes

Spectral Lines Database

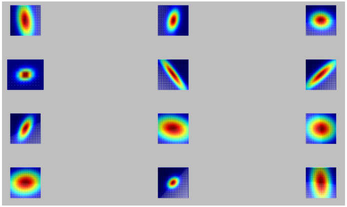
ASYDO Virtual Universe

FITS Metadata

ASYDO Database Creator

Splatalogue or SLAP Service

DB Generator Script



Elements:

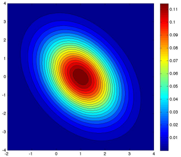
- Persistent object called Virtual Universe (VU)
- Sources that belongs to VU $(\alpha_S,\delta_S,z_S)$.
- Sources have several components (structures)
- Components use a models
  - Molecular Clouds, PDR, blackbody, continuum
- A model generates each $T_l$, $Z_l$ and $\phi_l$
- Arbitrary observations
  - Angular position $(\alpha,\delta)$
  - Angular resolution $\Delta\theta$ and the Field of View (FOV)
  - Central frequency $\nu$, spectral resolution $(\delta\nu)$ and spectral bandwidth (BW)

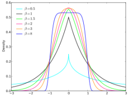
Modules:

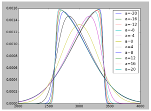
- `asydopy.vu` virtual universe (asydo core)
- `asydopy.db` line database manipulation (SLAP service)
- `asydopy.factory` generate randomized cubes

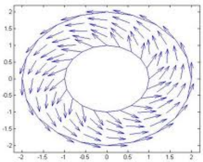
# Tools for the Models

- Spatial structures
  - Gaussian 2D
  - Generalized Gaussian 2D (saturated, Lorenzian, etc)
  - Exponential
  - Soft-edge rings (TODO)
  - Random Clouds (TODO)
- Spectral form
  - Skew-Normal Distribution (1D)
- Local shift functions
  - Linear
  - Exponantial



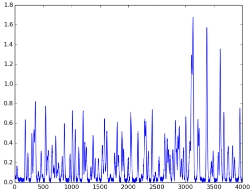


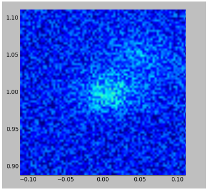


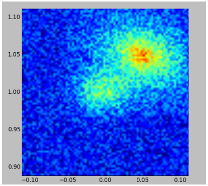


# What we save in the FITS?

- A 3D image (cube)
- For each component (and subcomponent) we have
  - 2D images with the original matrices
    - Temperature ($T_m$)
    - Red-shift ($Z_m$)
    - Broadening ($\Phi_m$)
  - A FITS table with each line of the component
  - This include:
    - unique line code
    - molecule name
    - chemical name
    - rest frequency
    - observed frequency
    - base red-shift
    - temperature







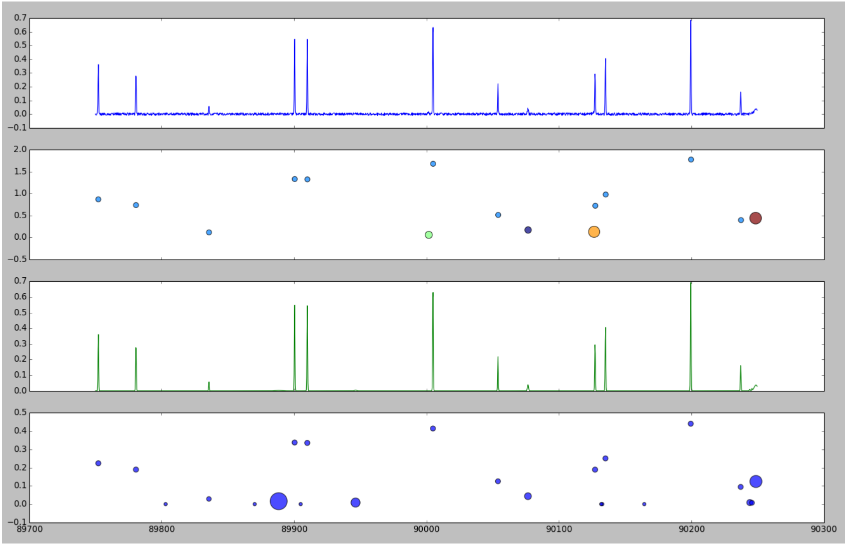
# Examples

## Supervised Learning Example:

- Pick a 2GHz frequency window ($\sim$ 300 GHz)
- Select randomly (p=0.3) if a cube have a molecule
- Force Phosphapropynylidyne existence and abscense (Binary Class)
- 30000 cubes, 25x25x1000 size each
- Naive approach: Use a SVM to train and test
- Result: 62% (something)
- The ML approach is insanely simple

