Exploring the Universe with Artificial Intelligence (UniversAI)
Proceedings IAU Symposium No. 397, 2026
I. Liodakis & V. Efthymiou, eds.
doi:10.1017/S174392132510481X

Application of Machine Learning Methods to the Identification of Star-Forming Regions

K. V. Plakitina and M. S. Kirsanova

Institute of Astronomy of the Russian Academy of Sciences
email: plakitina.kv@outlook.com

Abstract. Modern astronomy handles vast amounts of spectral data obtained from observations of molecular emissions across different sources, such as dark infrared clouds, protostars, and HII regions. Traditional analytical methods become increasingly ineffective when dealing with such large datasets. In this study, we investigate the use of machine learning techniques to analyse molecular emission lines and classify molecular clumps. Using data from the MALT90 survey, which observes the Galactic plane at 3 mm wavelengths, we explore various clustering algorithms and dimensionality reduction techniques to simplify the process of object identification. Our results show that machine learning approaches successfully distinguish molecular clumps with active star formation from other types of astronomical objects. This approach holds promise for processing data collected with advanced facilities like JWST and ALMA, leading to more efficient analysis of astronomical data.

Keywords. ISM: clouds, ISM: molecules, methods: data analysis

1. Introduction

In recent years, machine learning (ML) algorithms have been increasingly applied to complex observational datasets (Sen et al. 2022), uncovering patterns and dependencies that are difficult to detect with traditional statistical methods. ML is widely used in astronomy for tasks ranging from galaxy classification to transient detection (Aniyan and Thorat 2017; Dax et al. 2021) and is increasingly applied to star formation studies. For example, Marton et al. (2019) combined Gaia DR2, WISE, and Planck data to construct an all-sky probabilistic catalogue of over one million young stellar objects (YSO) candidates, while Angarita et al. (2025) used ML on millimetre spectra of massive YSO to identify spectral signatures linked to physical conditions. These studies demonstrate the ability of ML to extract meaningful information from large and complex datasets.

In this study, we explore ML methods for automated identification and classification of molecular clumps at different evolutionary stages using multi-wavelength millimetre and infrared data. For unsupervised analysis, clustering algorithms reveal natural groupings of clumps, while supervised classification uses labelled sources to classify objects with uncertain evolutionary stages.

2. Archival Datasets

We collected emission intensities of molecular lines and flux densities of the dust emission at 870 μm from the MALT90 dataset (Rathborne et al. 2016) using the VizieR database†. Infrared data at 3.6, 4.5, 8, and 24 μm were retrieved from the Spitzer Heritage

† https://cdsarc.cds.unistra.fr/viz-bin/cat/J/other/PASA/33.30

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union.

https://doi.org/10.1017/S174392132510481X Published online by Cambridge University Press

Archive‡. We also included data on class II CH_3OH masers within 5″ of sources from the maser database (Ladeyschikov et al. 2019) to our analysis.

3. Preliminary data reduction

The MALT90 dataset contains 269 features for 3556 molecular clumps, each assigned an infrared-based category: Q (quiescent), A (protostellar), C (compact HII regions), H (extended HII regions), P (photon-dominated region, PDR hereafter), and U (uncertain). In the first clustering stage, lines detected in fewer than 40% of sources were excluded, leaving HCO^+, HNC, N_2H^+, HCN, C_2H, and the 870 μm flux. In the second stage and classification, all MALT90 line intensities, Spitzer IR fluxes, and methanol maser indicators were used. Data were standardized with Scikit-learn's QuantileTransformer (Pedregosa et al. 2011) and reduced to two components via t-SNE (Maaten and Hinton 2008).

4. Methods

4.1. *Photometry*

We obtained Spitzer data at 3.6, 4.5, 8, and 24 μm from the Spitzer Heritage Archive†and performed photometry using IRAC (channels 1, 2, and 4) and MIPS (24 μm channel), following the respective instrument handbooks‡. The achieved accuracy (10–20%) is sufficient, since clustering relies on relative rather than absolute fluxes, and uncertainties at this level do not affect overall distributions.

