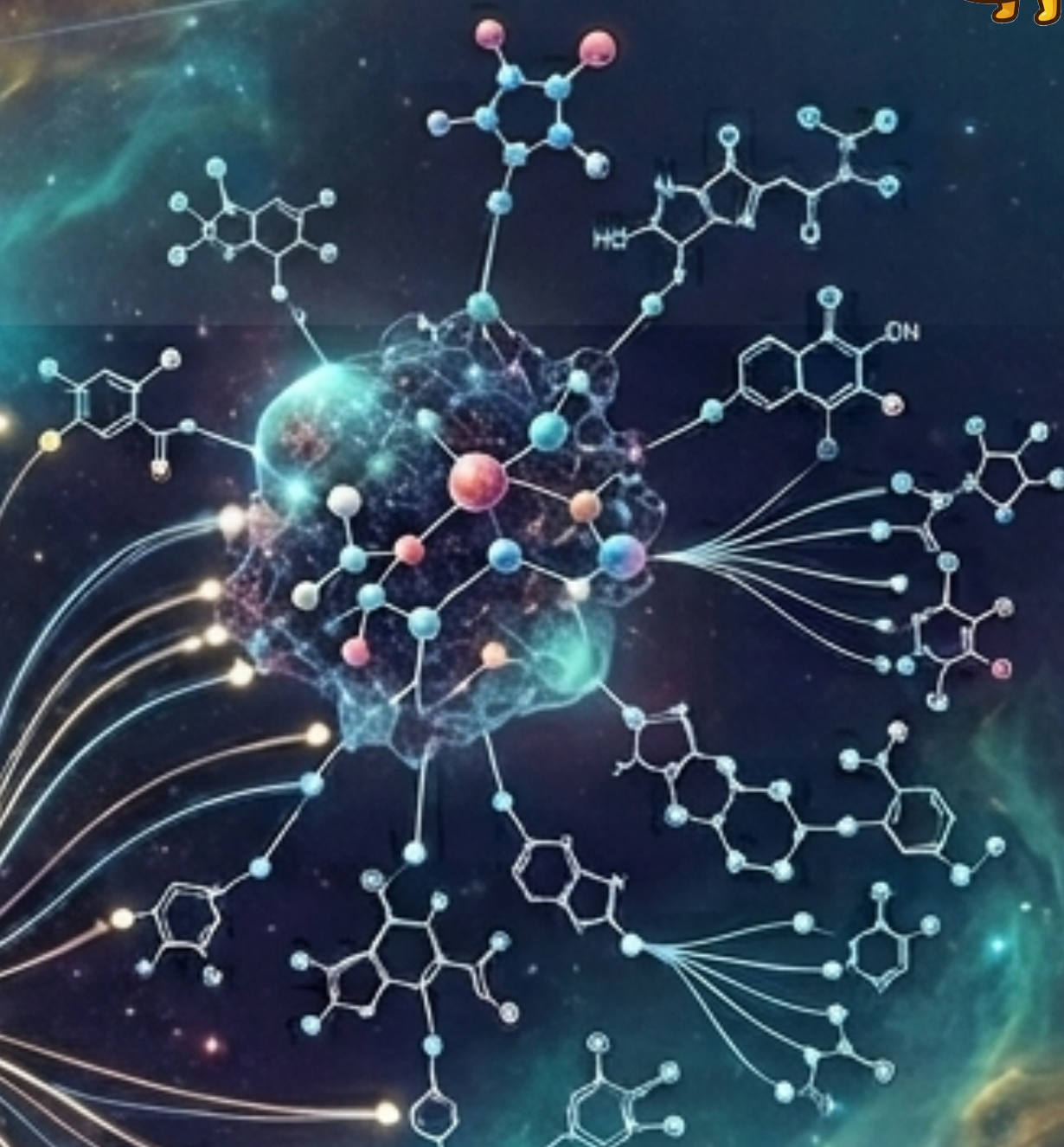


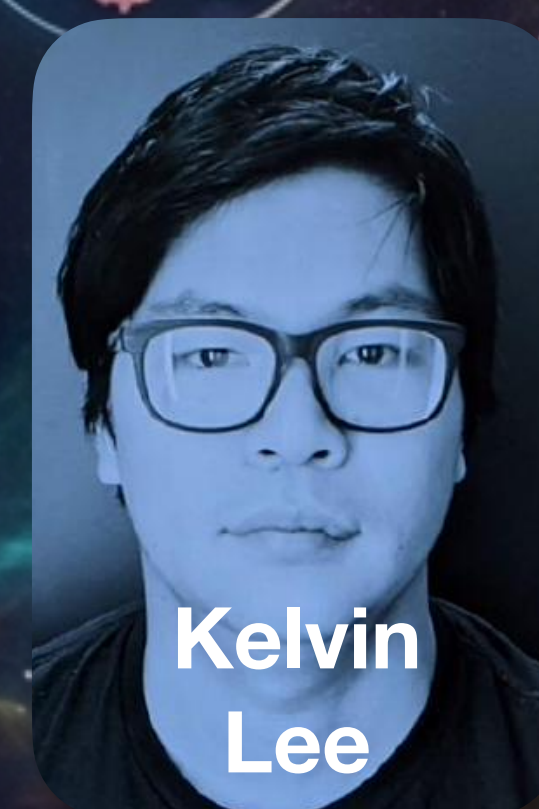


AstroAMASE: Leveraging Machine-Learned Chemical Intuition for Molecular Assignment and Discovery

Brett A. McGuire



**Zach
Fried**



**Kelvin
Lee**



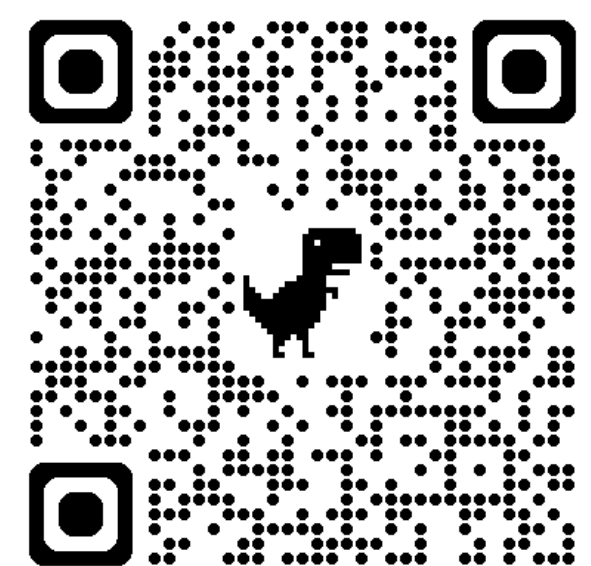
**Ryan
Loomis**



**Andy
Lipnicky**

+ Rich Teague & Leah Albrow

Fried & McGuire 2024
J Phys Chem A 128, 8254





AstroAMASE: Leveraging Machine-Learned Chemical Intuition for Molecular Assignment and Discovery

Brett A. McGuire



**Zach
Fried**



**Kelvin
Lee**



**Ryan
Loomis**



**Andy
Lipnicky**

+ Rich Teague & Leah Albrow

Fried & McGuire 2024
J Phys Chem A 128, 8254

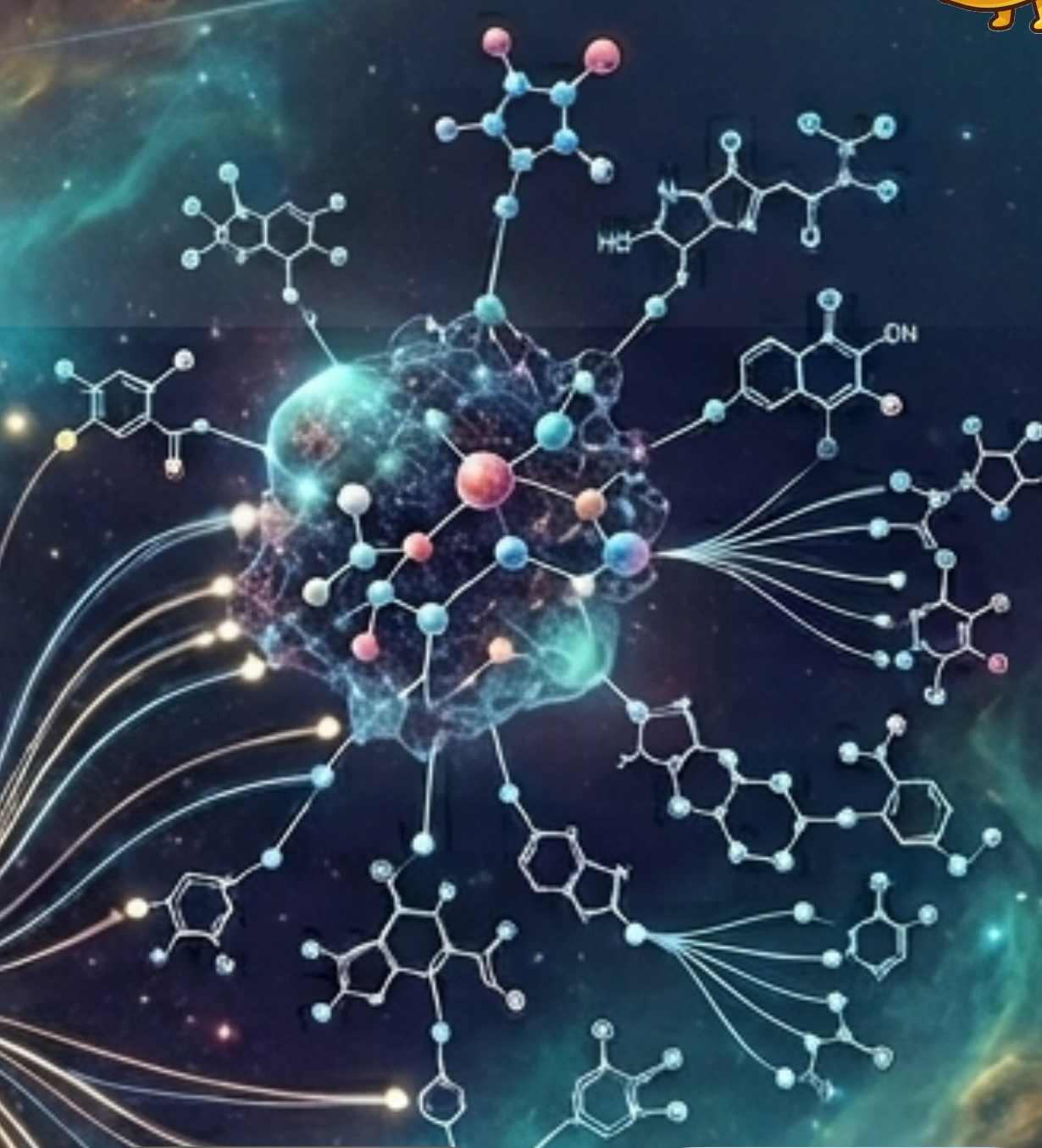




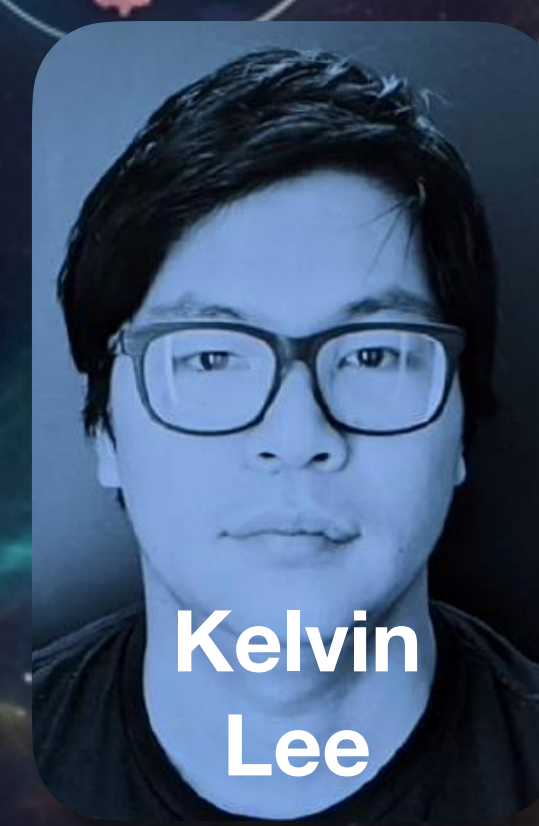
AstroAMASE: Leveraging Machine-Learned Chemical Intuition for Molecular Assignment and Discovery

Brett A. McGuire

Presidort of astrochemistry



**Zach
Fried**



**Kelvin
Lee**



**Ryan
Loomis**



**Andy
Lipnicky**



+ Rich Teague & Leah Albrow

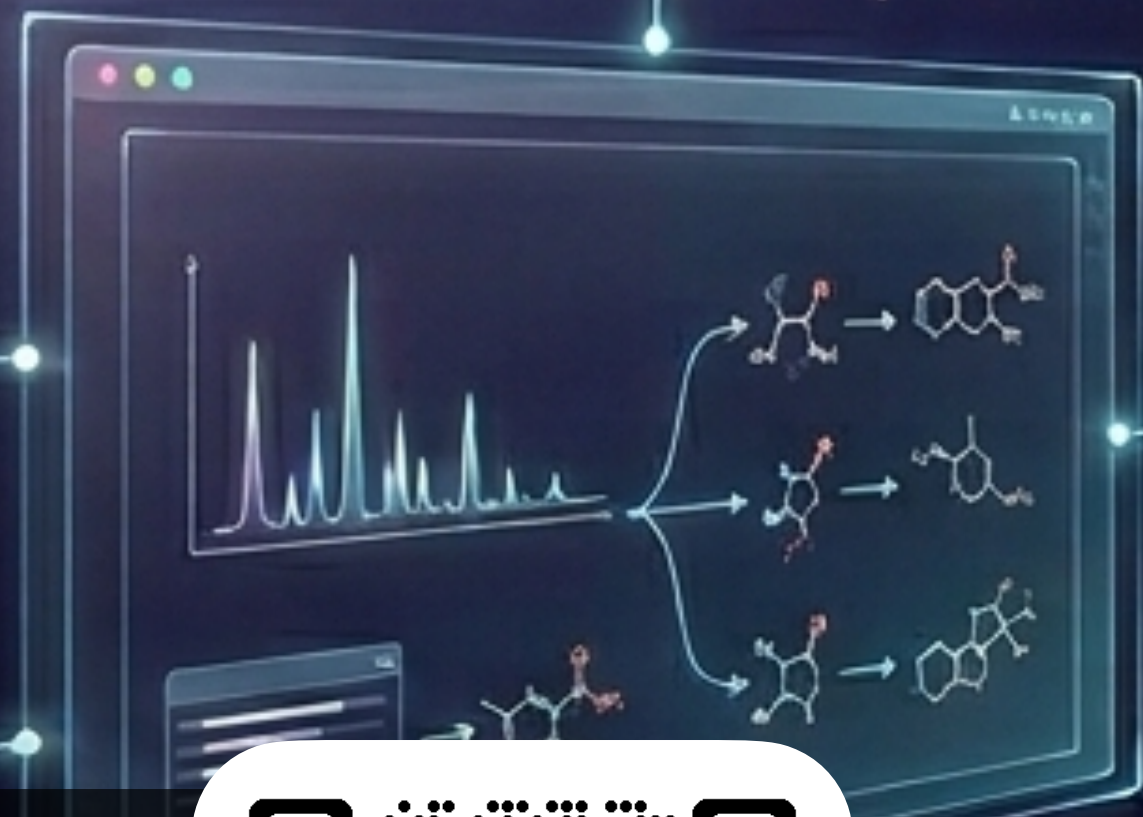
Fried & McGuire 2024
J Phys Chem A 128, 8254



AstroAMASE: Leveraging Machine-Learned Chemical Intuition for Molecular Assignment and Discovery

Brett A. McGuire

Presidort of astrochemistry



**Zach
Fried**



**Kelvin
Lee**



**Ryan
Loomis**



**Andy
Lipnicky**

+ Rich Teague & Leah Albrow

Fried & McGuire 2024
J Phys Chem A 128, 8254



Astrochemistry *astrō'keməstrē* (n)

The study of molecules in space:
where they are, how they got there,
and what they are doing.

