Astronomical and Astrophysical Transactions (*AApTr*), 2025, Vol. 35, Issue 4, pp. 439–448,
ISSN 1055-6796, Photocopying permitted by license only, ©Cambridge Scientific Publishers

Identification of chemical species with MOLLId in the molecular gas of the hot core RCW 120 S2

A.A. Farafontova*, M.S. Kirsanova

Institute of Astronomy, Russian Academy of Sciences, 119017, 48 Pyatnitskaya Str., Moscow, Russia

Received 15 June, 2025; accepted 2 August, 2025

In this work, we present the results of molecules identification using our developed program of automated identification MOLLId. In the program, we approximated the profiles of the lines with a Gaussian function and used the observed central frequencies in the matching function which is based on the search through spectroscopic databases (CDMS and JPL). The application of the program allowed us to identify 400 emission lines in six broadband observational spectra of the hot core RCW 120 YSO S2. In the observational spectra, most of the identified lines are due to the rotational transitions of simple molecules SO, OCS, SiO, HCO, H_2CO, SO_2, SO_2, H_2CS and complex organic molecules CH_3OH, CH_3CCH and CH_3CN, CH_3OCH_3, CH_3OCHO, CH_3CHO, CH_3CHO, C_2H_3CN and others. We found species which consist of up to 9 atoms, e.g., C_2H_5CN, isotopologs of $^{13}CH_3OH$, ^{33}SO, $^{33}SO_2$, ^{34}SO, $OC^{34}S$, $H_2C^{33}S$, $H_2C^{34}S$, and deuterated molecules such as HDO. Using the program, the time of spectral analysis was reduced from 40 h per spectrum (manual processing) to $\leq$10 min with the line detection limit of $\geq 7\sigma$.

Keywords: Star formation, molecules, line identification

1 Introduction

Emission lines of small and complex molecules appear in different mediums: from diffuse interstellar gas and stellar envelopes to dense cores in star-forming regions. These lines can serve as a tool in the estimation of chemical composition, physical parameters, and kinematic structure of these regions. Earliest stages of star formation are the most diverse in molecular species; therefore, their analysis can provide information essential to investigate the chemical pathways.

Hot cores represent one of the earliest stages of star evolution. In this phase, the protostar is deeply embedded in surrounding gas and the temperature is high enough to sublimate molecules from the surface of dust grains and form the complex species in the gas phase. Spectra of such hot cores are rich in molecular lines. There, the

*Email: farafontova@inasan.ru

complex molecules, e.g. CH_3CN, $CH_2(OH)CHO$, C_2H_5CN, $HCOOCH_3$ (Calcutt et al., 2014) etc. were found and studied in detail. Identification of molecules hinges on spectroscopic databases, for example, CDMS (Müller et al., 2001) and JPL (Pickett et al., 1998). Taking into account the complexity of a spectrum and the number of spectral transitions presented in the databases, the identification process takes a lot of resources and time. Therefore, different programs and methods are used in order to ease the spectral analysis.

Most simple methods of spectral analysis and line identification include manual extraction of observed frequencies and matching to the frequencies in databases. The excitation conditions are taken into account by researchers, and then the most probable molecules are chosen. More complex methods include the usage of Gaussian fit or Gaussian decomposition methods for the extraction of spectral line parameters and search of a best fit of a spectrum calculated in LTE (Local Thermodynamical Equilibrium) conditions and the observed one. For example, tools like `GaussPy` (Lindner et al., 2015) and `SPIF` (Juvela and Tharakkal, 2024) fit one or multiple Gauss components into spectral lines and derive the parameters of these components. In general, they are used to decompose the complex profile. Programs, like `XCLASS` (Möller et al., 2017) and `PySpecKit` (Ginsburg et al., 2022), use a predefined or given list of molecules and calculate line profiles at the particular excitation temperature and column density. Calculation of LTE spectra and identifying lines this way is generally more accurate than manual identification, but it takes some predefined conditions like the chemical composition of an object and the approach that all molecules' emission arises in the same physical conditions, which can be incorrect in some ways.