4.2. *Machine learning techniques*

We used a Random Forest (Breiman 2001) classifier to identify key features and predict the evolutionary stage of uncertain clumps (Rathborne et al. 2016), training only on pre-classified objects. A Gradient Boosting classifier (Friedman 2001) was applied in the same way. For unsupervised analysis, we employed HDBSCAN (Campello et al. 2013), which groups sources by local density and naturally treats outliers as noise. All analyses were performed with the Scikit-learn library for Python (Pedregosa et al. 2011).

5. Results

We optimized clustering by testing different scalers (Pedregosa et al. 2011), t-SNE perplexities (30–100) with 2–3 components, and HDBSCAN parameters (minimum cluster size 30–150, minimum samples scaled accordingly). Each setup was scored with the silhouette metric (Rousseeuw 1987). The best silhouette score was consistently obtained with QuantileTransformer and 2-component t-SNE, which we adopt throughout unless stated otherwise.

5.1. *Clustering results*

HDBSCAN identified eleven clusters based on integrated intensities of C_2H, N_2H^+, HCN, HNC, and HCO^+, with the best silhouette score (0.68) at a minimum cluster size of 75 and sample size of 45. We show the most stable clusters content in the left

‡ https://irsa.ipac.caltech.edu/applications/Spitzer/SHA/?__action=layout.showDropDown&

† https://irsa.ipac.caltech.edu/applications/Spitzer/SHA/?__action=layout.showDropDown&

‡ https://irsa.ipac.caltech.edu/data/SPITZER/docs/irac/iracinstrumenthandbook/IRAC_Instrument_Handbook.pdf; https://irsa.ipac.caltech.edu/data/SPITZER/docs/mips/mipsinstrumenthandbook/MIPS_Instrument_Handbook.pdf

https://doi.org/10.1017/S174392132510481X Published online by Cambridge University Press

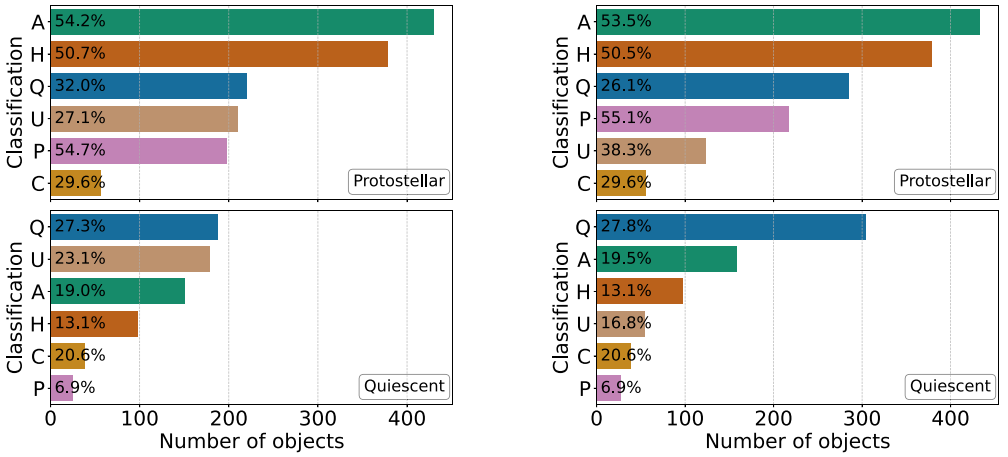


Figure 1. Distribution of clump **categories** in the two most stable HDBSCAN clusters from t-SNE on five molecular lines (HCO^+, HNC, N_2H^+, HCN, C_2H). Bars show MALT90 classification (Q, A, C, H, P, U); percentages give fractions and numbers show counts. **Left panel:** before classification of uncertain clumps. **Right panel:** after classification.