NGC 6334I
'Cat's Paw Nebula'



NGC 6334I
'Cat's Paw Nebula'

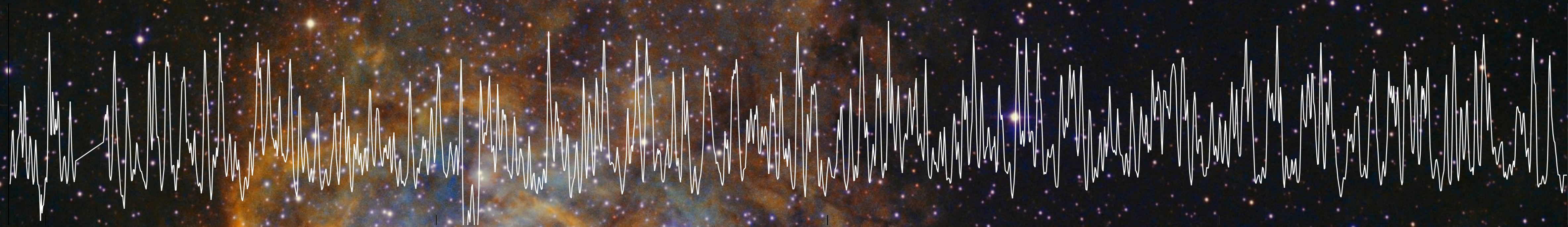
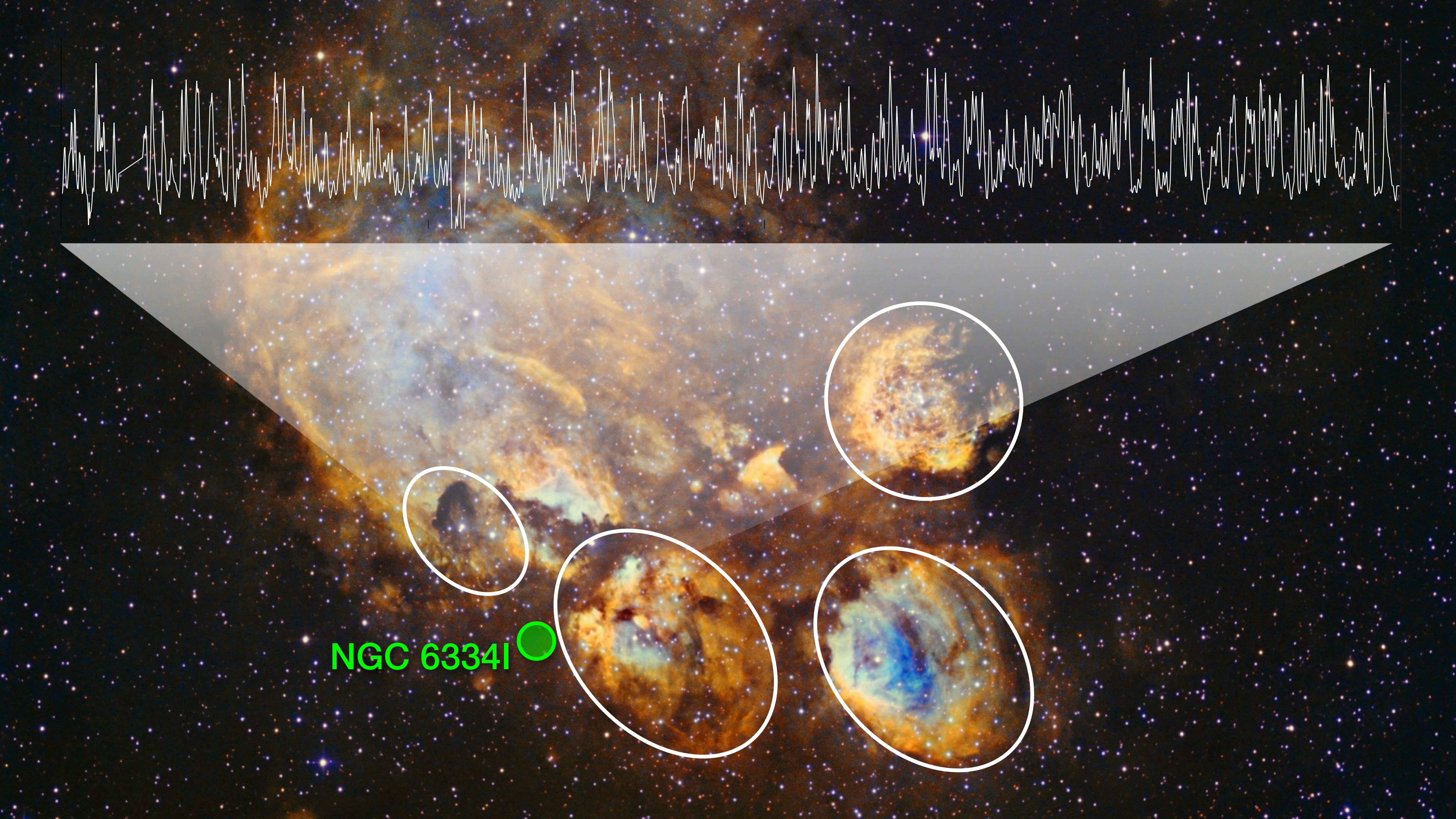
NGC 6334I



NGC 6334I
'Cat's Paw Nebula'

NGC 6334I

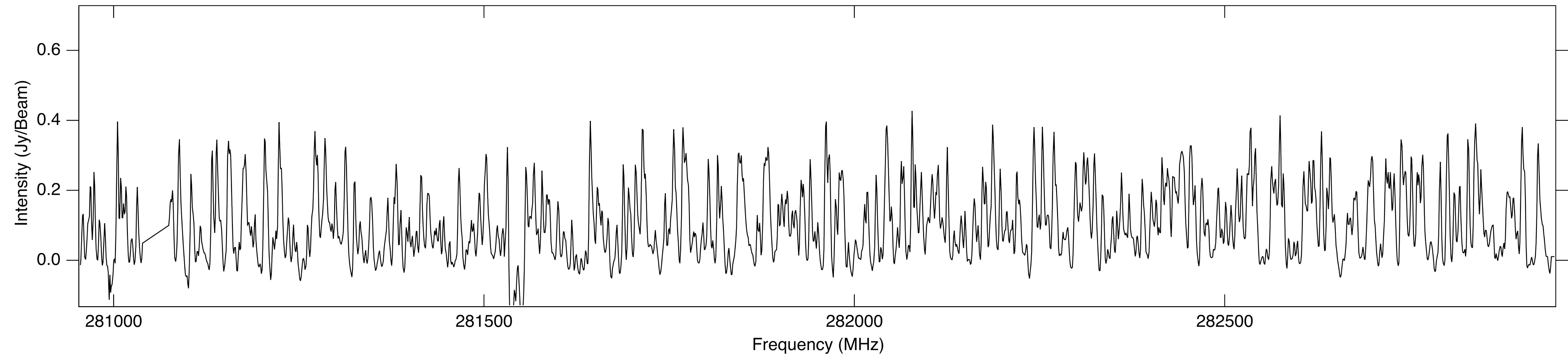




NGC 6334I ●

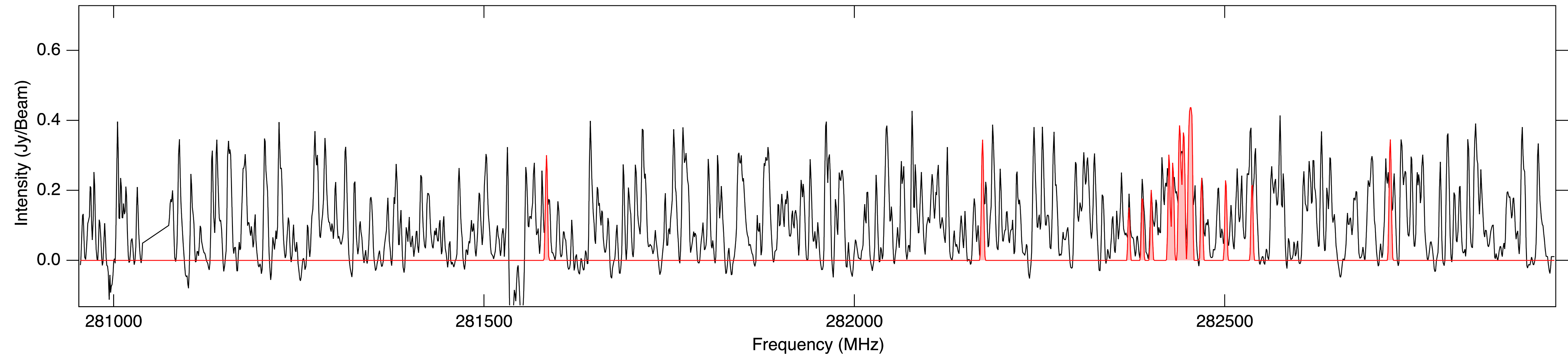
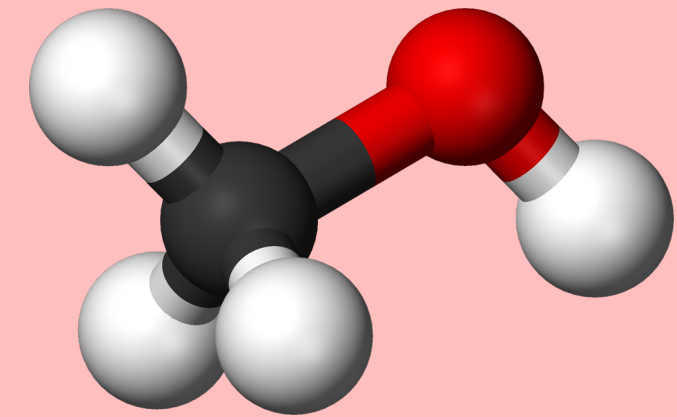
MOLECULAR IDENTIFICATION

MOLECULAR IDENTIFICATION



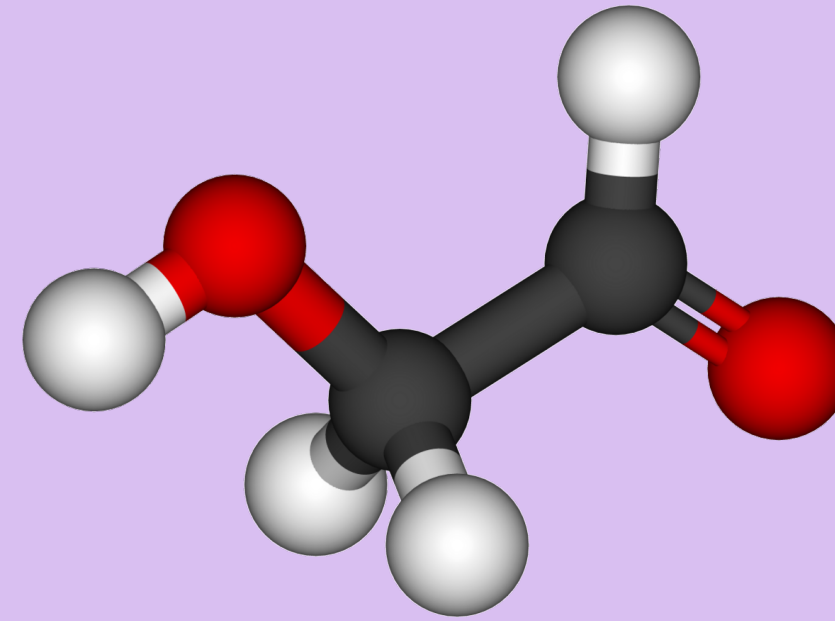
MOLECULAR IDENTIFICATION

(^{13}C) METHANOL

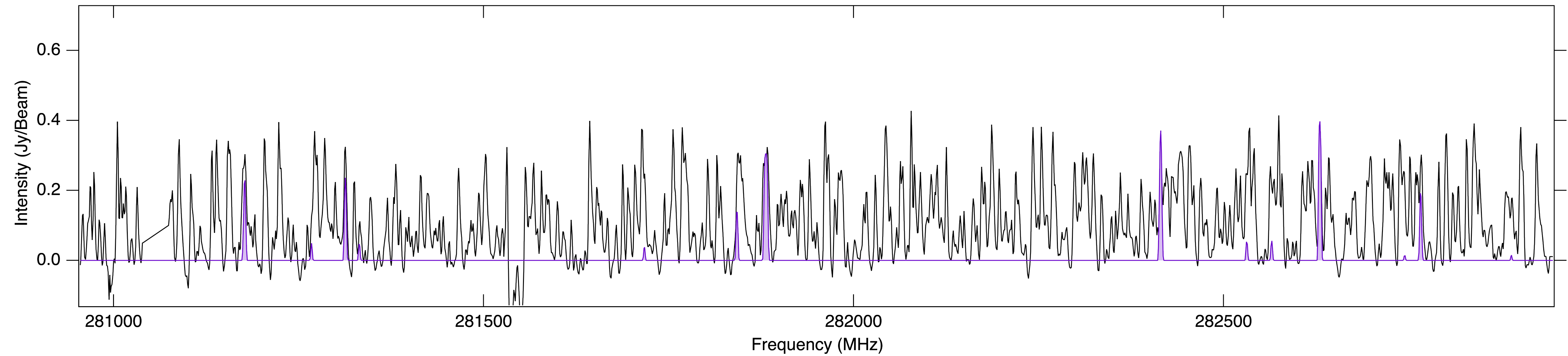
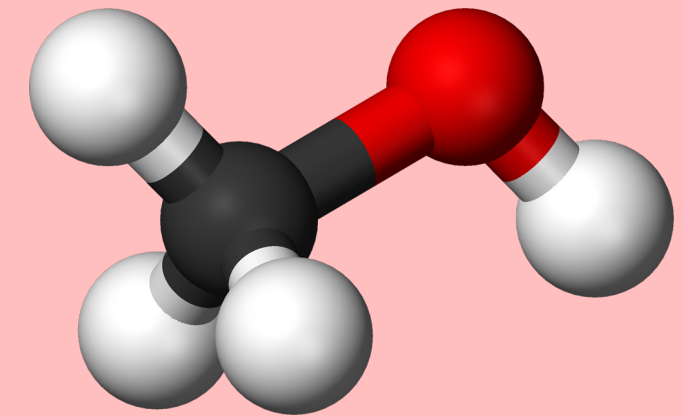


MOLECULAR IDENTIFICATION

GLYCOLALDEHYDE

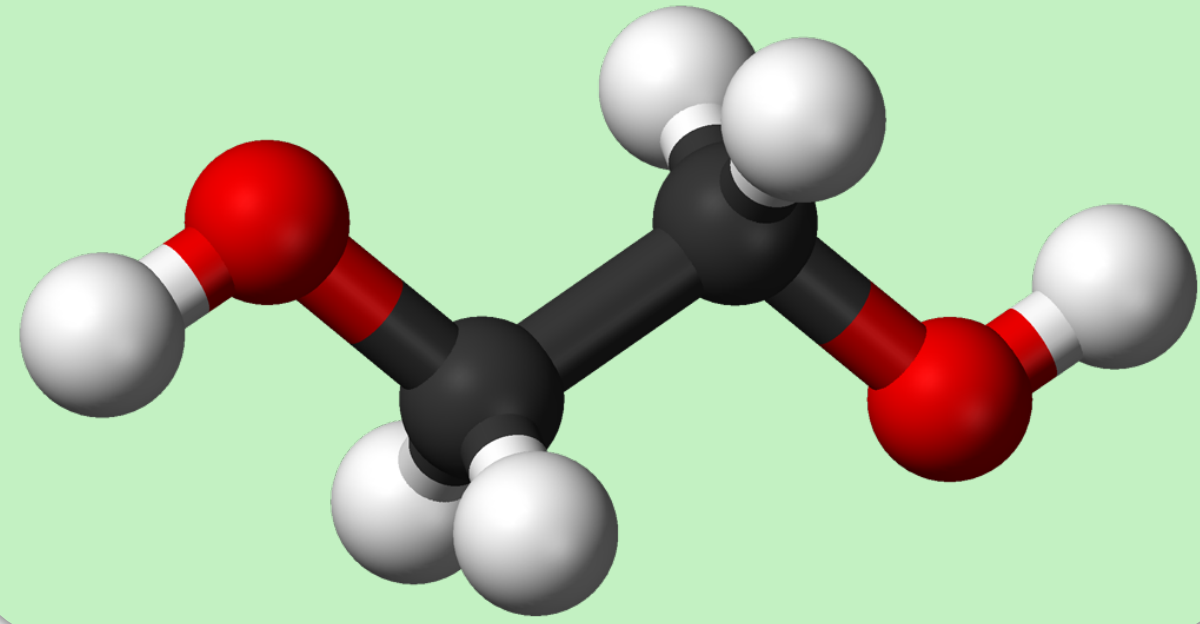


(¹³C) METHANOL

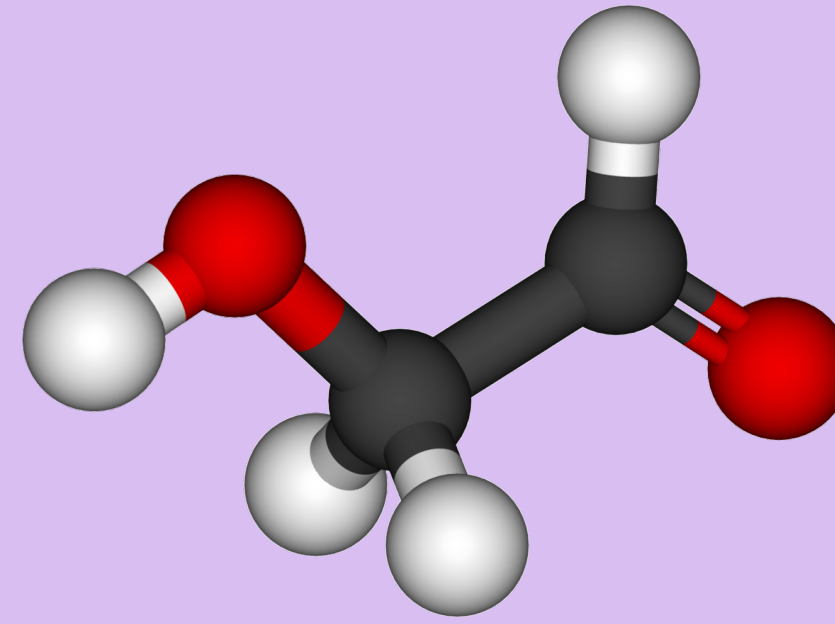


MOLECULAR IDENTIFICATION

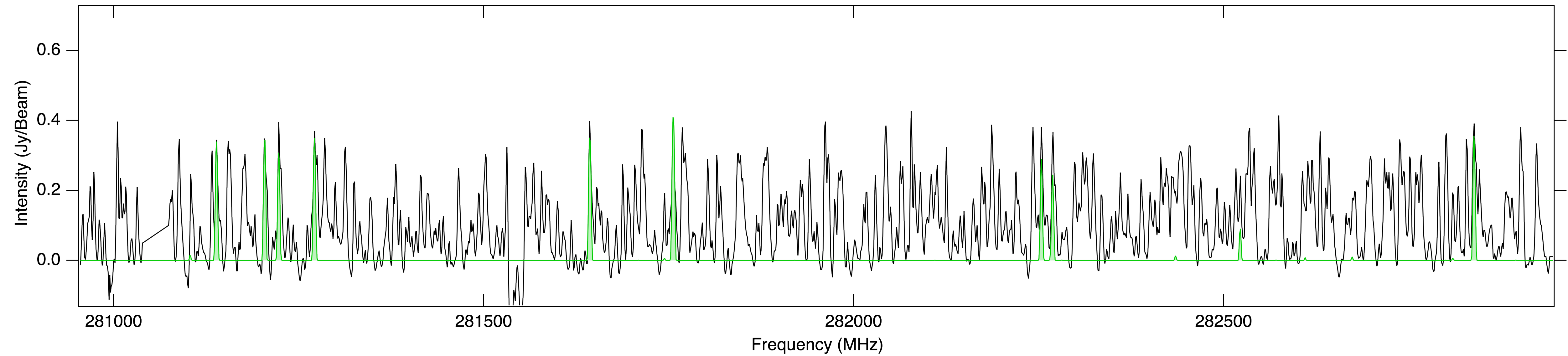
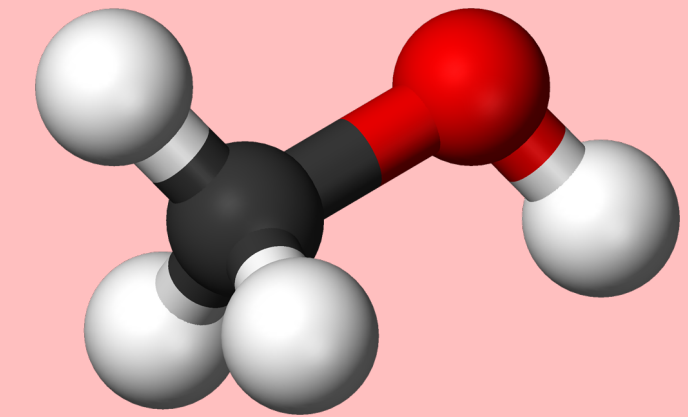
ETHYLENE GLYCOL



