Global Scientific JOURNALS

GSJ: Volume 13, Issue 11, November 2025, Online: ISSN 2320-9186
www.globalscientificjournal.com

Predicting Interstellar Molecular Detectability Using Machine Learning on Chemical Descriptors

Caden L. Reedy

Abstract

Understanding the presence and distribution of molecular species in interstellar environments is fundamental to astrochemistry and the study of chemical evolution in planetary systems. This work applies to a machine learning framework to quantify and predict molecular detectability in space based on intrinsic chemical properties. A curated dataset of 49 molecules, comprising both detected and undetected species, was analyzed using molecular descriptors including molecular weight, polarity (LogP and topological polar surface area), hydrogen-bonding potential, rotatable bonds, ring structures, and heavy atom count. Logistic regression was employed to generate probabilistic predictions of detectability. Results indicate that smaller, polar, and acyclic molecules are more readily detectable, whereas larger, aromatic, and hydrophobic molecules exhibit lower detectability. This predictive model provides a systematic approach for prioritizing candidate molecules for spectroscopic investigation, advancing our understanding of chemical complexity in interstellar space.

Methods

A) Dataset Composition

A total of 49 molecules were selected based on their documented detection status in interstellar environments: 25 detected and 24 undetected molecules. Each molecule was characterized using the following descriptors:

- Molecular Weight (MW)
- LogP (octanol-water partition coefficient, proxy for polarity)
- Hydrogen bond donors (HBD)
- Hydrogen bond acceptors (HBA)
- Topological Polar Surface Area (TPSA)
- Number of rotatable bonds
- Number of ring structures
- Number of heavy atoms

Detected molecules included small polar species such as Ammonia, Methanol, and Hydrogen cyanide, whereas undetected molecules tended to be larger, aromatic, or halogenated species, including Benzene, Toluene, and Iodobenzene.

B) Machine Learning Methodology

Logistic regression was employed to model the probability $P(\text{Detected})$ that a molecule is detectable in space:

$$P(\text{Detected}) = \frac{1}{1+e^{-z}}, z = \sum_i w_i x_i + b$$

where x_i represents each molecular descriptor and w_i the corresponding coefficient learned during model training. The model transforms a weighted linear combination of molecular features into a probabilistic measure of detectability, ranging from 0 (low probability) to 1 (high probability).

Data preprocessing:

- Features were standardized to zero mean and unit variance.
- The binary target variable: 1 for detected molecules, 0 for undetected molecules.
- Probabilities were computed for all molecules in the dataset.

Results

A) Predicted Detection Probabilities

The logistic regression model produced the following predicted probabilities for detectability:

Molecule	Detected	Predicted Probability			
Ammonia	1	0.99	Propynal	1	0.655
Hydrogen isocyanide	1	0.97	Methanethiol	1	0.65
Hydrogen cyanide	1	0.97	Glycolaldehyde	1	0.614

Water	1	0.961	Acetic acid	0	0.59
Methylamine	1	0.951	Cyclopropenylidene	1	0.528
Formaldehyde	1	0.93	Acetone	0	0.528
Isocyanic acid	1	0.918	Propanol	0	0.507
Methanol	1	0.913	Isopropanol	0	0.507
Carbon monoxide	1	0.907	Methyl acetate	0	0.289
Ethylene oxide	1	0.899	Tetrahydrofuran	0	0.284
Acetonitrile	1	0.857	Butanol	0	0.24
Formyl cyanide	1	0.835	Diethyl ether	0	0.157
Cyanoacetylene	1	0.835	tert-Butanol	0	0.152
Ketene	1	0.829	Ethyl acetate	0	0.1
Acetamide	1	0.814	Aniline	0	0.035
Acetaldehyde	1	0.799	Nitrobenzene	0	0.019
Vinyl alcohol	1	0.794	Benzene	0	0.018
Formamide	1	0.772	Methylaniline	0	0.014
Ethyl cyanide	1	0.75	Phenol	0	0.014
Propargyl alcohol	1	0.721	Acetanilide	0	0.007
Ethanol	0	0.661	Anisole	0	0.006
Propynal	1	0.655	Toluene	0	0.004
Methanethiol	1	0.65	Chlorobenzene	0	0.004
Glycolaldehyde	1	0.614	Styrene	0	0.002

Acetic acid	0	0.59	Phenylacetylene	0	0.001
Cyclopropenylidene	1	0.528	Xylene	0	0.001
Acetone	0	0.528	Bromobenzene	0	0.0003
			Iodobenzene	0	0.00003

B) Molecular Property Trends

Analysis of feature influence reveals several critical trends:

1. Molecular Weight:

   1.1 Detectable molecules are generally low molecular weight (<60 g/mol).

   1.2 Larger molecules, especially aromatic or halogenated species, are poorly detectable.

2. Polarity (LogP and TPSA):

   2.1 Detectable molecules exhibit moderate polarity (LogP ≤ 0.5; TPSA 17–40 Å²).

   2.2 Hydrophobic species show near-zero detectability probabilities.

3 Ring Structures:

   3.1 Acyclic molecules dominate high-probability detections.

   3.2 Ringed molecules are strongly correlated with undetectability.

4 Hydrogen Bonding and Rotatable Bonds:

   4.1 Detected molecules possess 0–2 hydrogen bond donors/acceptors and ≤1 rotatable bond, enhancing volatility and spectroscopic observability.

5 Heavy Atoms:

5.1 Fewer heavy atoms correspond to increased detectability; complex molecules are less detectable.

Summary Table:

Property	Detected (1)	Undetected (0)
MW	17–60	60–200+
LogP	-1.5 → 0.5	0.5 → 3+
HBD/HBA	0–2	0–2
TPSA	17–40	0–42
Rings	0	1–2+
Rotatable Bonds	0–1	1–3+
Heavy Atoms	1–4	5–12+

C) Implications for Astrochemical Discovery

- Small, polar, acyclic molecules are optimal candidates for detection.

- Machine learning probabilistic outputs enable prioritization of previously undetected molecules for observational campaigns.
- Intermediate-probability molecules suggest candidates that could be conditionally detectable, depending on abundance and observational sensitivity.

Discussion

Logistic regression provides an interpretable framework to quantify detection likelihood. The results reinforce the chemical intuition that molecular simplicity and polarity enhance detectability in interstellar environments. Despite the limited dataset, the model highlights robust trends:

- Detection probability is a multivariate function of molecular features.
- Feature importance can guide future molecular target selection.
- The model may be extended to larger datasets or more sophisticated algorithms for enhanced predictive accuracy.

Conclusion

This study demonstrates that machine learning can reliably model molecular detectability in astrochemical contexts. Predicted probabilities reveal that small, polar, acyclic molecules dominate interstellar detections, whereas larger, aromatic, or hydrophobic species are largely undetected. The logistic regression framework provides a probabilistic roadmap for selecting novel molecular targets, enabling strategic prioritization for future astrochemical observations.

References

- Herbst, E., & van Dishoeck, E. F. (2009). Complex Organic Interstellar Molecules. *Annual Review of Astronomy and Astrophysics, 47*, 427–480.
- McGuire, B. A. (2018). 2018 Census of Interstellar, Circumstellar, Extragalactic, Protoplanetary Disk, and Exoplanetary Molecules. *ApJS, 239*(2), 17.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.