The RCW 120 is a well-known Galactic HII region, appearing as a bubble-shaped structure in the image plane and located 1.34 kpc from the Sun (Deharveng et al. 2009). In the work of Zavagno et al. (2007) numerous Class 0-I and Class II sources were found towards the millimeter condensations that border the ionized region. Later works, e.g. by Zavagno et al. (2010), Figueira et al. (2018) detected massive YSOs towards the condensation 1 near the south-west border and proposed that the most massive source in this condensation is a massive Class 0 object. In our work, we use APEX observations of the most massive YSO in the condensation 1, namely source ID 2 in Table 5 of Figueira et al. (2018). For this source, the name YSO S2 will be used below.

In the study, we apply an automated tool `MOLLId` developed by us for the line identification and spectral lines' parameter derivation (Farafontova et al., in prep.). We aim to test the developed program and analyze the chemical composition of the YSO S2.

2 Observational data

The observations were conducted using the APEX telescope (Güsten et al. 2006) in Chile on May, June August 2022 (project E-0109.C-0623A-2022) as part of the ESO-operated share. The primary goal of these observations was to obtain high-quality spectra with a noise level of 4 mK at a spectral resolution of 0.3 $\mathrm{km\,s^{-1}}$. The receiver used was the nFLASH230 (Belitsky et al., 2018). In this article, we present results of observational setups covering the sidebands shown in Fig. 1.

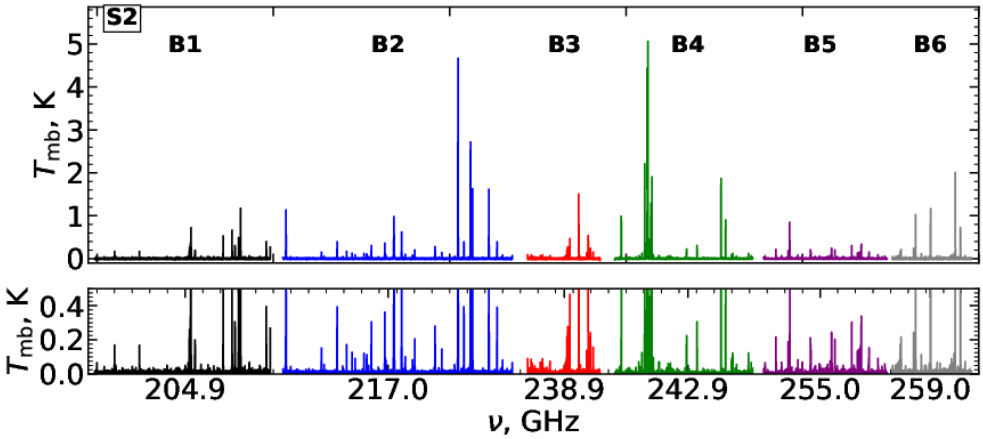


Figure 1 Observational bands B1-B6 for the YSO S2 source. Lower panel demonstrates zoomed view of the y-axis range from 0 to 0.45 K.

The data were calibrated to antenna temperature in real-time using the standard *apexOnlineCalibrator* package. Lately, we additionally applied factors of $\eta_{\rm mb} = 0.81$ and 0.73 for May/June and August, respectively[1], to scale the data at the main beam temperature scale. The spatial resolution of the data was $26''$, corresponding to 0.17 pc at the distance of RCW 120.

We observed position at $\alpha = 17^{\rm h}12^{\rm m}08.700^{\rm s}$ and $\delta = -38°30'47.4''$ (J2000). The observations were carried out in position-switching mode with the off-position at $\alpha = 17^{\rm h}12^{\rm m}08.000^{\rm s}$, $\delta = -38°36'03.00''$ (J2000). The PWV level ranged between 0.3–0.9 mm, $T_{\rm sys} \approx 100$ K. For the data presented in this article, observational time differs from 75 min to 3.5^h to reach the target noise level.

3 Methods

A two-part software package was developed to process the observational data. The first module performs spectral line profile fitting using a single-component Gaussian function, while the second module is dedicated to molecular line identification via cross-matching with spectroscopic databases. The derived central frequencies from the Gaussian fits were used as input for the second stage, where they were compared against rest frequencies from established spectroscopic catalogs.