GLYCOLALDEHYDE

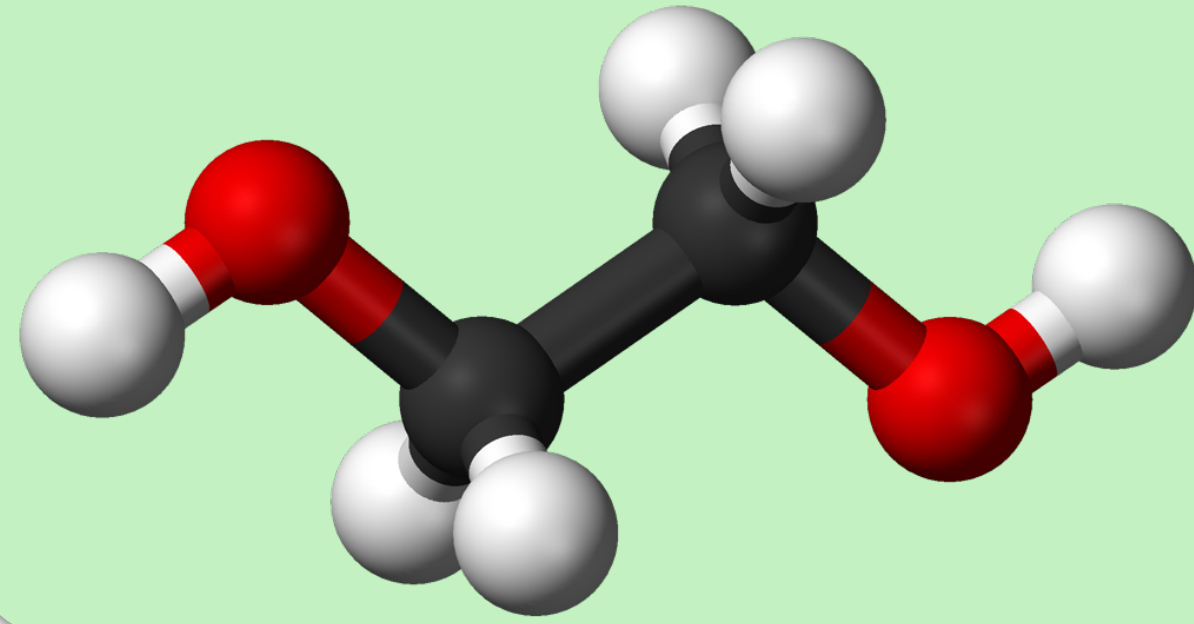


(¹³C) METHANOL

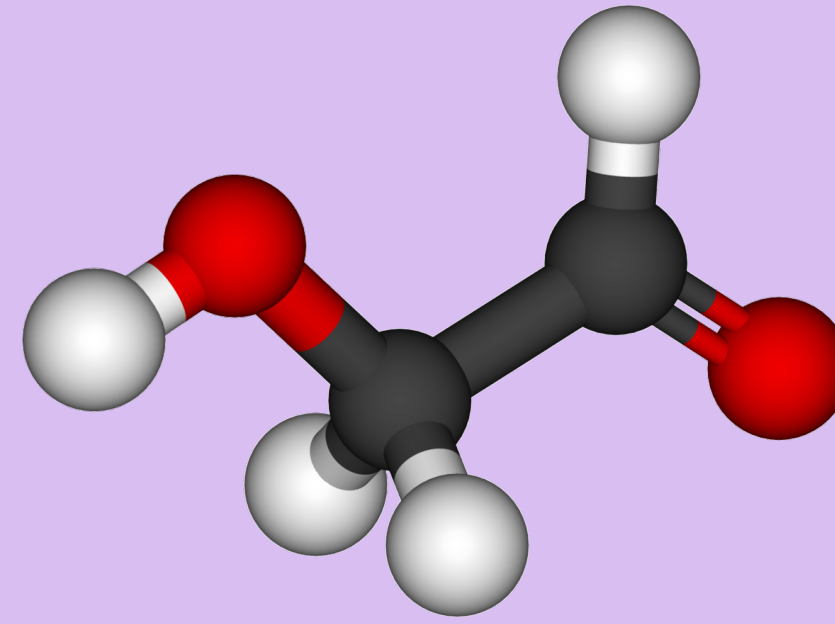


MOLECULAR IDENTIFICATION

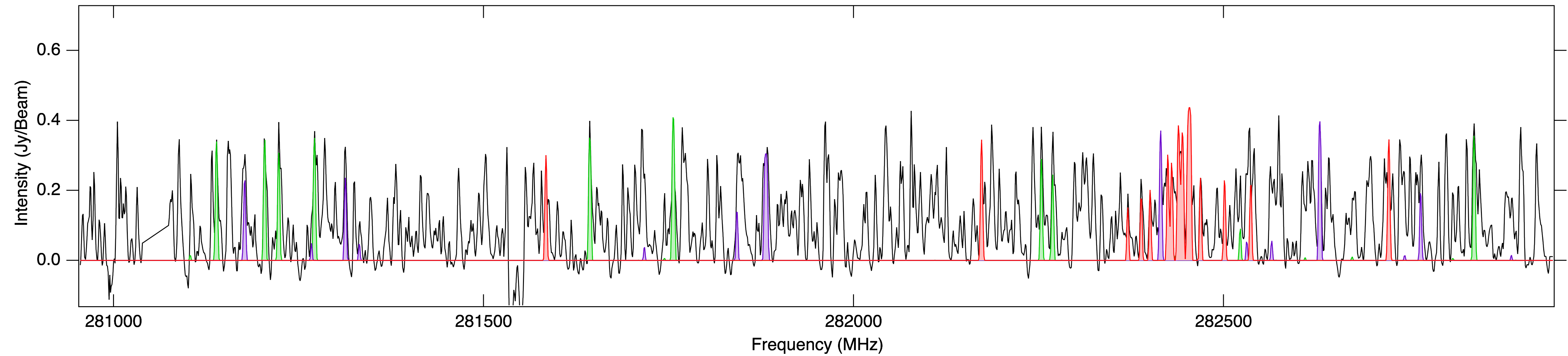
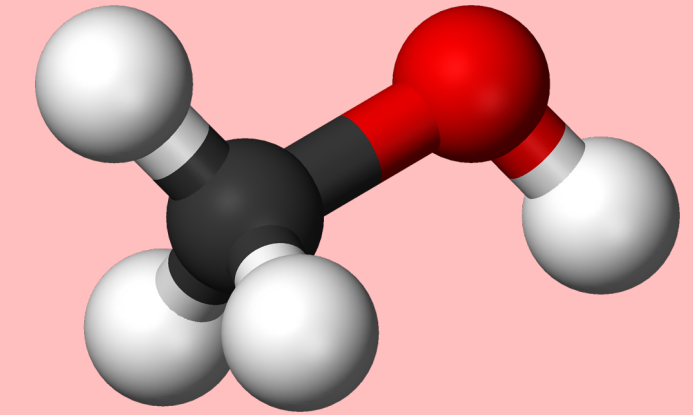
ETHYLENE GLYCOL