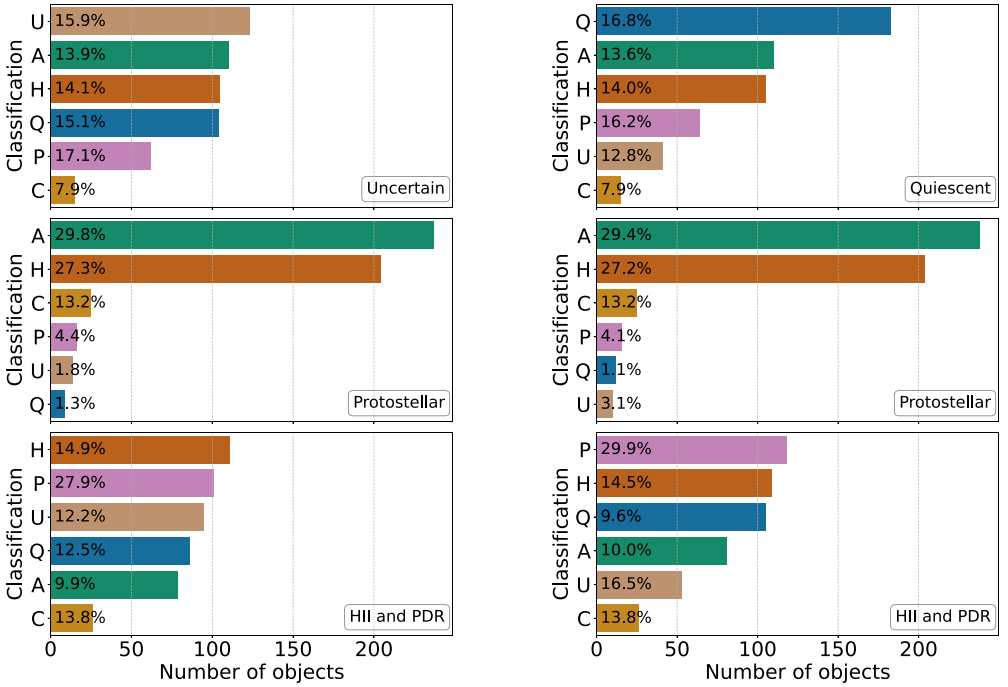


Figure 2. Distribution of clump classes in the two most stable HDBSCAN clusters from t-SNE on sixteen molecular lines (HCO^+, HNC, N_2H^+, HCN, C_2H, H^13CO^+, HN^13C, ^13CS, $^13C^34S$, HC^13CCN, HNCO, CH_3CN, HC_3N, HNCO, H 41α. **Left panel:** before classification of uncertain clumps. **Right panel:** after classification.

panel of Fig. 1. The former cluster is dominated by protostars and HII regions, while the latter cluster contain mainly IR-quiescent and uncertain clumps. Feature analysis with RF showed C_2H and N_2H^+ as the most important molecular lines to identify the clusters. Our result is reasonable since C_2H traces UV-irradiated gas and N_2H^+ is the tracer of dense gas. Adding 870μm flux had negligible impact, confirming clustering stability.

https://doi.org/10.1017/S174392132510481X Published online by Cambridge University Press

Following analysis of the clustering was performed using integrated intensities of 16 molecular lines, Spitzer infrared fluxes, and maser detection flags. The best silhouette score (0.65) was obtained with t-SNE perplexity 100, minimum cluster size 100, and minimum samples 50. As shown in the left panel of Fig. 2, the first cluster is dominated by the objects of the uncertain category, the second one – by protostars and HII regions, and the third one – by HII regions and PDRs. The most influential features were maser presence, N_2H^+, and C_2H.

5.2. *Classification results*

Over 20% of MALT90 sources were labelled as uncertain by Rathborne et al. (2016). To refine them, we applied RF and GB classifiers using 16 molecular lines and IR features. Both showed modest accuracy (~ 0.6). With a 50% threshold, RF reassigned 527/774 clumps and GB 611/774, with 505 consistent between models and only 6 in disagreement. About 80% were IR-quiescent. Notably, IR fluxes, previously low-ranked, emerged as key predictors.