Before applying the program package to the observational data, baseline subtraction was performed using the `arpls` function from the `pybaselines` package (Erb, 2024). This function is based on the least squares approach of asymmetric weighted smoothing.

[1] https://www.apex-telescope.org/telescope/efficiency/index.php

3.1 Approximation of observational line profiles

Spectral line profiles are mostly studied using simplified models. The most common approach is to approximate the profile using a Gaussian function. The function used for describing the line profile is a single Gaussian with three independent parameters: center frequency, line width, and line intensity maximum. In this work, the first part of the software package which is developed in the programming language `python` is aimed at successive approximation of spectral line profiles using the Gaussian function. Subsequently, center frequencies from the fit are used in the second part of the software package, i.e. in the identification of molecules. Application of the program allows the calculation of the widths, amplitudes, and central frequencies of all emission lines up to a given noise level in the spectrum. The program algorithm starts with the line corresponding to the maximum intensity in the spectrum; it is approximated by a Gaussian profile and subtracted from the spectrum. Then the next line has the intensity maximum because the previous one has been subtracted. The sequential determination of the intensity maximum in the spectrum, fitting the Gaussian profile into the line with the highest intensity, and subtracting this line from the spectrum is performed for each spectral line until the level of 3σ is reached, where σ is the noise level. To optimize the parameters, the program uses the function `least_squares` with the method `trf` of `scipy.optimize`, which minimizes the difference $\sum_{i=1}^{n}[y_i - f(x_i)]^2$, where x_i are the frequency values, y_i are the line intensity values from the observations, and $f(x_i)$ is the Gaussian function. In the sum of quadratic deviations, an array of frequencies is different at each of the steps because, after finding the optimal parameters for the line profile, this line is subtracted from the spectrum and then the frequencies whose indices correspond to the new data set y_i are taken into account.

When estimating the line parameters, an adaptive approach is used and the width of the first line found is taken into account to improve the initial approximation for the remaining lines. The errors of the resulting Gaussian parameters are computed by means of the Jacobian of the optimization function. The Jacobian of the optimization function (J) is a matrix of partial derivatives:

$$J_{ij} = \frac{\partial f(x_i, \vec{p})}{\partial p_j}, \tag{1}$$

where $\vec{p}$ is the vector of parameters (A, μ, σ) at each i-th step. The standard deviations of the parameters (σ_i) are calculated using the expression:

$$\sigma_i = \sqrt{\mathrm{diag}((J_{ij}^T J_{ij})^{-1} \cdot \frac{r_i^T \cdot r_i}{n-p}}), \tag{2}$$

where n is the number of observation data points, r_i is the difference between the data from the observations and the Gaussian approximation for a given solution, diag denotes the extraction of the diagonal elements of the matrix.

3.2 Identification of molecules

The main method used in this part of the software package is the frequency matching algorithm. The algorithm has been implemented in the method of matching molecules

Table 1 Levels used for identification of the molecules. Levels are selected by difference between molecular frequency from database and observed frequency ($\Delta\nu$), upper level energy ($E_{\rm u}$) and logarithm of Einstein coefficient of spontaneous decay($\log A_{ij}$).

Level	$\Delta\nu$ (GHz)	$E_{\rm u}$ (K)	$\log A_{ij}$
1	$\pm$0.0005	≤ 200	$\geq 1\times10^{-7}$
2	$\pm$0.001	≤ 400	$\geq 1\times10^{-8}$
3	$\pm$0.002	≤ 800	$\geq 1\times10^{-9}$
4	$\pm$0.003	≤ 1000	$\geq 1\times10^{-10}$

according to several criteria. The initial selection of molecules is done by frequency, upper level energy ($E_{\rm u}$) and the logarithm of the Einstein coefficient of the spontaneous transition ($\log A_{ij}$). The values of $E_{\rm u}$ and $\log A_{ij}$ for each selection level are given in Table 1. Molecules are first searched in a narrow spectral range ($\Delta\nu$ $\pm$0.0005) from the center line frequency and for the low energies of the upper level ($E_{\rm u} \leq 200$). The search boundaries are then extended if no molecules for a given frequency have been matched at this stage. If several candidate molecules have been matched for the same line, the candidate with the smallest frequency difference and value $E_{\rm u}$ and the largest value $\log A_{ij}$ is selected among them. The selection is made according to the list of molecules identified in the interstellar medium[2], as well as taking into account the isotopologues and isotopomers of these molecules.