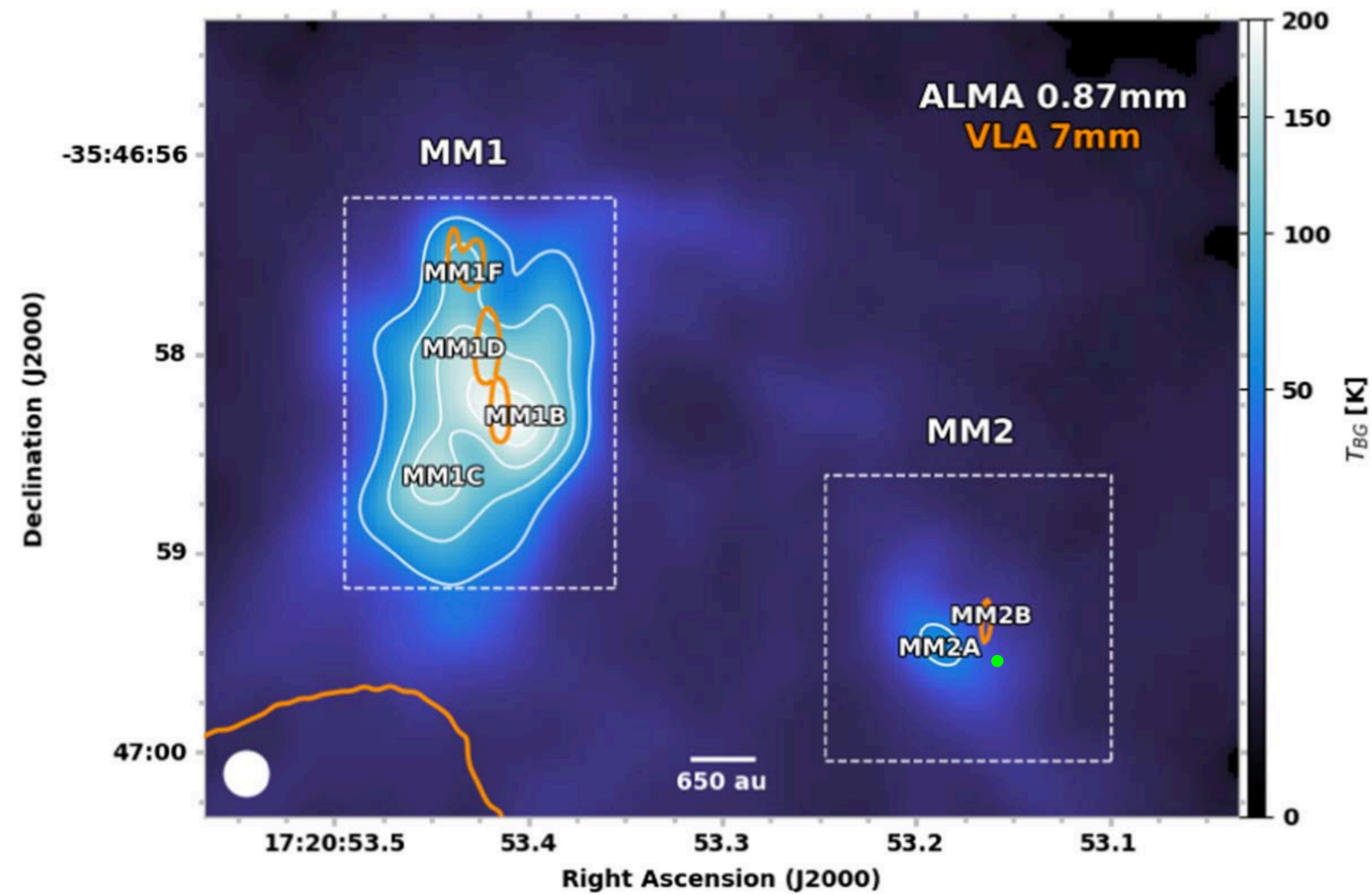


GLYCOLALDEHYDE

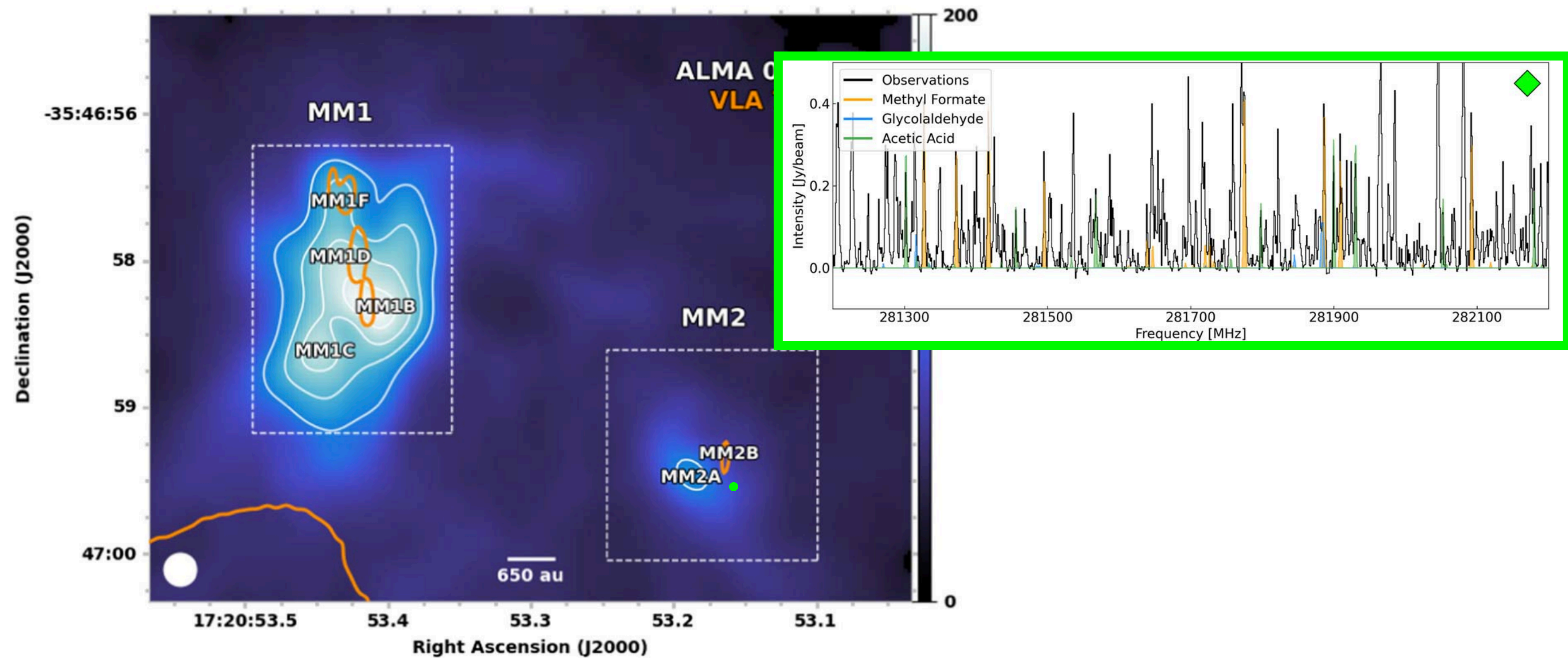


(¹³C) METHANOL

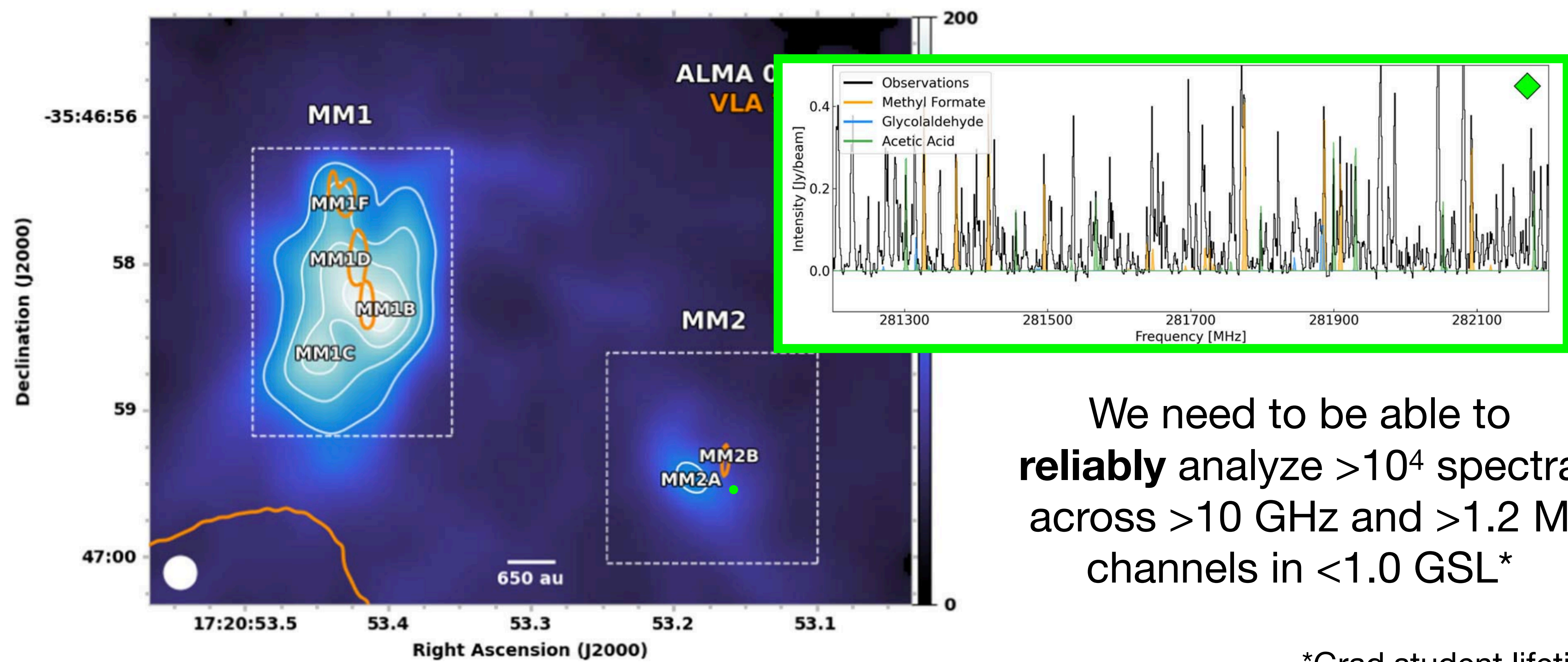




THE PROBLEM



THE PROBLEM

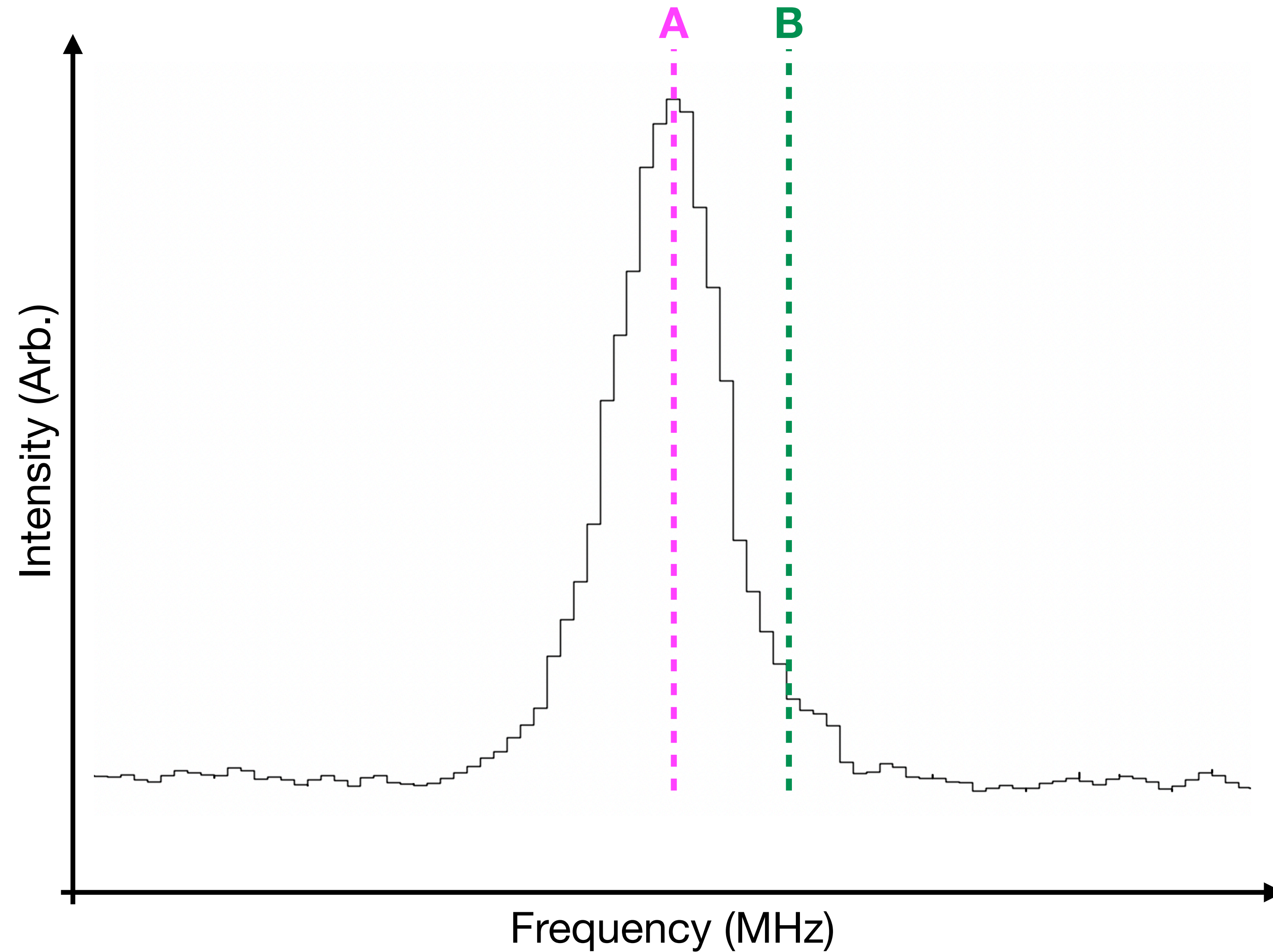


We need to be able to **reliably** analyze $>10^4$ spectra across >10 GHz and >1.2 M channels in <1.0 GSL*

*Grad student lifetimes

PICKING APART A SPECTRUM

Fried & McGuire 2024 *JPCA* 128, 8254

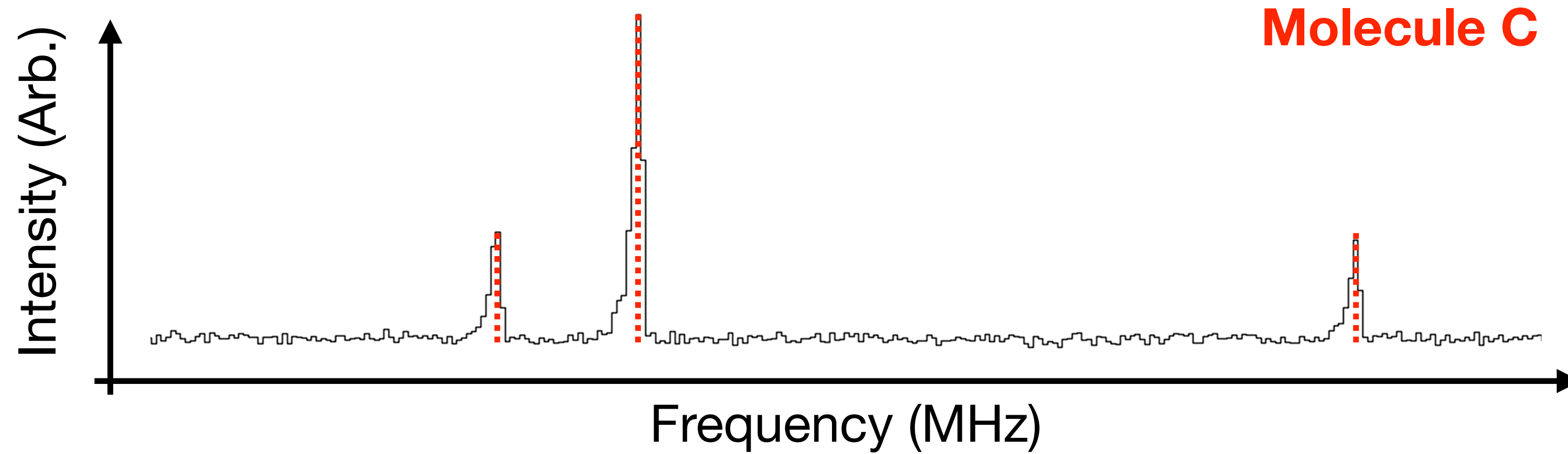


Go line-by-line and
check for closest
match in database

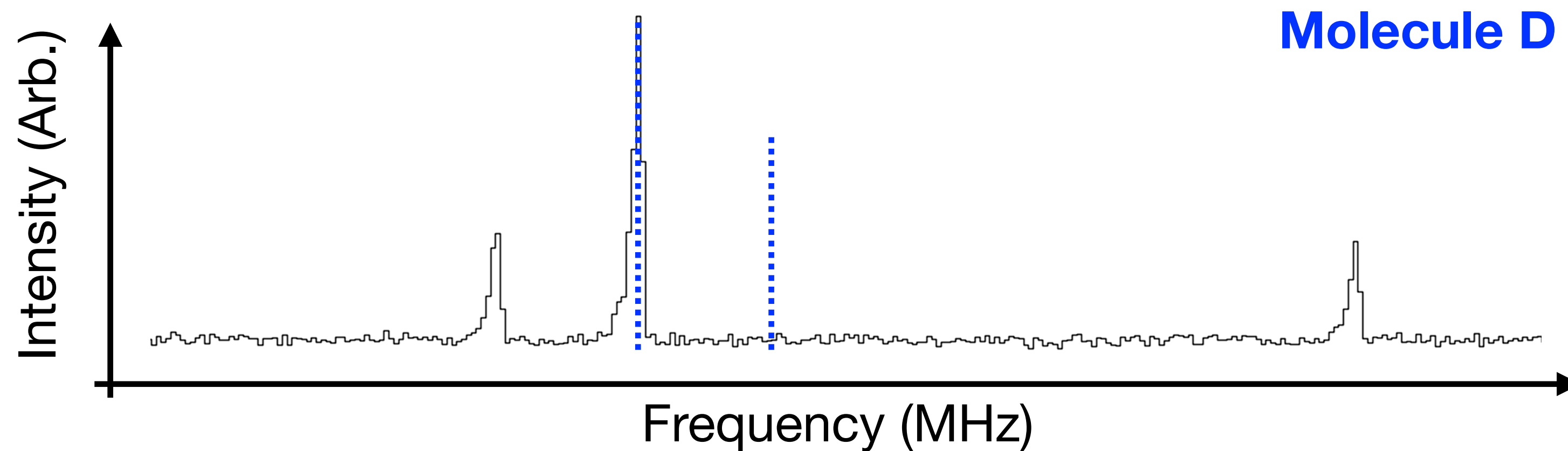
Trivial conventional
computational
problem

PICKING APART A SPECTRUM

Fried & McGuire 2024 *JPCA* 128, 8254



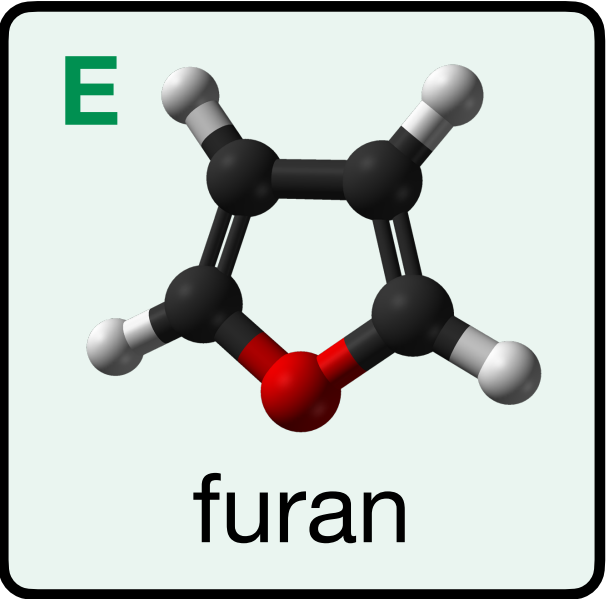
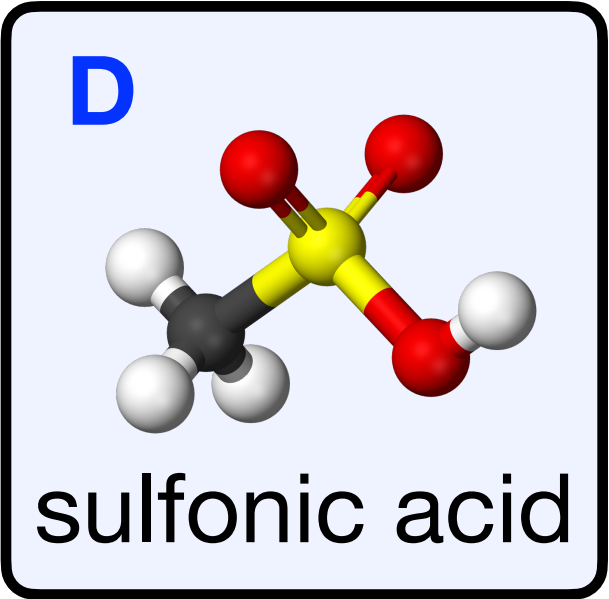
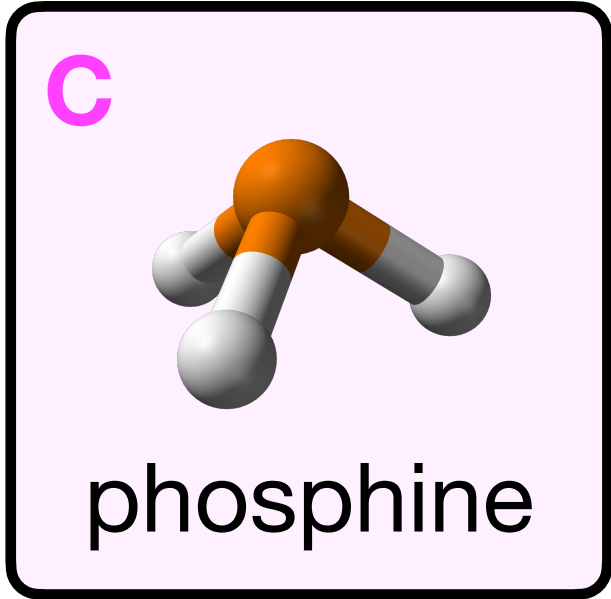
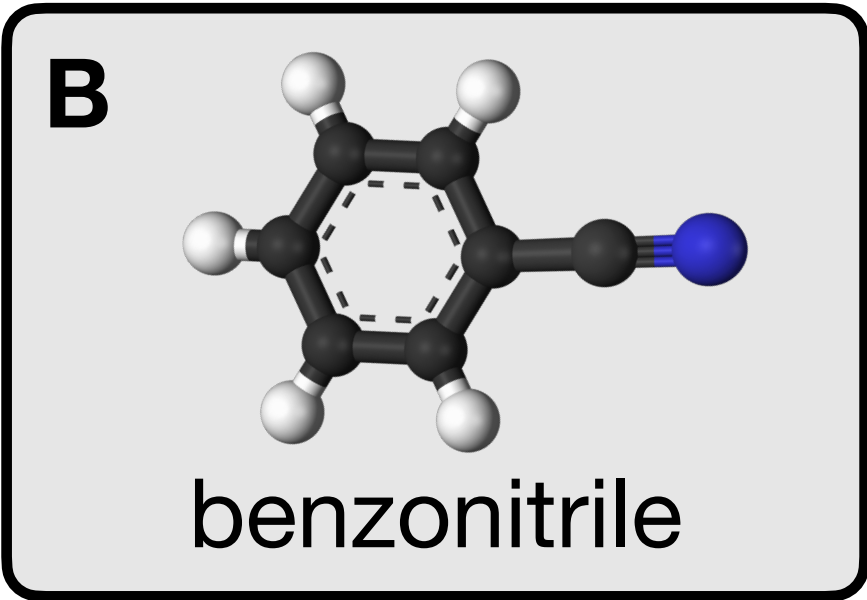
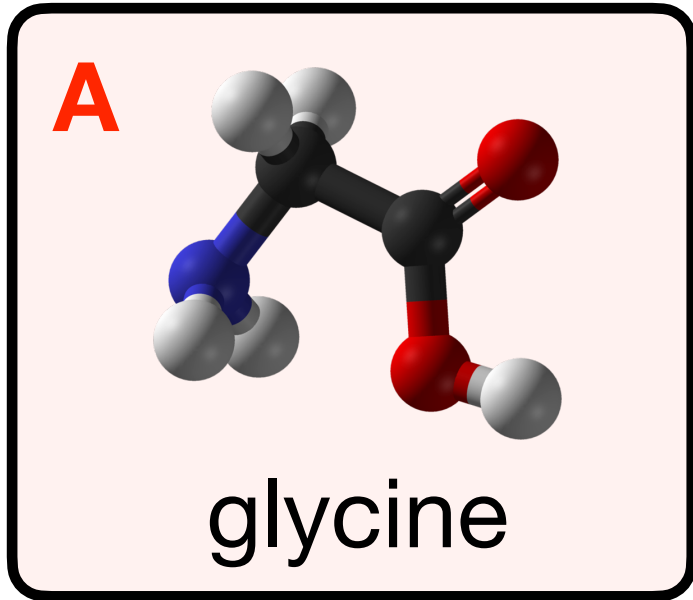
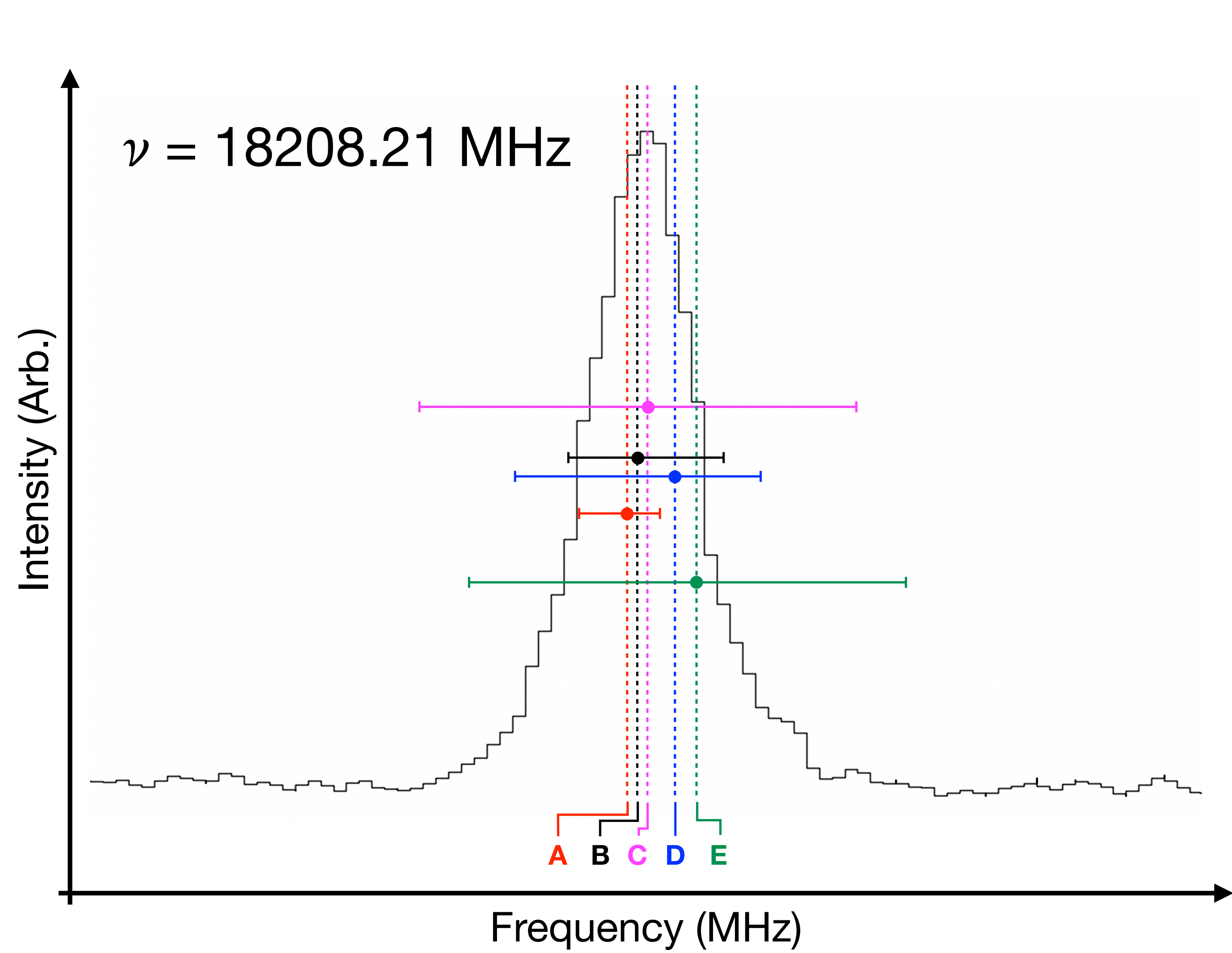
Check for other
confirming lines in the
spectrum