Next, we replaced uncertain category with classifications from the RF and GB models and repeated clustering. Results are shown in the right panels of Fig. 1 and Fig. 2. Quiescent-dominated clusters became more distinct, while protostar- and HII+PDR-dominated clusters remained stable, enhancing cluster homogeneity and evolutionary trends.

It was noted by Rathborne et al. (2016) that there are plans to refine object classifications, including the use of maser emission. As shown in Fig. 2, adding class II methanol maser flags improved cluster separation and highlighted a protostar-dominated cluster.

6. Conclusion

This study analyzes clustering and classification of MALT90 molecular clumps. Using five molecular lines (HCO^+, HNC, N_2H^+, HCN, C_2H) separates protostellar and IR-quiescent clumps, while including IR fluxes (3.6–24 μm), all 16 integtrted line intensities, and class II CH_3OH masers reveals an additional cluster dominated by PDRs and HII regions. Clusters contained mixed MALT90 types, reflecting subjective IR-based classifications or continuous clump evolution. Random Forest and Gradient Boosting classified over 500 previously uncertain clumps, finding $\sim 80\%$ quiescent. The two algorithms largely agreed, with only six discrepancies, mainly between protostars and quiescent clumps.

References

Angarita, Y., Chaparro, G., Lumsden, S. L., Walsh, C., Avison, A., Asabre Frimpong, N., & Fuller, G. A. 2025, Pattern finding in millimetre-wave spectra of massive young stellar objects. *Astronomy & Astrophysics*, 694, A20.

Aniyan, A. K. & Thorat, K. 2017, Classifying Radio Galaxies with the Convolutional Neural Network. *ApJS*, 230(2), 20.

Breiman, L. 2001, Random forests. *Machine learning*, 45, 5–32.

Campello, R. J., Moulavi, D., & Sander, J. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining* 2013, pp. 160–172. Springer.

Dax, M., Green, S. R., Gair, J., Macke, J. H., Buonanno, A., & Schölkopf, B. 2021, Real-Time Gravitational Wave Science with Neural Posterior Estimation. *Phys. Rev. Lett.*, 127(24), 241103.

Friedman, J. H. 2001, Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

https://doi.org/10.1017/S174392132510481X Published online by Cambridge University Press

Ladeyschikov, D. A., Bayandina, O. S., & Sobolev, A. M. 2019, Online Database of Class I Methanol Masers. *AJ*, 158(6), 233.

Maaten, L. v. d. & Hinton, G. 2008, Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov), 2579–2605.

Marton, G., Ábrahám, P., Szegedi-Elek, E., Varga, J., Kun, M., Kóspál, Á., Varga-Verebélyi, E., Hodgkin, S., Szabados, L., Beck, R., & Kiss, C. 2019, Identification of Young Stellar Object candidates in the Gaia DR2 x AllWISE catalogue with machine learning methods. *MNRAS*, 487(2), 2522–2537.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., *et al.* 2011, Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12, 2825–2830.

Rathborne, J. M., Whitaker, J. S., Jackson, J. M., Foster, J. B., Contreras, Y., Stephens, I. W., Guzmán, A. E., Longmore, S. N., Sanhueza, P., Schuller, F., Wyrowski, F., & Urquhart, J. S. 2016, Molecular Line Emission Towards High-Mass Clumps: The MALT90 Catalogue. *PASA*, 33, e030.

Rousseeuw, P. 1987, Rousseeuw, p.j.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. comput. appl. math. 20, 53-65. *Journal of Computational and Applied Mathematics*, 20, 53–65.

Sen, S., Agarwal, S., Chakraborty, P., & Singh, K. P. 2022, Astronomical big data processing using machine learning: A comprehensive review. *Experimental Astronomy*, 53(1), 1–43.

https://doi.org/10.1017/S174392132510481X Published online by Cambridge University Press