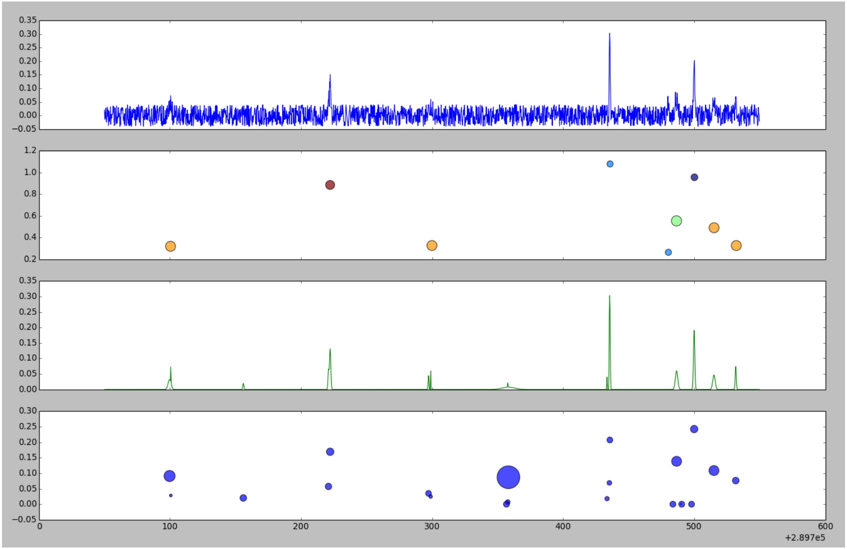
# Examples

## Line Identification Example
(wavelet-peak-detect, levenberg-marquardt gaussian fitting)



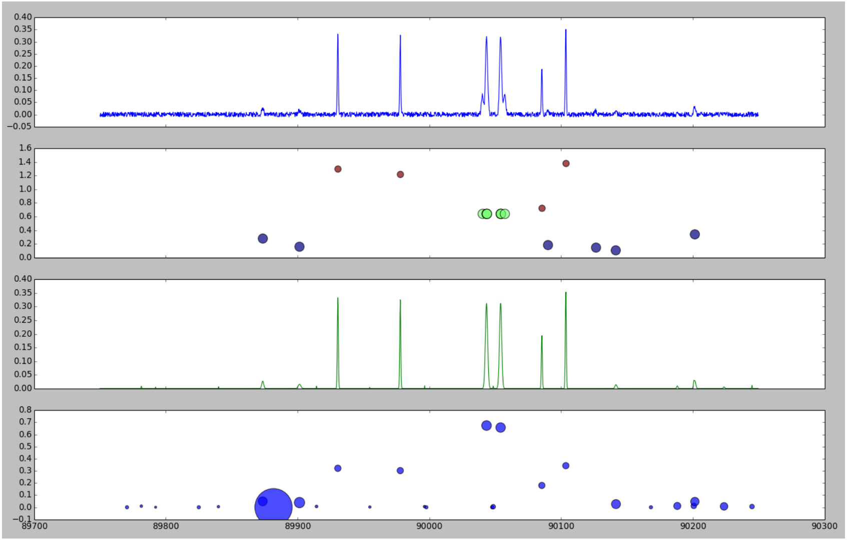
# Examples

## Line Identification Example (wavelet-peak-detect, levenberg-marquardt gaussian fitting)



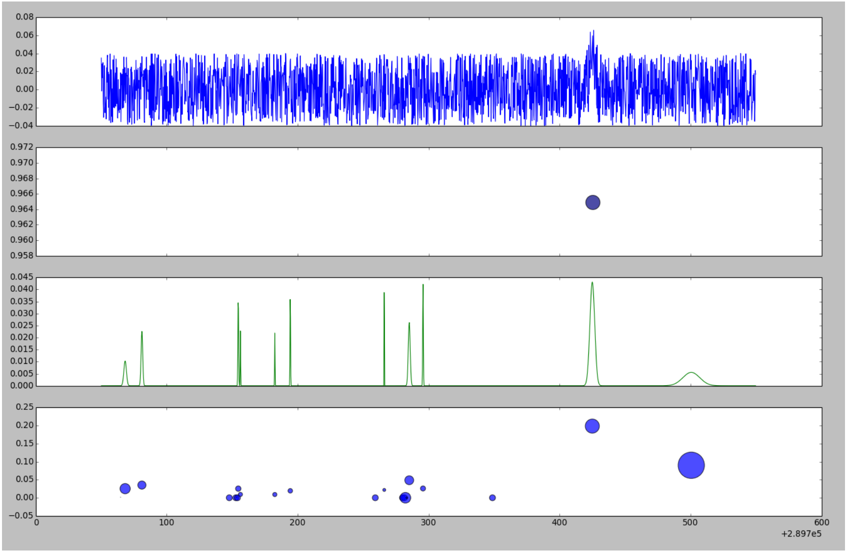
# Examples

## Line Identification Example
(wavelet-peak-detect, levenberg-marquardt gaussian fitting)



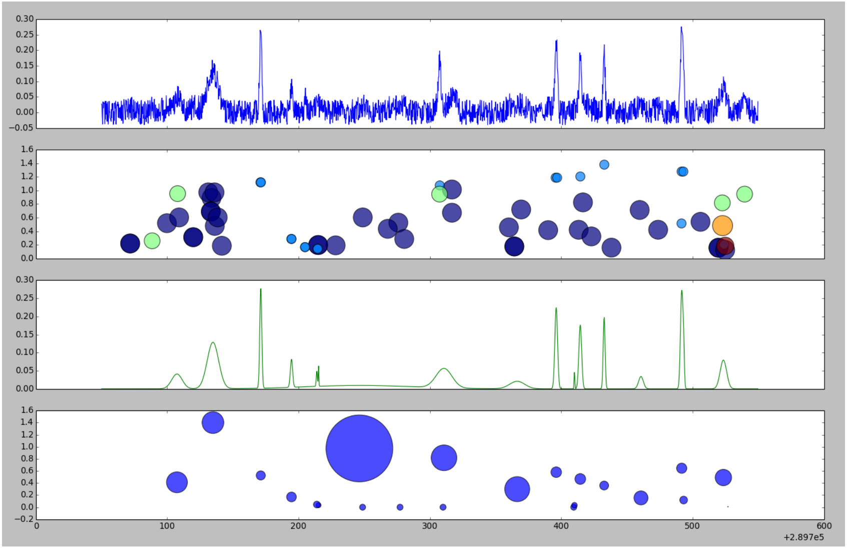
# Examples

## Line Identification Example (wavelet-peak-detect, levenberg-marquardt gaussian fitting)



# Examples

## Line Identification Example (wavelet-peak-detect, levenberg-marquardt gaussian fitting)



# Future Work

- Virtual Universe Service!

- IVOA-like synthetic data generation standard (web)
- Include more models and tools
- Integration with `astropy` and/or CASA
- Train using **synthetic data**, test using **real data**

# Exorcising the Ghost in the Machine

## Synthetic Spectral Data Cubes for Assessing Big Data Algorithms

**Mauricio Araya**, Mauricio Solar,
Diego Mardones and Teodoro Hochfärber

Laboratory of Interdisciplinary Research in Astroengineering
Universidad Técnica Federico Santa María
Chile