~trivial conventional
computational
problem

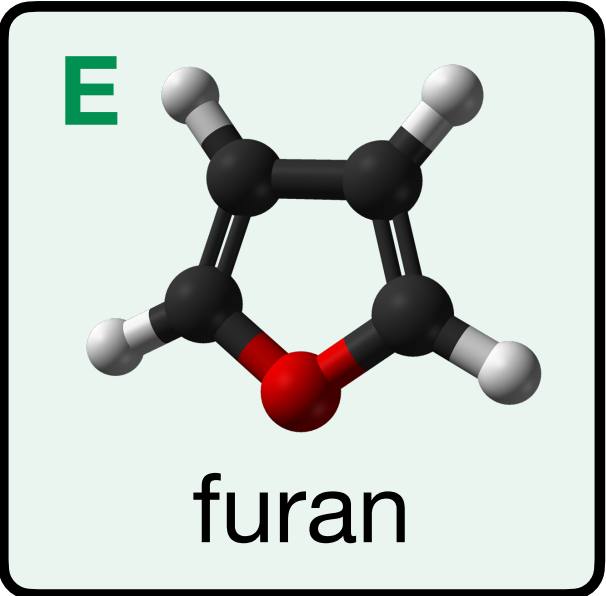
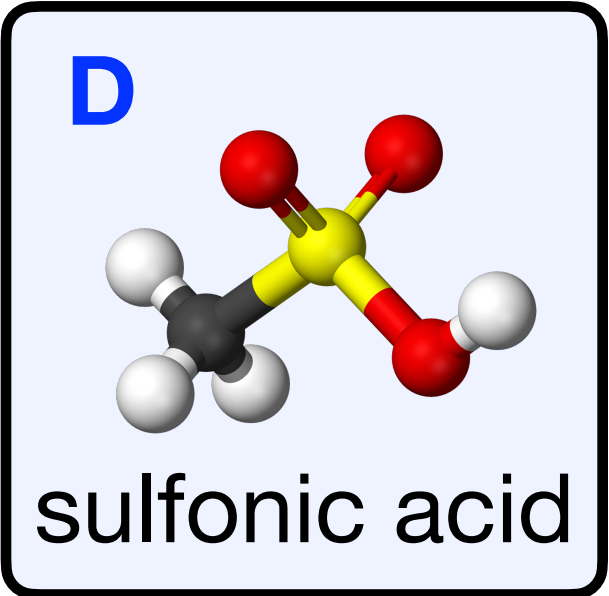
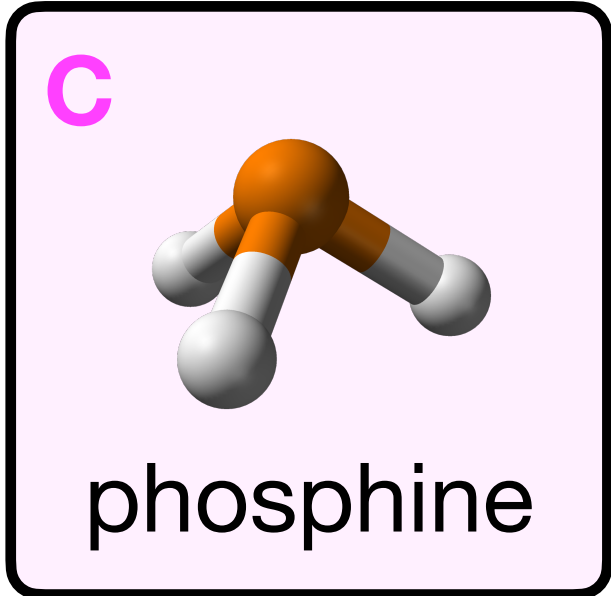
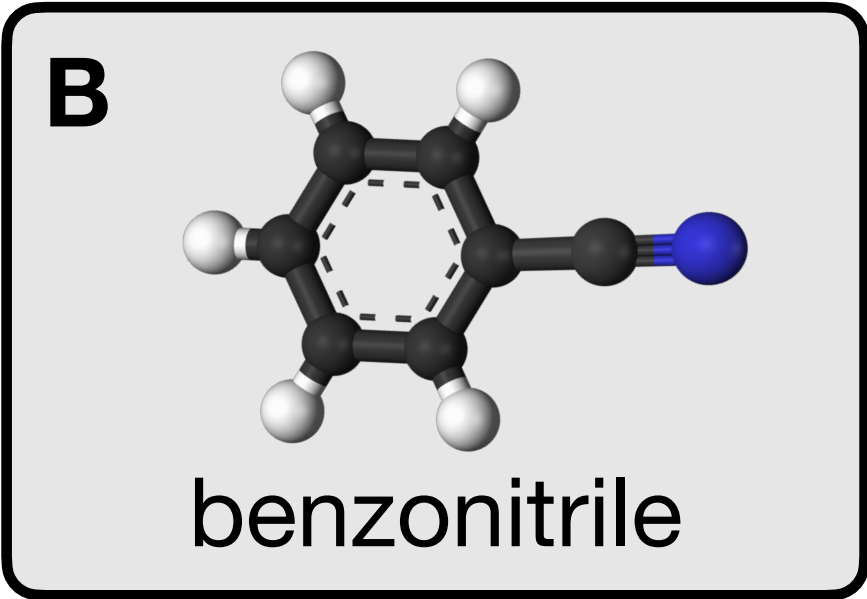
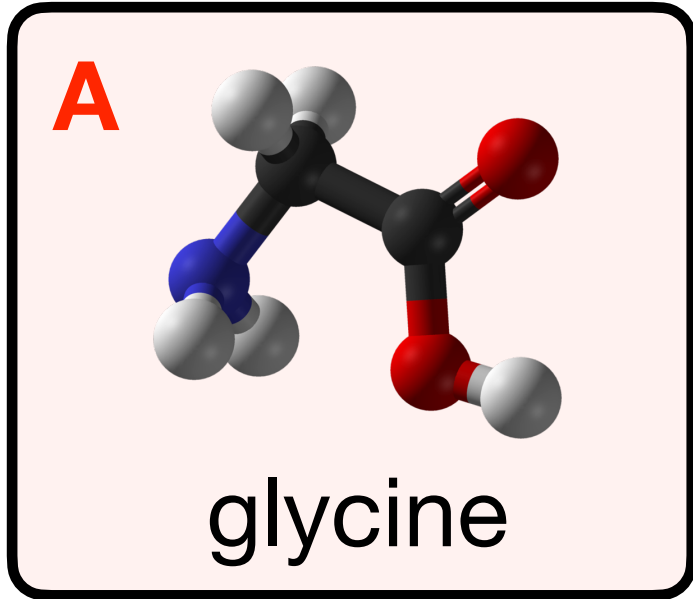
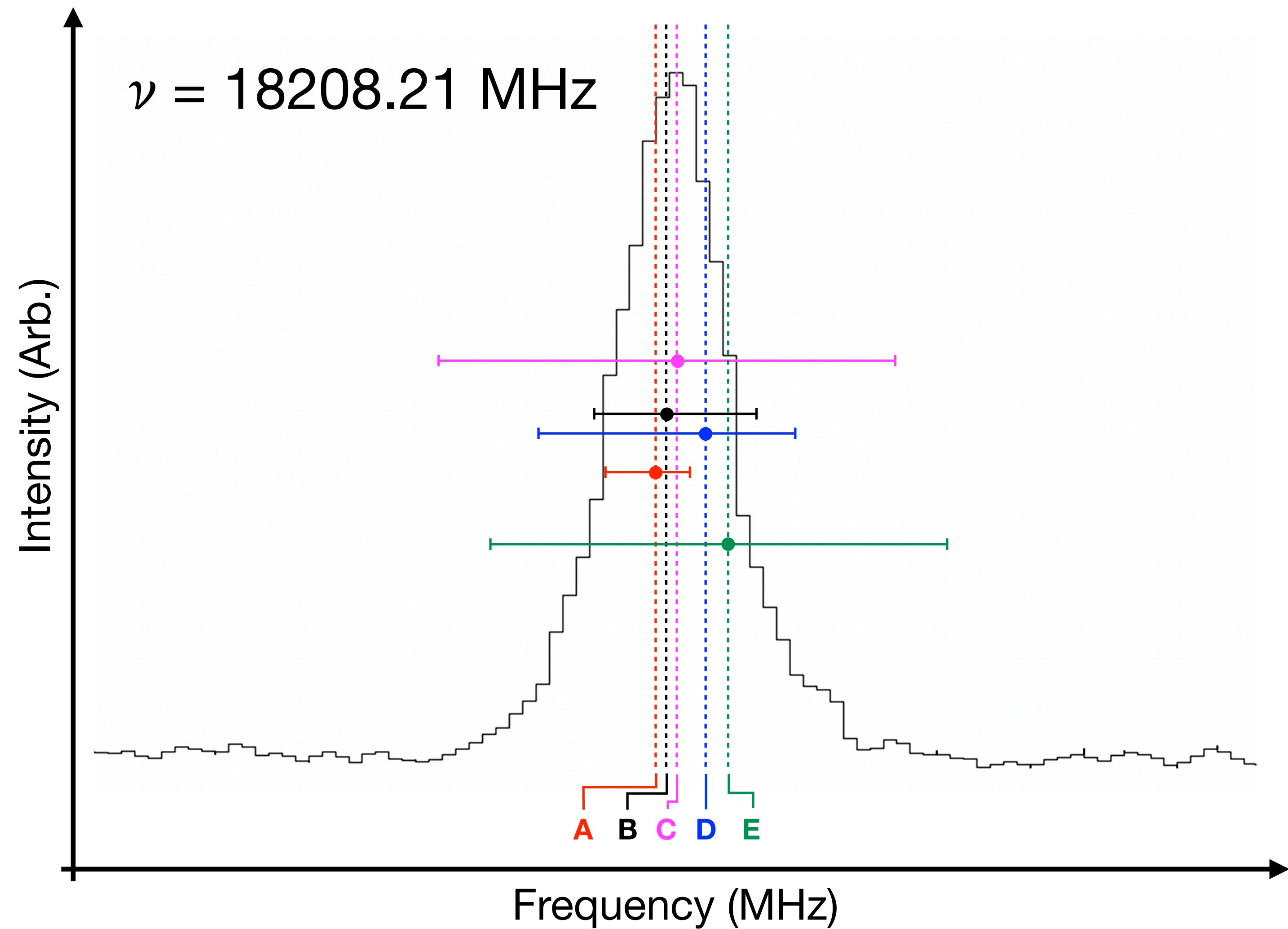
A PRACTICAL EXAMPLE

Fried & McGuire 2024 *JPCA* 128, 8254



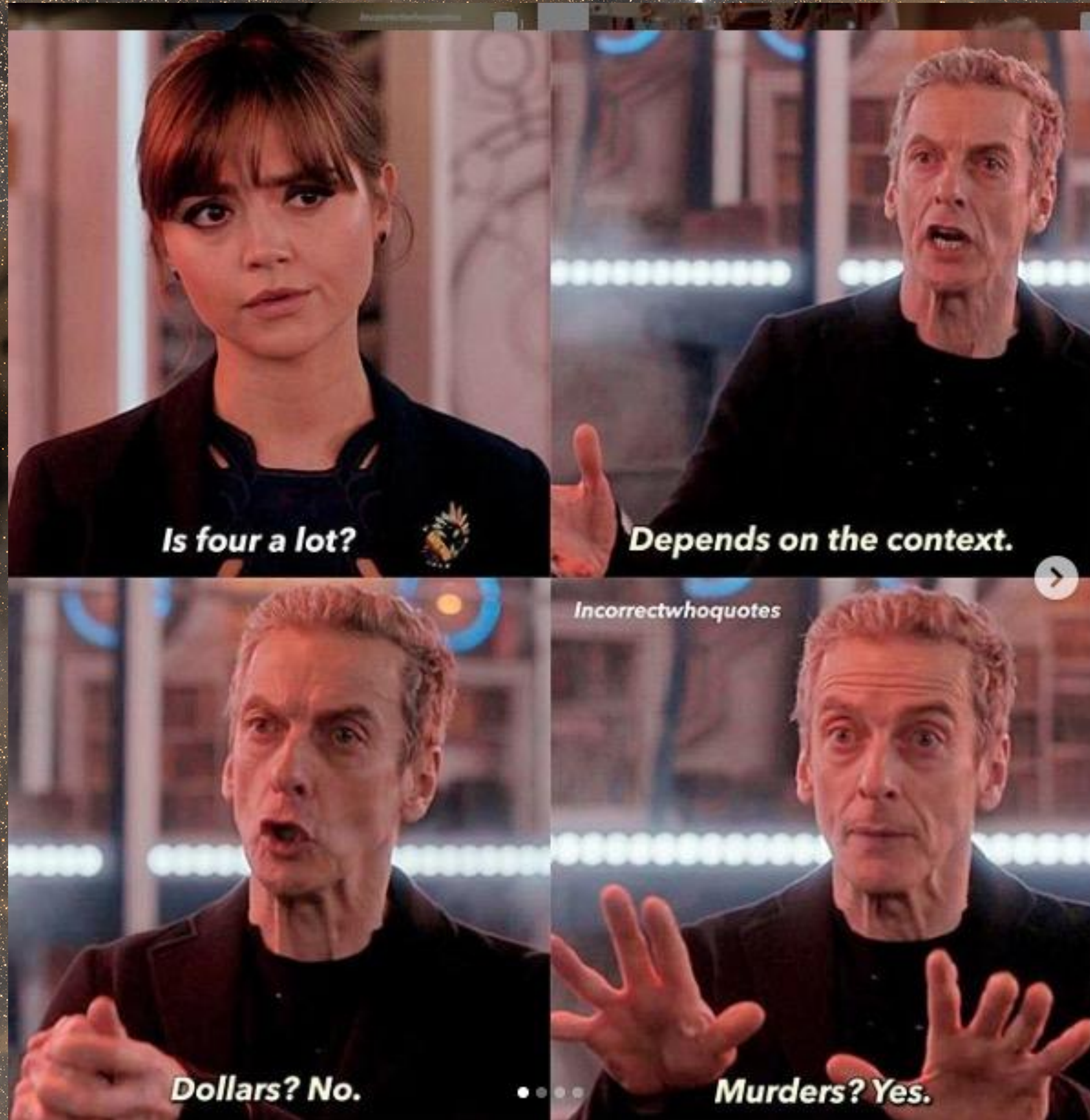
Candidate	Catalog - Exp.
A	0.201 MHz
B	0.121 MHz
C	0.020 MHz
D	0.356 MHz
E	0.549 MHz