4 Results

4.1 Identified species

As a result of the algorithm, 407 lines of 79 molecules were identified in the observational spectra of YSO S2 in a time of 8 min per spectral range. Distribution of the identified lines over the frequency is shown in Fig. 2. The source of spectroscopic data for 78% of the lines was the JPL database. Methanol (CH_3OH) is the most abundant molecule, accounting for 14% of the total number of molecules detected (see Fig. 4). Methyl formate (CH_3OCHO) accounts for 10% of the total and is the second most abundant, CH_3OCH_3 accounts for 6%, CH_3CHO accounts for 5% of the total number of identified molecules. The number of identified lines of methylacetylene (CH_3CCH) is 4%, H_2CS, CH_3CN, and sulfur dioxide (SO_2) are 4.2%, 3.7%, and 3.4%, respectively. The isotopologue $^{13}CH_3OH$ occupies 3.4% among all identified molecules, the presence of $SO^{18}O$ is 2.7%. The deuterated compounds CH_2DCN account for 1.5% (6 lines), HDCO 1.0% (4 lines) and CH_2DCHO 0.7% (3 lines), HDO (see Kirsanova & Farafontova, 2025) 0.5% (2 lines) of the total number of molecules.

Classification of the identified molecules by groups (see Fig. 3) shows a large number of carbon-bearing molecules, accounting for 84.5% of the total number of lines. Oxygen-bearing compounds correspond to 70.8%, CH_3-group is present in 47.7% of the total number of lines found. The sulfur-bearing species H_2CS, SO_2, OCS, HSCN, and their isotopologues account for 23.8%. Nitrogen-bearing species

[2]https://cdms.astro.uni-koeln.de/classic/molecules

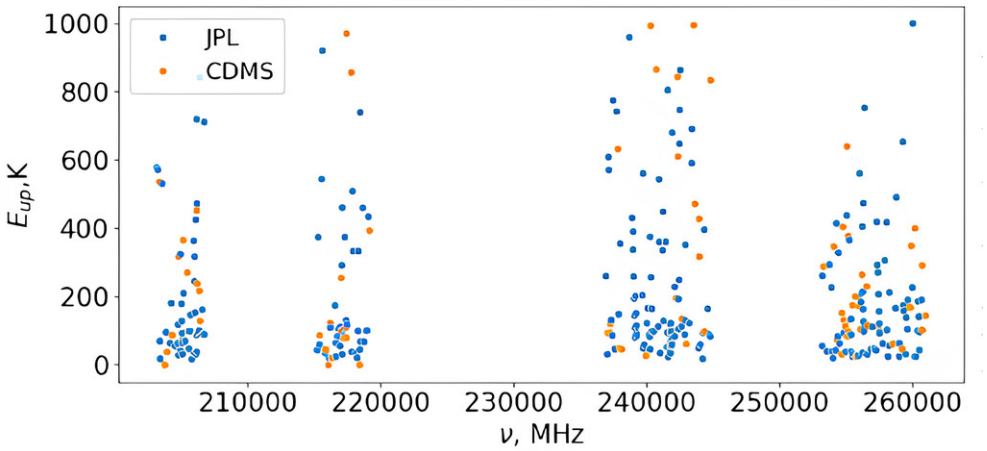


Figure 2 Scatter plot with dependence between upper-level energy and frequency for the identified molecules. Different colors show the source spectroscopic database for every identified molecule.