# Skew-normal Distribution

- The pdf of the skew normal (SN) distribution is:

$$f(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\left(\frac{x-\xi}{\omega}\right)\right) \tag{4}$$

  where $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ is the standard normal pdf and $\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$ the standard normal cdf.
- The parameters of $SN(\xi, \omega, \alpha)$ are $\xi$=location, $\omega$=scale and $\alpha$=shape
- We propose first to reparametrize as follows

$$\mu = E[X] = \xi + \omega\delta\sqrt{\frac{2}{\pi}} \tag{5}$$

$$\sigma^2 = E[(X-\mu)^2] = \omega^2\left(1 - \frac{2\delta^2}{\pi}\right) \tag{6}$$

$$\delta = \frac{\alpha}{\sqrt{1+\alpha^2}} \tag{7}$$

  which gives the following form

$$SN'(\mu, \sigma, a) = SN\left(\mu - \frac{\sigma\delta}{\sqrt{1-\frac{2\delta^2}{\pi}}} \cdot \sqrt{\frac{2}{\pi}}, \frac{\sigma}{\sqrt{1-\frac{2\delta^2}{pi}}}, \alpha\right) \tag{8}$$

# Example: simple model

Defining $T_l$, $Z_l$ y $\Phi_l$ matrices for each line is tedious. Group by molecules:

- molecule intensity maps $T_m$,
- local molecule redshift maps $Z_m$,
- and molecule broadening maps $S_m$.

The molecular model need to define (simple example):

$$T_l = f(T_m, l) = T_m exp\left(-\frac{|e_l - t|}{t}\right)$$

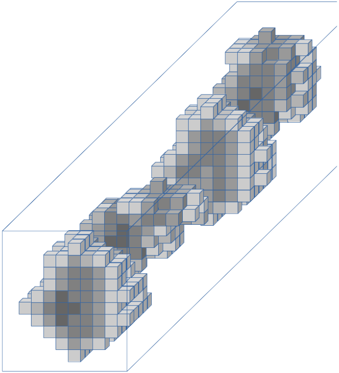
$$Z_l = g(Z_m, l) = Z_m$$

$$\phi_l = h(\phi_m, l) = \frac{S_m \nu_l}{2\sqrt{2\ln 2}c}$$

$$T_m = 2DGauss(\sigma_x, \sigma_y, \theta)$$

$$Z_m = Linear(\alpha, \beta, \theta)$$

$$S_m = s_m$$



# Spectral Line Database

$\mathcal{L}$ is the set of all lines, and $\nu_l$ the central frequency of $l \in \mathcal{L}$. If the frequency range is constrained to $\mathcal{R} = [\nu_{min}, \nu_{max}]$, the set of potentials peaks in $\mathcal{R}$ is:

$$\mathcal{L}_\mathcal{R} = \{l \in \mathcal{L} | \nu_l \in \mathcal{R}\}$$

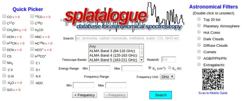
A line $l$ has other associated values in the DB such as the transition temperature $e_l$ or the molecule species $m_l$ (formula). For example, the set of species in an arbitrary region $\mathcal{R}$ is defined as

$$\mathcal{M}_\mathcal{R} = \{m \in \mathcal{M} | \exists l \in \mathcal{L}_\mathcal{R}, m = m_l\},$$

where $\mathcal{M}$ is the set of all the molecules in the DB.

## Assumptions

- **S1**: the DB contains all the observable lines
- **S2**: each transition has an associated molecule



Quick Picker

□ CO $v = 0$ □ $^{13}$CO $v = 0$
□ C$^{17}$O □ C$^{18}$O
□ CH$_3$OH $v_t = 0$ □ H$_2$CO
□ HCN $v = 0$ □ HNC $v = 0$
□ H$^{13}$CN $v = 0$ □ HC$^{15}$N $v = 0$
□ DCN $v = 0$ □ HCO$^+$ $v = 0$
□ CS □ H$^{13}$CO$^+$
□ NH$_3$ □ C I
□ C II □ O I
□ O III □ N II
□ H$_2$O $v = 0$ □ HDO
□ SiO $v = 0$

splatalogue
database for astronomical spectroscopy

Search: ex: ammonia, carbon monoxide, methanol, water, CO, NH3 etc

Telescope Bands: Any
ALMA Band 3 (84-116 GHz)
ALMA Band 4 (125-163 GHz)
ALMA Band 5 (163-211 GHz)

Redshift: 0.0

Energy Range: Min Max ● $E_L$ (cm$^{-1}$) ○ $E_L$ (K)