A PRACTICAL EXAMPLE



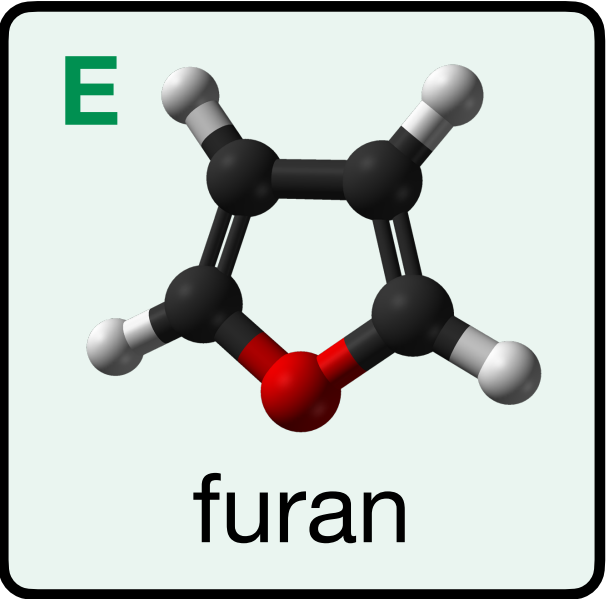
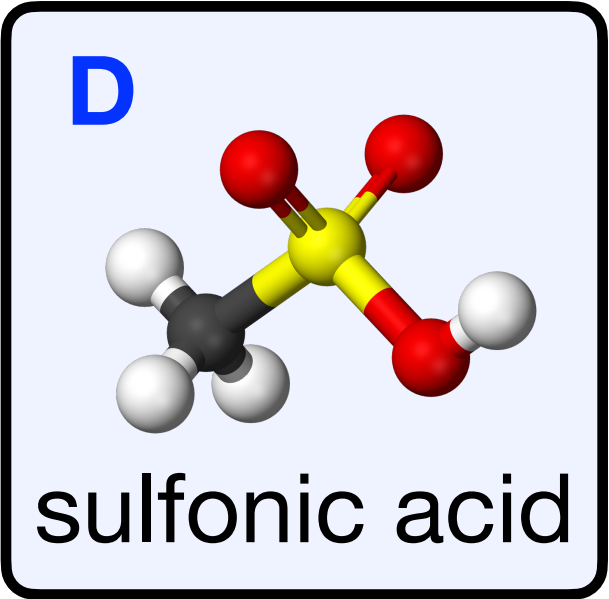
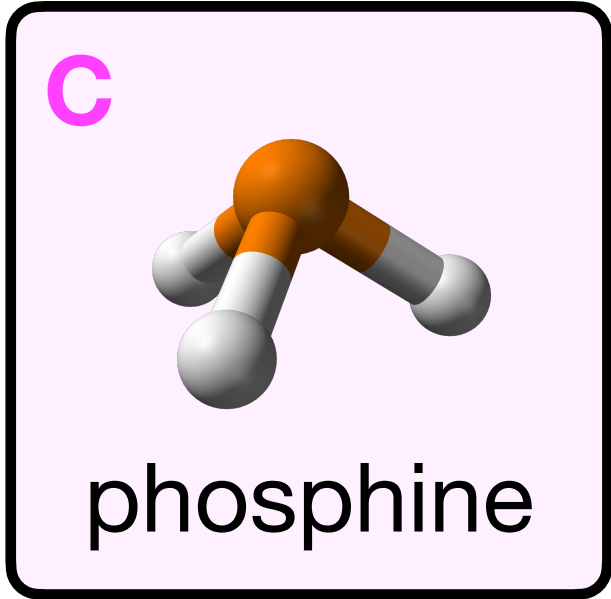
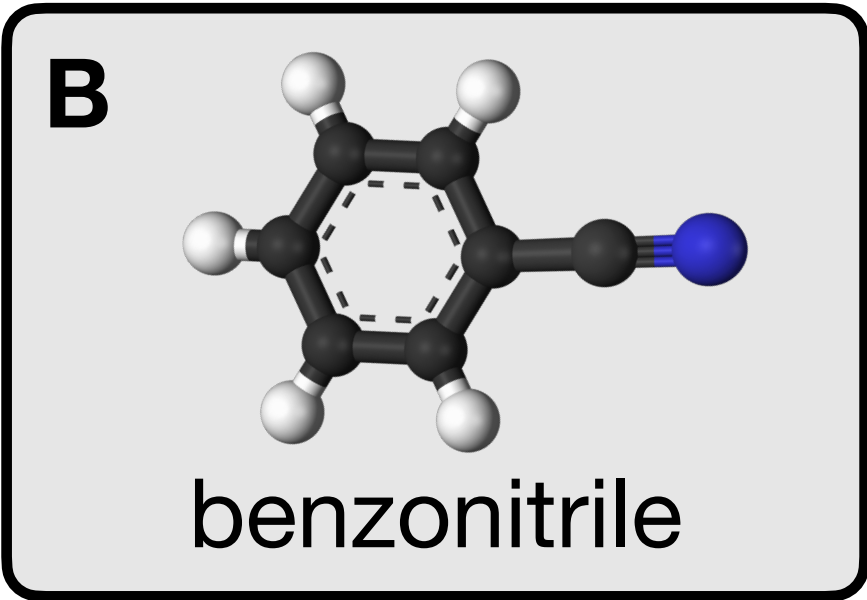
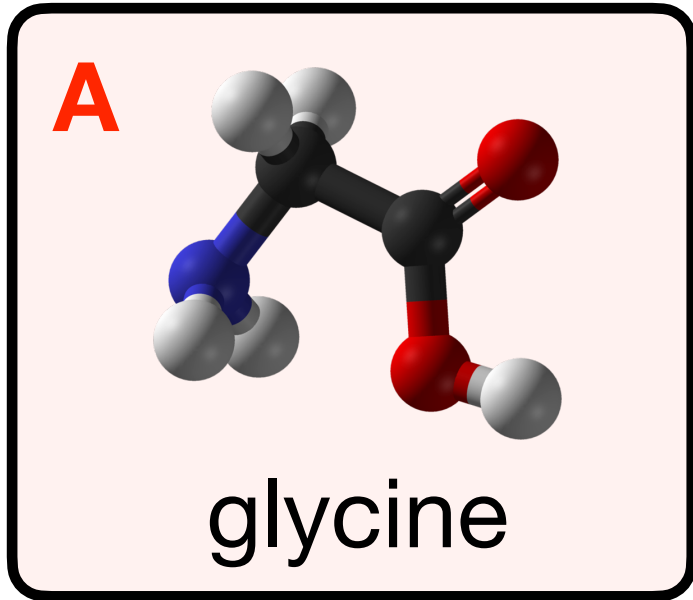
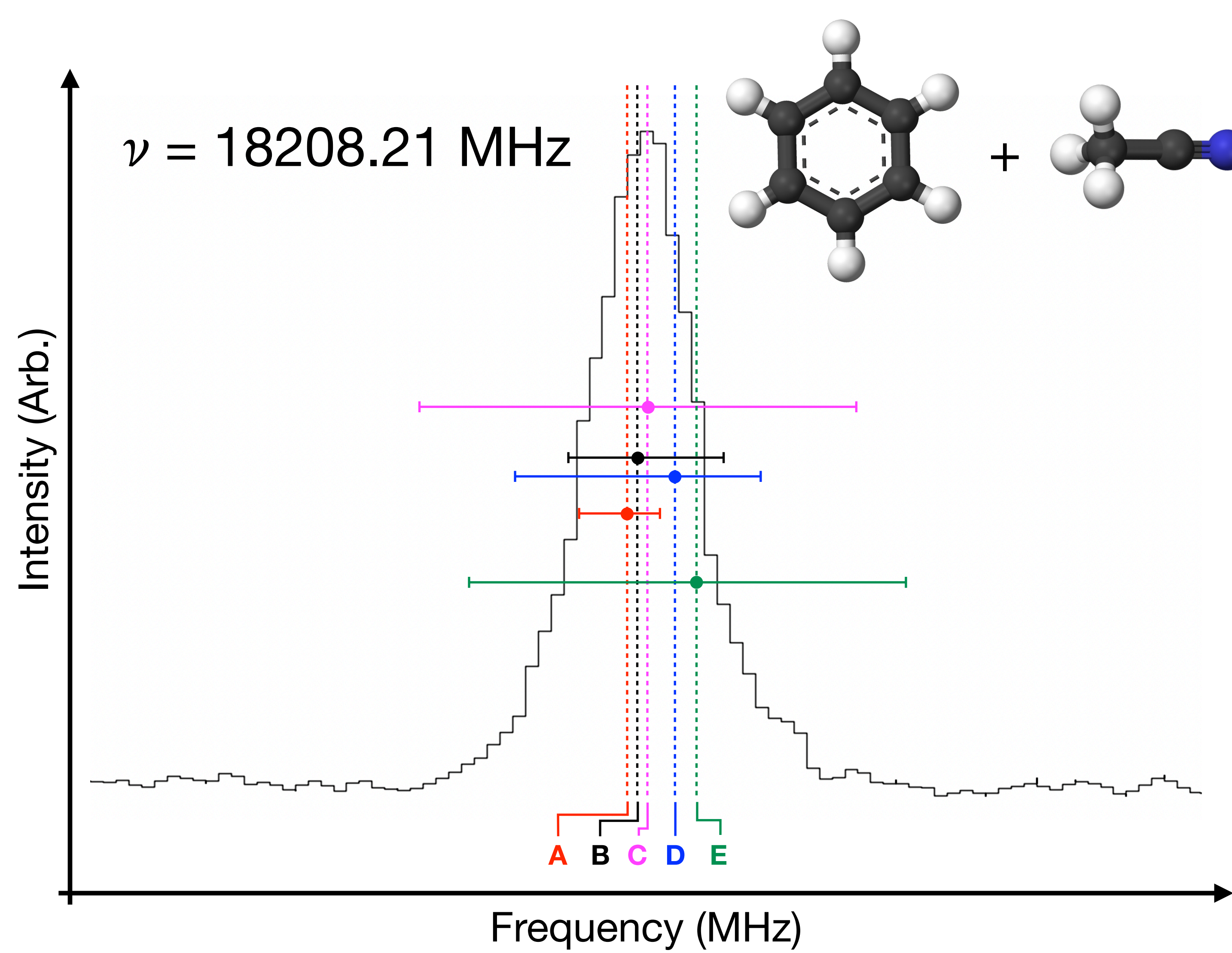
Candidate	Catalog - Exp.
A	0.201 MHz
B	0.121 MHz
C	0.020 MHz
D	0.356 MHz
E	0.549 MHz





A PRACTICAL EXAMPLE

Fried & McGuire 2024 *JPCA* 128, 8254



Candidate	Catalog - Exp.
A	0.201 MHz
B	0.121 MHz
C	0.020 MHz
D	0.356 MHz
E	0.549 MHz

How do we teach a computer to
have “chemical intuition”?

CONTEXT IS KEY

Efficient Estimation of Word Representations in Vector Space

word2vec

Tomas Mikolov

Google Inc., Mountain View, CA

tmikolov@google.com

Kai Chen

Google Inc., Mountain View, CA

kaichen@google.com

Greg Corrado

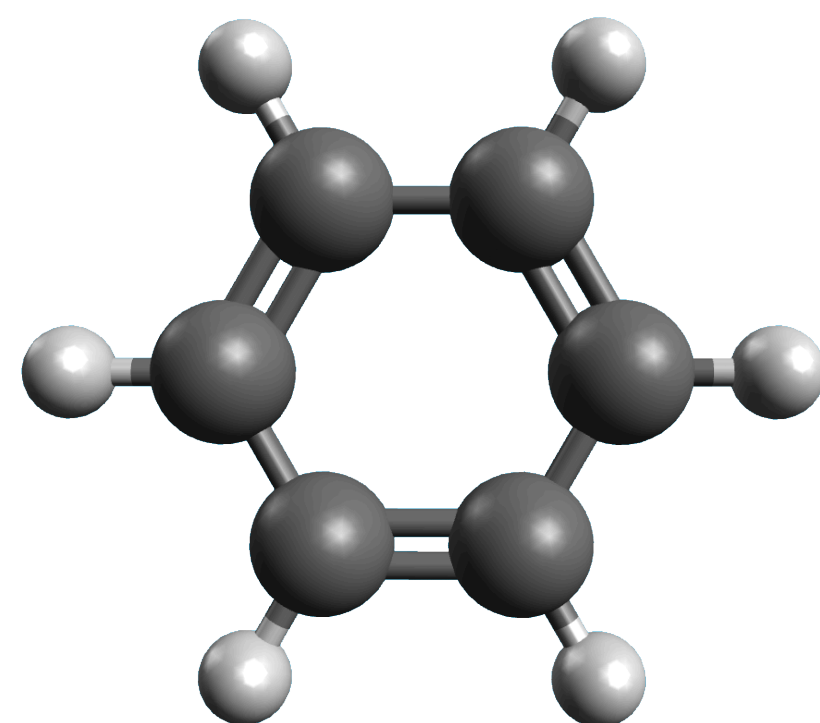
Google Inc., Mountain View, CA

gcorrado@google.com

Jeffrey Dean

Google Inc., Mountain View, CA

jeff@google.com



SMILES c1ccccc1

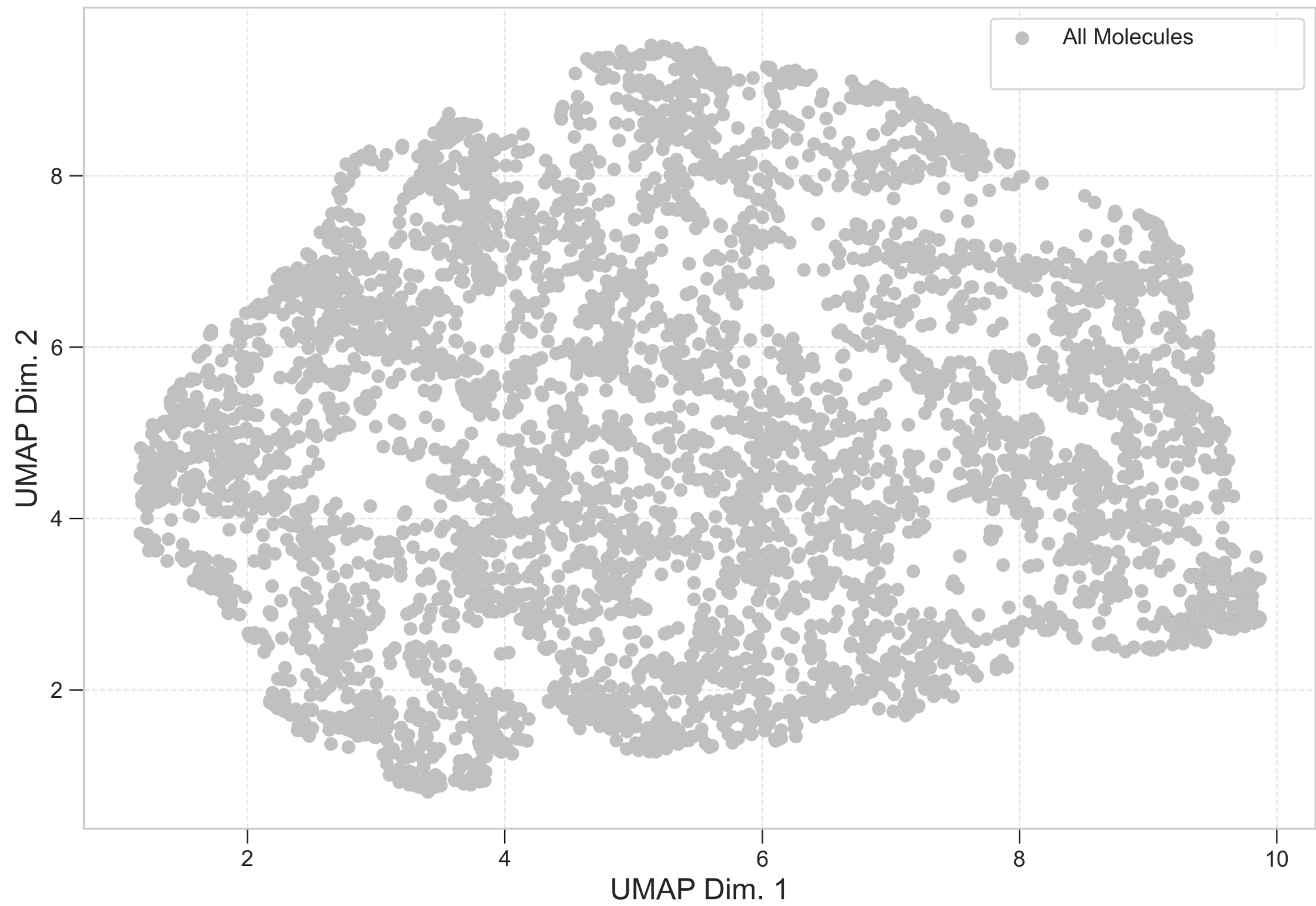
SELFIES

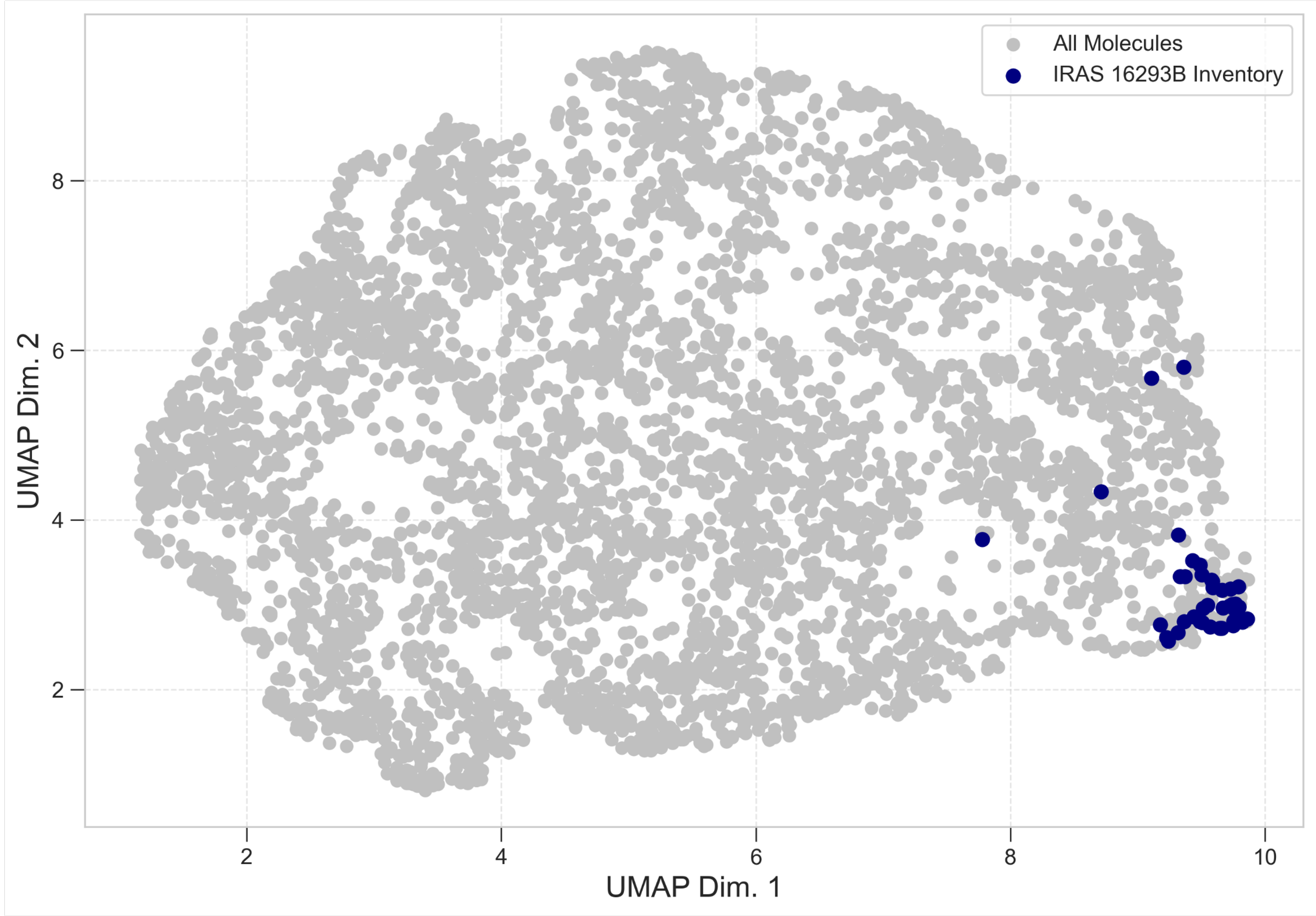
[C][=C][C][=C][C][=C][Ring1][=Branch1]