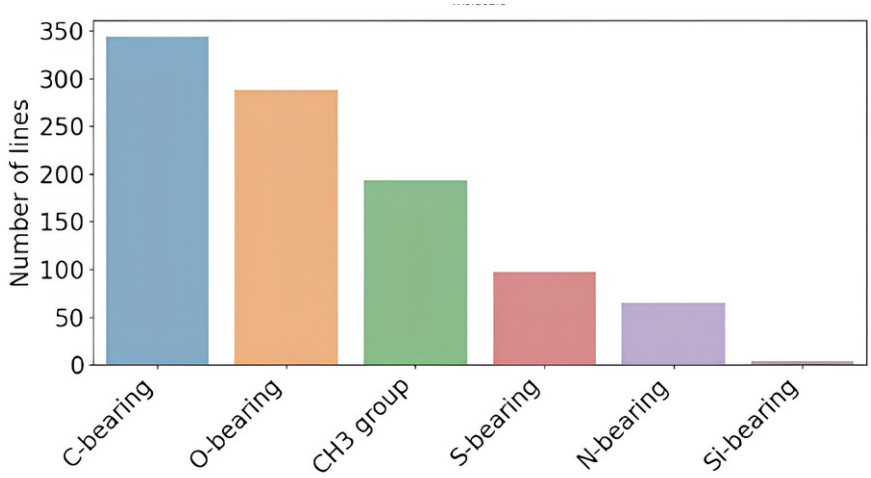


Figure 3 Column diagram with division by groups for identified molecules. Molecules were distributed in such a way that any molecule could belong to several categories at the same time. On y-axis the number of identified species which contain particular group or atom are shown.

account for 16%. Silicon monoxide (SiO) and its isotopologues ^{29}SiO, ^{30}SiO make up 1% or 4 lines.

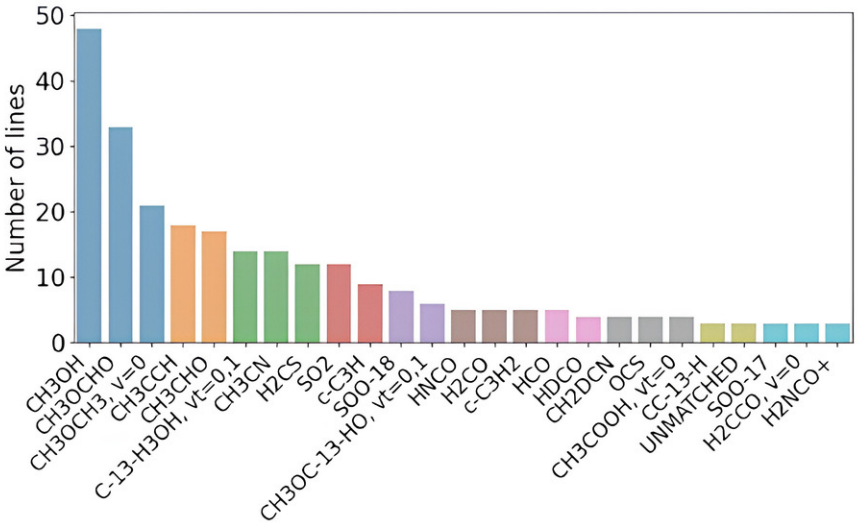


Figure 4 Column diagram with number of lines identified for each molecule.

5 Discussion

Molecular identification often requires a rigorous search through databases or usage of the LTE calculation approach to reproduce the observed spectrum. The methods of manual search, for example, in Hernández-Hernández et al. (2014) the SPLATALOGUE[3] web-site is used to identify by eye the main lines in the spectra. In Gieser et al., 2019, 67 lines were identified by manual search through CDMS and JPL database. Unfortunately, a manual search takes a lot of time and does not account for excitation conditions, so the selection of one molecule among all candidates can be a tedious task. In our software, the multi-level search allows us to match the species sequentially and filter out high upper-level energy ($E_{\rm up} \geq 1000$ K) transitions which are supposedly less probable in our type of source.

The LTE calculation can significantly improve the accuracy of the identification procedure. For example, the initial manual search is deployed in Golshan et al. (2024), Hernández-Hernández et al. (2014), and after having a complete set of molecules, including definite and tentative detections, the fit of the spectra is made using the XCLASS toolbox. In Sewiło et al. (2018) the MADCUBA software (Martín et al. 2011) is used to automatically find the transitions and perform the best fit of the observed spectrum. Further application of the LTE approximation in our identification software could allow us to separate close transitions ($\Delta\nu \leq 0.0005$ GHz), blended lines, and more precisely choose the most suitable transition under the particular physical conditions.