Frequency Range: Frequency Unit: GHz

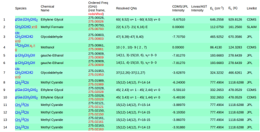
Min Max

+ Frequency - Frequency Search

Astronomical Filters
(Double click to unselect)

○ Top 20 list
○ Planetary Atmosphere
○ Hot Cores
○ Dark Clouds
○ Diffuse Clouds
○ Comets
○ AGB/PPN/PN
○ Extragalactic

Scan to Mobile Splat



| | Species | Chemical Name | Ordered Freq (GHz) (rest frame, redshifted) | Resolved QNs | CDMS/JPL Intensity | Lovas/AST Intensity | $E_L$ (cm$^{-1}$) | $E_L$ (K) | Linelist |
|---|---|---|---|---|---|---|---|---|---|
| 1 | g'Ga-(CH$_2$OH)$_2$ | Ethylene Glycol | 275.00026, 275.00026 | 60( 8,52) v= 1 - 60( 8,53) v= 0 | -5.67510 | | 646.2558 | 929.8126 | CDMS |
| 2 | CH$_3$OCHO v=0 | Methyl Formate | 275.00793, 275.00793 | 22( 6,17)- 21( 6,16) E | 0.00000 | | 112.0750 | 161.2500 | SLAIM |
| 3 | cis-CH$_2$OHCHO v=0 | Glycolaldehyde | 275.00803, 275.00803 | 47( 8,39)-47( 8,40) | -7.70750 | | 465.9252 | 670.3586 | JPL |
| 4 | $^{13}$CH$_3$OH $v_t$ = 0 | Methanol | 275.00861, 275.00861 | 10 ( 0 , 10)- 9 ( 2 , 7) | 0.00000 | | 86.4130 | 124.3283 | CDMS |
| 5 | g-CH$_3$CH$_2$OH | gauche-Ethanol | 275.00909, 275.00909 | 14(11, 3)-15(10, 6), $v_t$= 0- 0 | -7.81270 | | 193.6683 | 278.6439 | JPL |
| 6 | g-CH$_3$CH$_2$OH | gauche-Ethanol | 275.00909, 275.00909 | 14(11, 4)-15(10, 5), $v_t$= 0- 0 | -7.81270 | | 193.6683 | 278.6439 | JPL |
| 7 | cis-CH$_2$OHCHO v=0 | Glycolaldehyde | 275.01953, 275.01953 | 37(12,26)-37(11,27) | -3.62870 | | 324.3232 | 466.6261 | JPL |
| 8 | CH$_3$$^{13}$CN | Methyl Cyanide | 275.01989, 275.01989 | 15(12)-14(12), F=14-14 | -6.24000 | | 777.4904 | 1118.6288 | JPL |
| 9 | g'Ga-(CH$_2$OH)$_2$ | Ethylene Glycol | 275.02028, 275.02028 | 45( 2,43) v= 1 - 45( 2,44) v= 0 | -5.59110 | | 332.2653 | 478.0529 | CDMS |
| 10 | g'Ga-(CH$_2$OH)$_2$ | Ethylene Glycol | 275.02028, 275.02028 | 45( 3,43) v= 1 - 45( 1,44) v= 0 | -5.48190 | | 332.2653 | 478.0529 | CDMS |
| 11 | CH$_3$$^{13}$CN | Methyl Cyanide | 275.02121, 275.02121 | 15(12)-14(12), F=15-14 | -3.88970 | | 777.4904 | 1118.6288 | JPL |
| 12 | CH$_3$$^{13}$CN | Methyl Cyanide | 275.02150, 275.02150 | 15(12)-14(12), F=14-15 | -9.19350 | | 777.4904 | 1118.6288 | JPL |
| 13 | CH$_3$$^{13}$CN | Methyl Cyanide | 275.02161, 275.02161 | 15(12)-14(12), F=16-15 | -3.86070 | | 777.4904 | 1118.6288 | JPL |
| 14 | CH$_3$$^{13}$CN | Methyl Cyanide | 275.02163, 275.02163 | 15(12)-14(12), F=14-13 | -3.91880 | | 777.4904 | 1118.6288 | JPL |