Line identification methods can have the fraction of unidentified lines up to 25%, see, e.g., Rolffs et al. (2011), 20% in Allen et al. (2017). In some works, the number of unidentified lines is quite low, e.g., 5% in Beuther et al. (2006) and 1.8% in

[3]https://splatalogue.online/#/home

Zernickel et al. (2012). All works use the XCLASS line identification method as the main tool to analyze the chemical composition of the source. Its efficiency depends on the processing algorithms and availability of the spectroscopic data. In the case of the manual search, efficiency of the identification mostly depends on the quality and availability of the spectroscopic data. In our identification, the number of unidentified molecules is equal to 3 out of all 407 lines (0.7%). We suggest that this number can be reduced by addition of vibrational states of different molecules or by expansion of the upper-level energy search range.

6 Conclusions

The software package, developed to simplify the analysis of large amounts of spectral data and the identification of molecular emission lines in the sub- and millimeter range, was discussed in this work. The software was applied to six bands from 202 to 260 GHz of observational spectra of hot core YSO S2 from the condensation at the border of the RCW 120 PDR. In the observational spectra of the YSO S2, we identified 407 lines of 79 molecules. The identified species differ from simple 2-atom molecules (like SO, SiO, NS) to complex up to 9 atoms (like C_2H_5CN) organic species. The most common molecule among all species is CH_3OH. The most common group of molecules is C-bearing molecules. After that group, the CH_3-group, S-bearing, and N-bearing species follow. The percentage of unidentified species is equal to 0.7% out of all lines. In further work, we aim to increase the accuracy of the fit by adding the LTE approximation of the observed spectra.

Acknowledgements

We are thankful to S. V. Salii for fruitful discussions. This work was supported by the RSF grant no. 24-22-00097.

References

Allen, V., van der Tak, F. F. S., Sánchez-Monge, Á., et al. 2017, A&A, 603, A133
Belitsky, V., Bylund, M., Desmaris, V., et al. 2018, A&A, 611, A98
Beuther, H., Zhang, Q., Sridharan, T. K., Lee, C. F., & Zapata, L. A. 2006, A&A, 454, 221
Deharveng, L., Zavagno, A., Schuller, F., et al. 2009, A&A, 496, 177
Erb, D. 2024, pybaselines: A Python library of algorithms for the baseline correction of experimental data
Figueira, M., Bronfman, L., Zavagno, A., et al. 2018, A&A, 616, L10
Gieser, C., Semenov, D., Beuther, H., et al. 2019, A&A, 631, A142
Ginsburg, A., Sokolov, V., de Val-Borro, M., et al. 2022, AJ, 163, 291
Golshan, Roya Hamedani, Sánchez-Monge, Álvaro, Schilke, Peter, et al. 2024, AA, 688, A3

Güsten, R., Nyman, L. Å., Schilke, P., et al. 2006, A&A, 454, L13
Hernández-Hernández, V., Zapata, L., Kurtz, S., & Garay, G. 2014, ApJ, 786, 38
Juvela, M. & Tharakkal, D. 2024, A&A, 685, A164
Kirsanova & Farafontova, 2025, accepted by Astronomy Letters, https://doi.org/10.48550/arXiv.2505.21945
Lindner, R. R., Vera-Ciro, C., Murray, C. E., et al. 2015, AJ, 149, 138
Martín, S., Martín-Pintado, J., Blanco-Sánchez, C., et al. 2019, A&A, 631, A159
Möller, T., Endres, C., & Schilke, P. 2017, A&A, 598, A7
Müller, H. S. P., Thorwirth, S., Roth, D. A., & Winnewisser, G. 2001, A&A, 370, L49
Pickett, H. M., Poynter, R. L., Cohen, E. A., et al. 1998, J. Quant. Spectr. Rad. Transf., 60, 883
Rolffs, R., Schilke, P., Zhang, Q., & Zapata, L. 2011, A&A, 536, A33
Sewiło, M., Indebetouw, R., Charnley, S. B., et al. 2018, The Astrophysical Journal Letters, 853, L19
Zavagno, A., Pomarès, M., Deharveng, L., et al. 2007, A&A, 472, 835
Zavagno, A., Russeil, D., Motte, F., et al. 2010, A&A, 518, L81
Zernickel, A., Schilke, P., Schmiedeke, A., et al. 2012, A&A, 546, A87