mol2vec

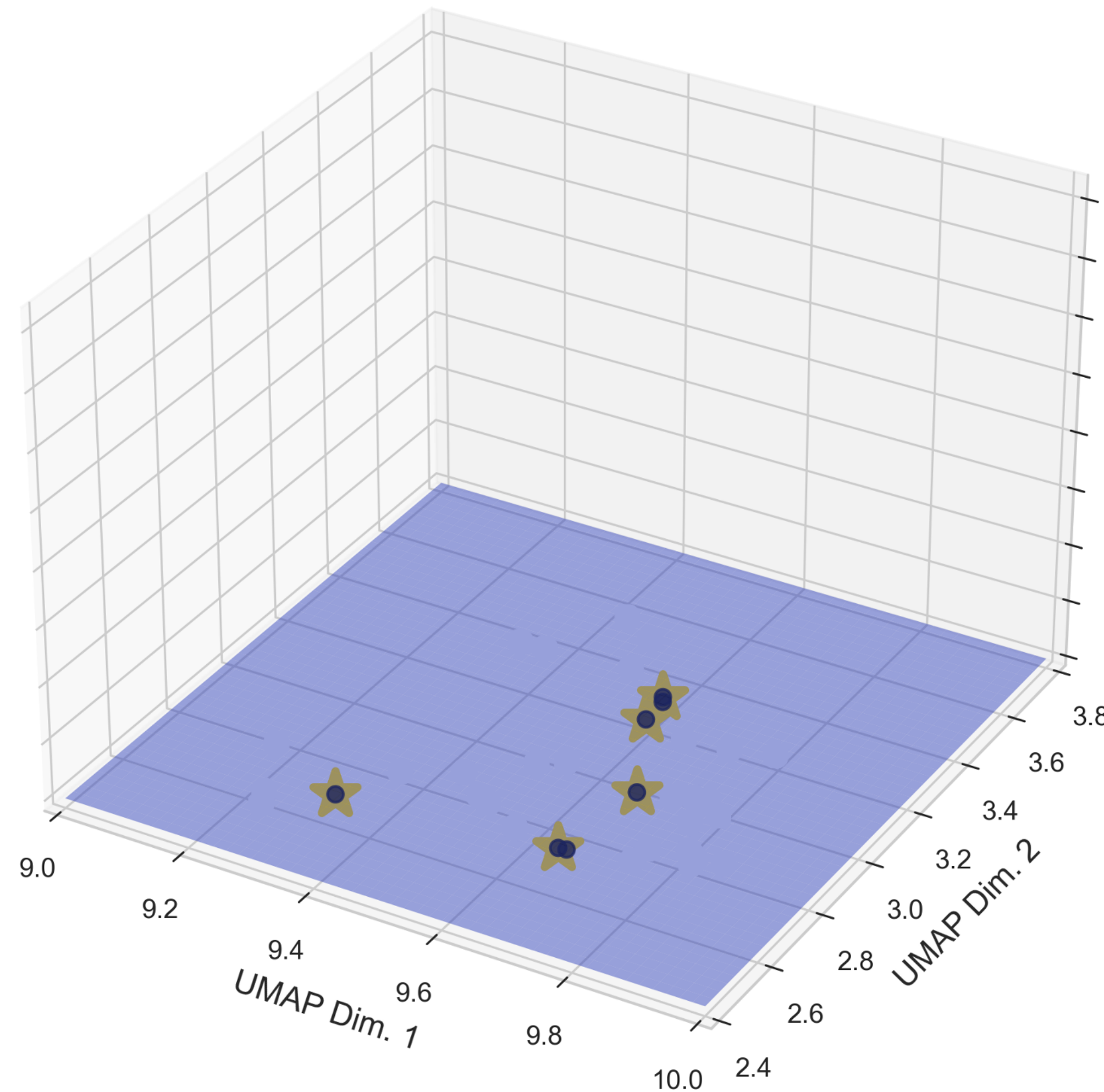
VICGAE

$$\begin{bmatrix} n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_x \end{bmatrix}$$





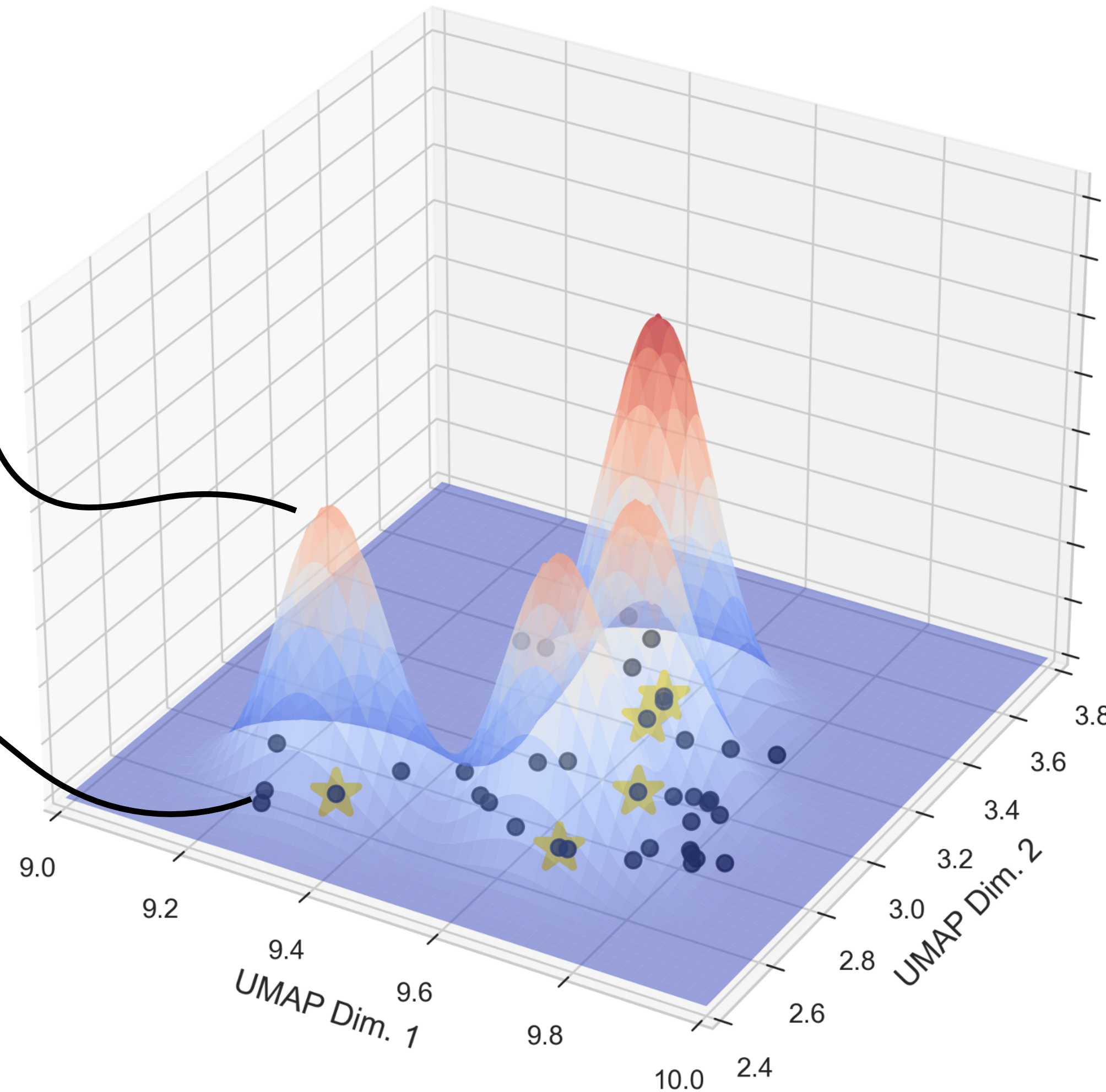
Molecules with
strongest transitions in
PILS ...



OCS
H₂CO
CH₃CHO
CH₃OCH₃
OCS
H₂CCO

Gaussian weight
added around each
strong molecule

Other known
molecules already
fall in non-zero
weighted regions

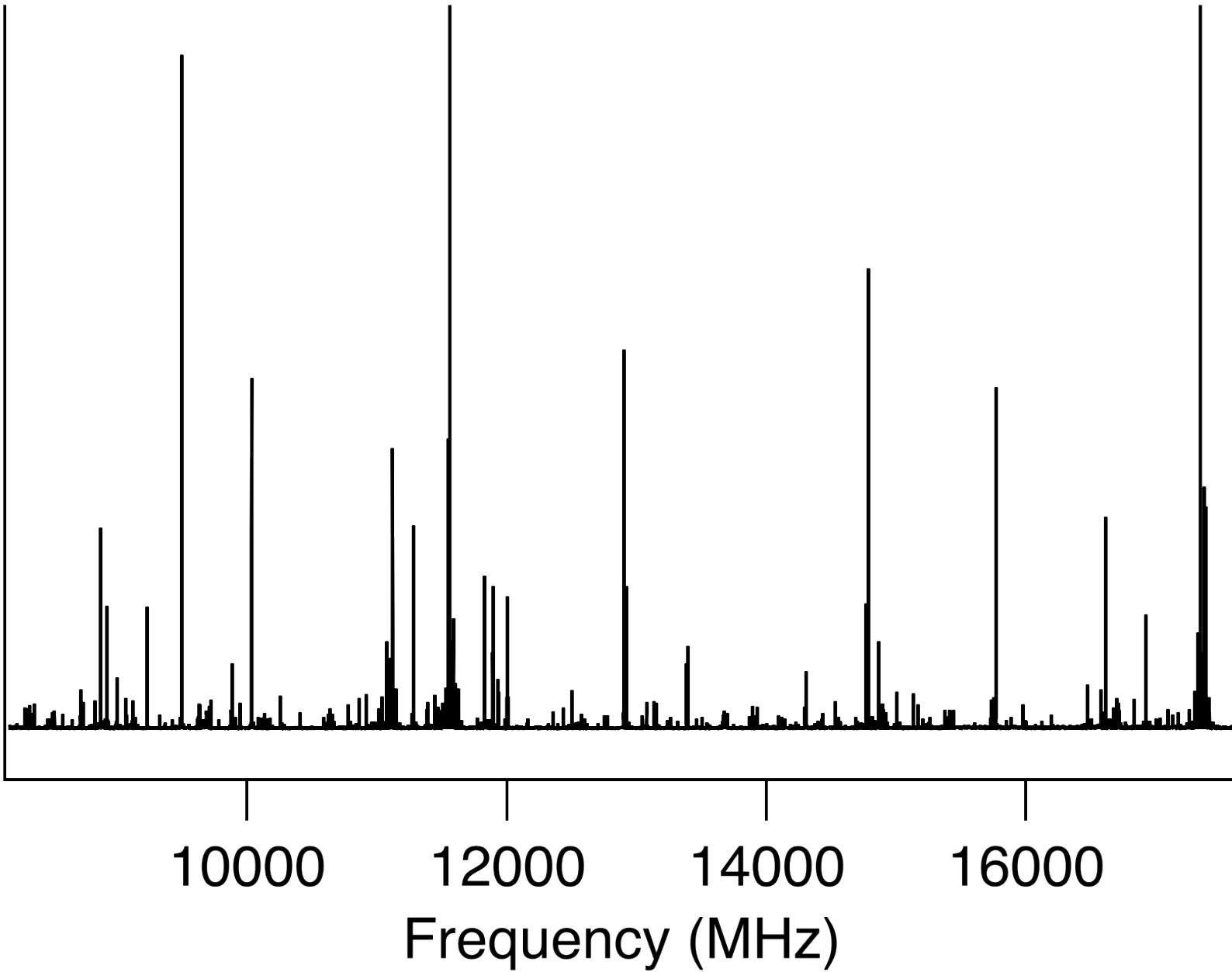


Molecules with
strongest transitions in
PILS ...

OCS
H₂CO
CH₃CHO
CH₃OCH₃
OCS
H₂CCO

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

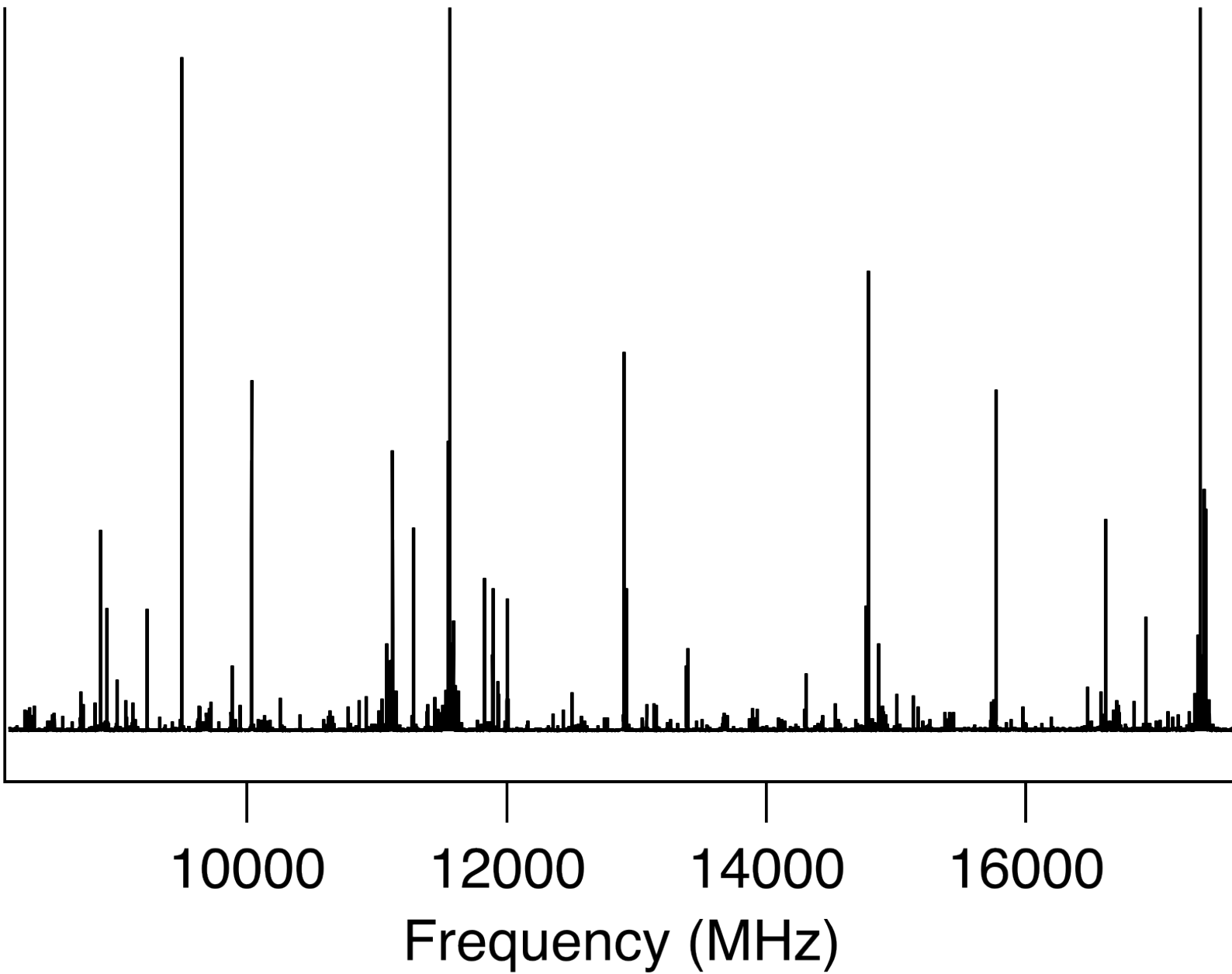
Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision



Line Survey

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision



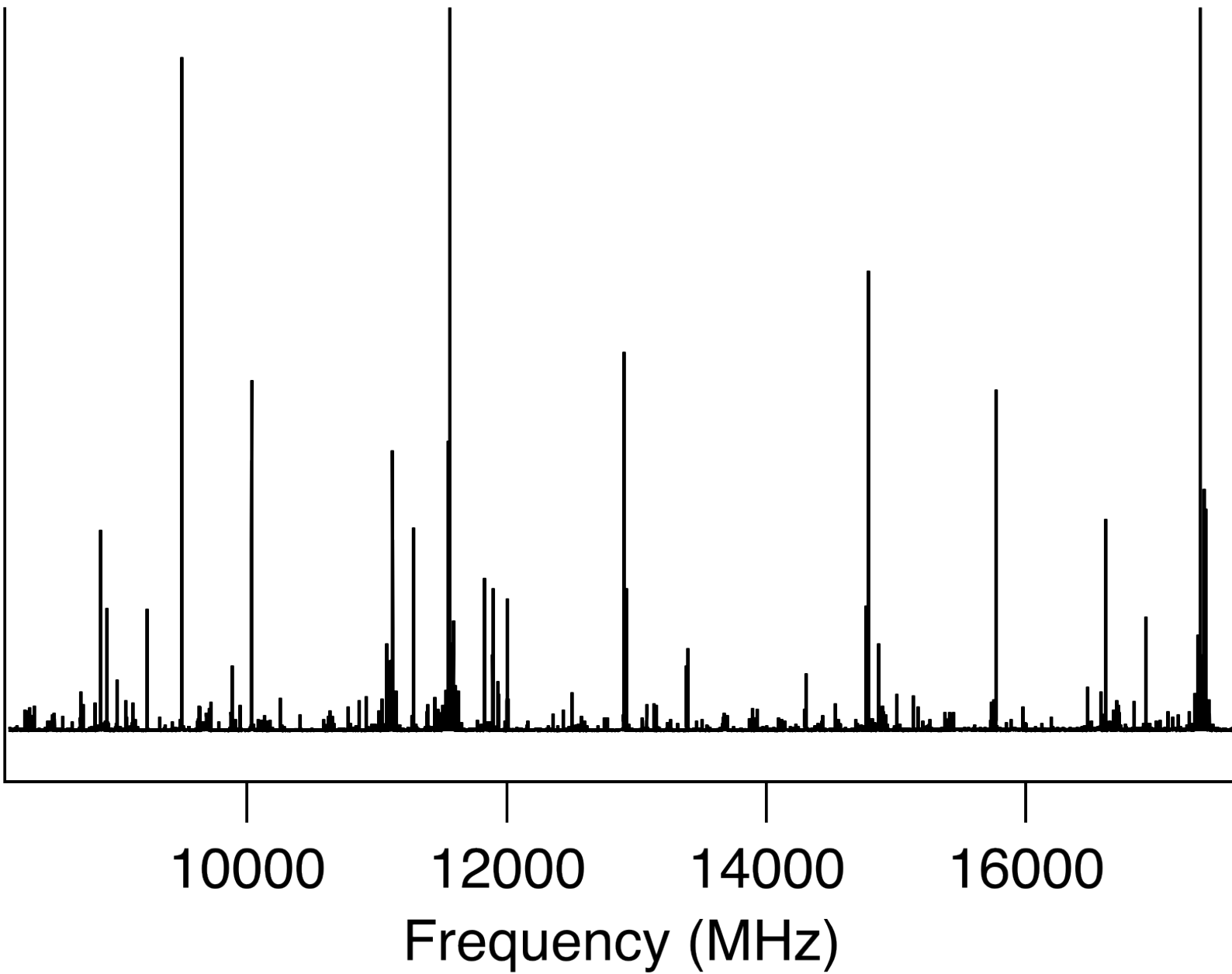
Line Survey

Peak
Finder →

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

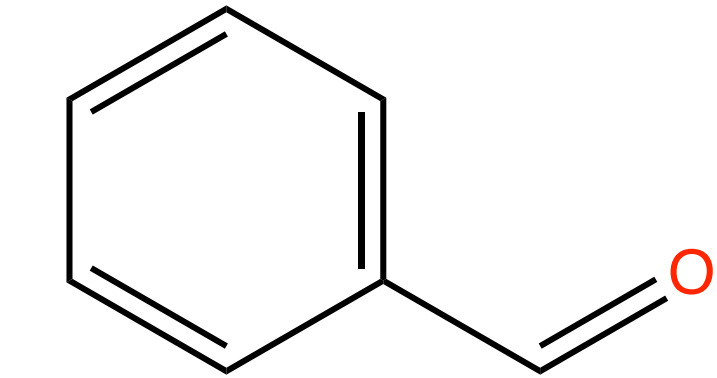


Line Survey

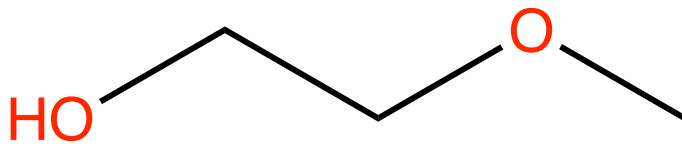
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

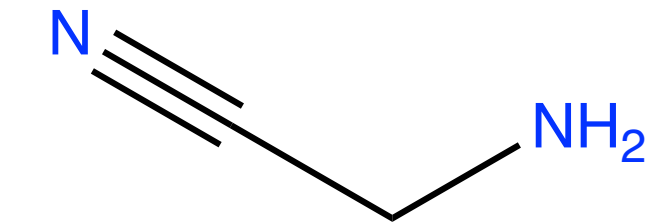
Database Query



8314.92 MHz



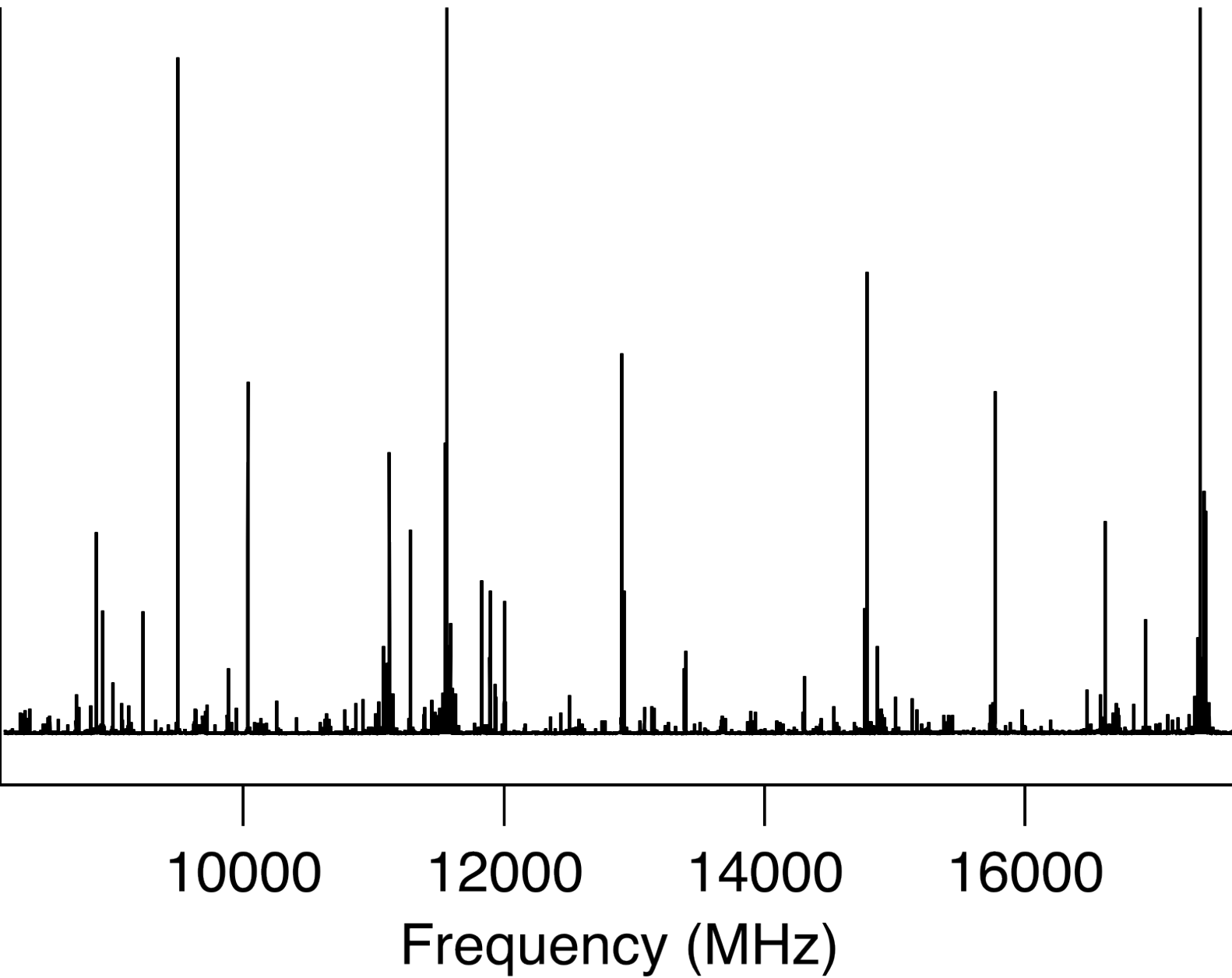
8135.07 MHz



8135.14 MHz

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

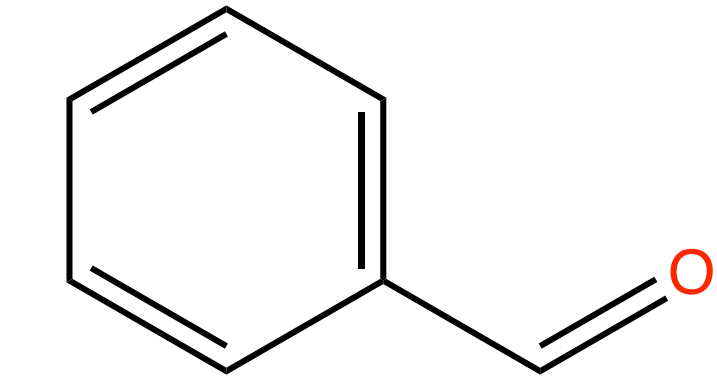


Line Survey

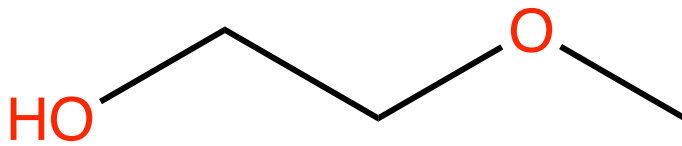
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

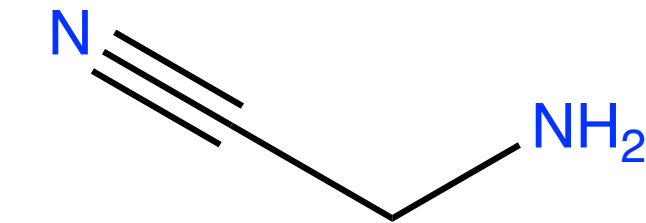
Database Query



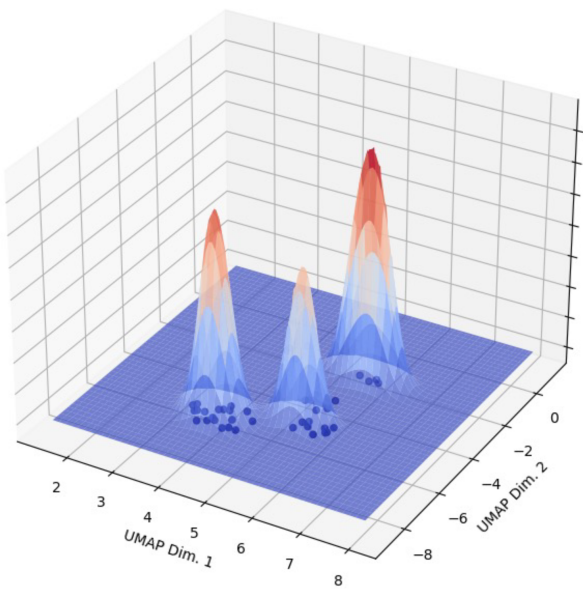
8314.92 MHz



8135.07 MHz



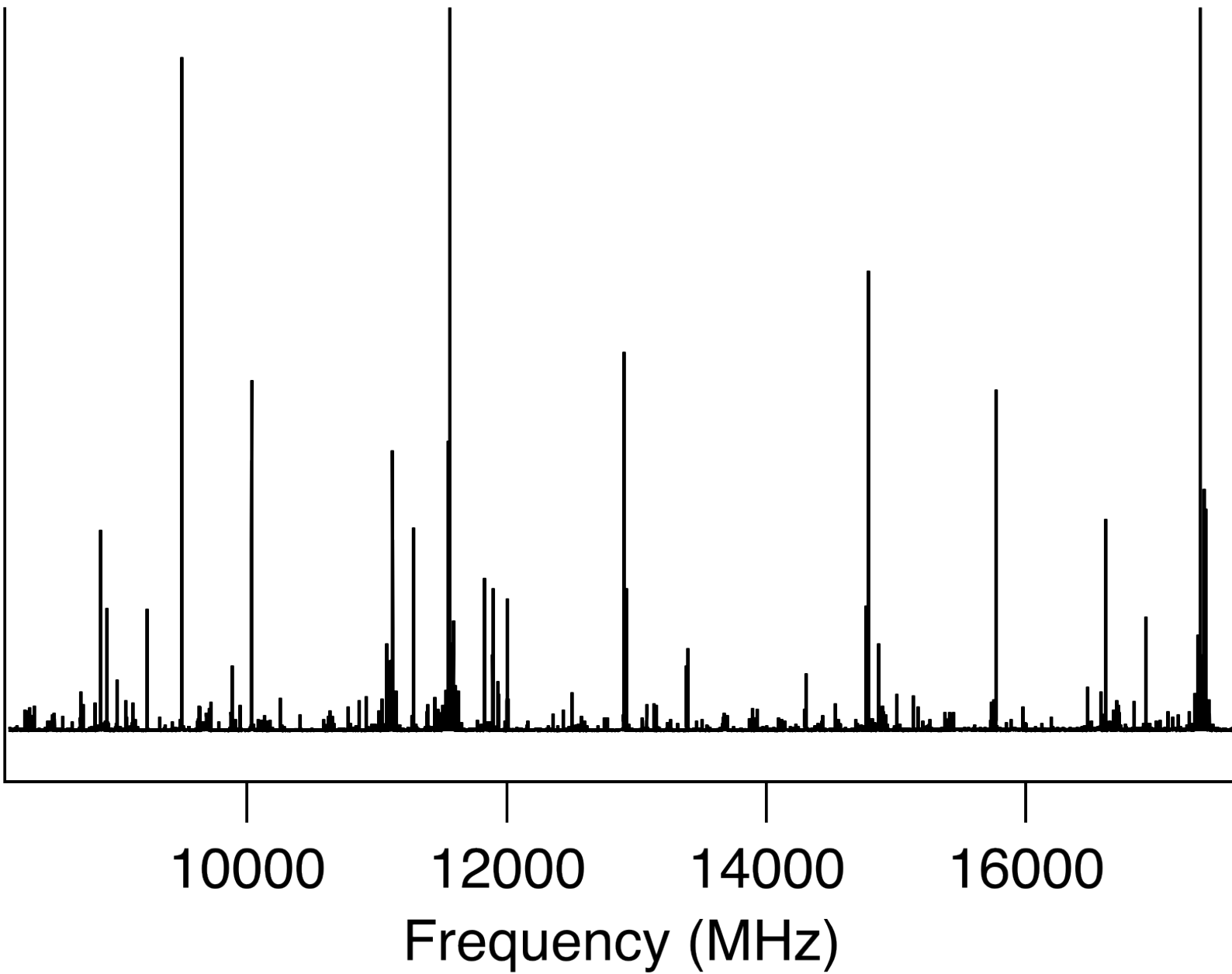
8135.14 MHz



Chemical Intuition Calculation

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

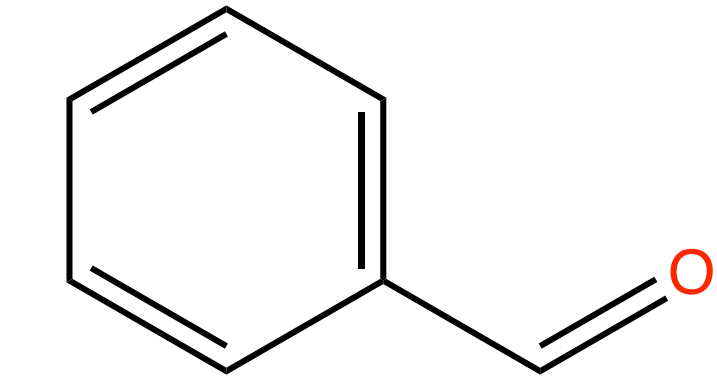


Line Survey

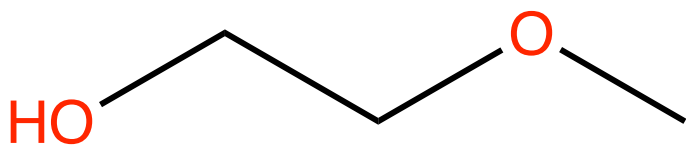
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

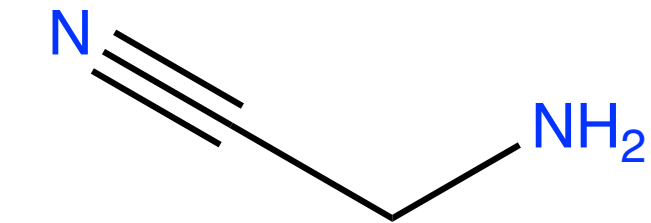
Database Query



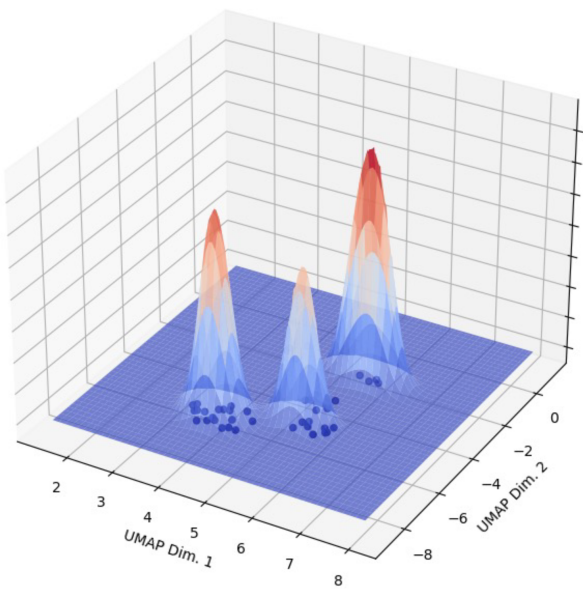
8314.92 MHz



8135.07 MHz



8135.14 MHz

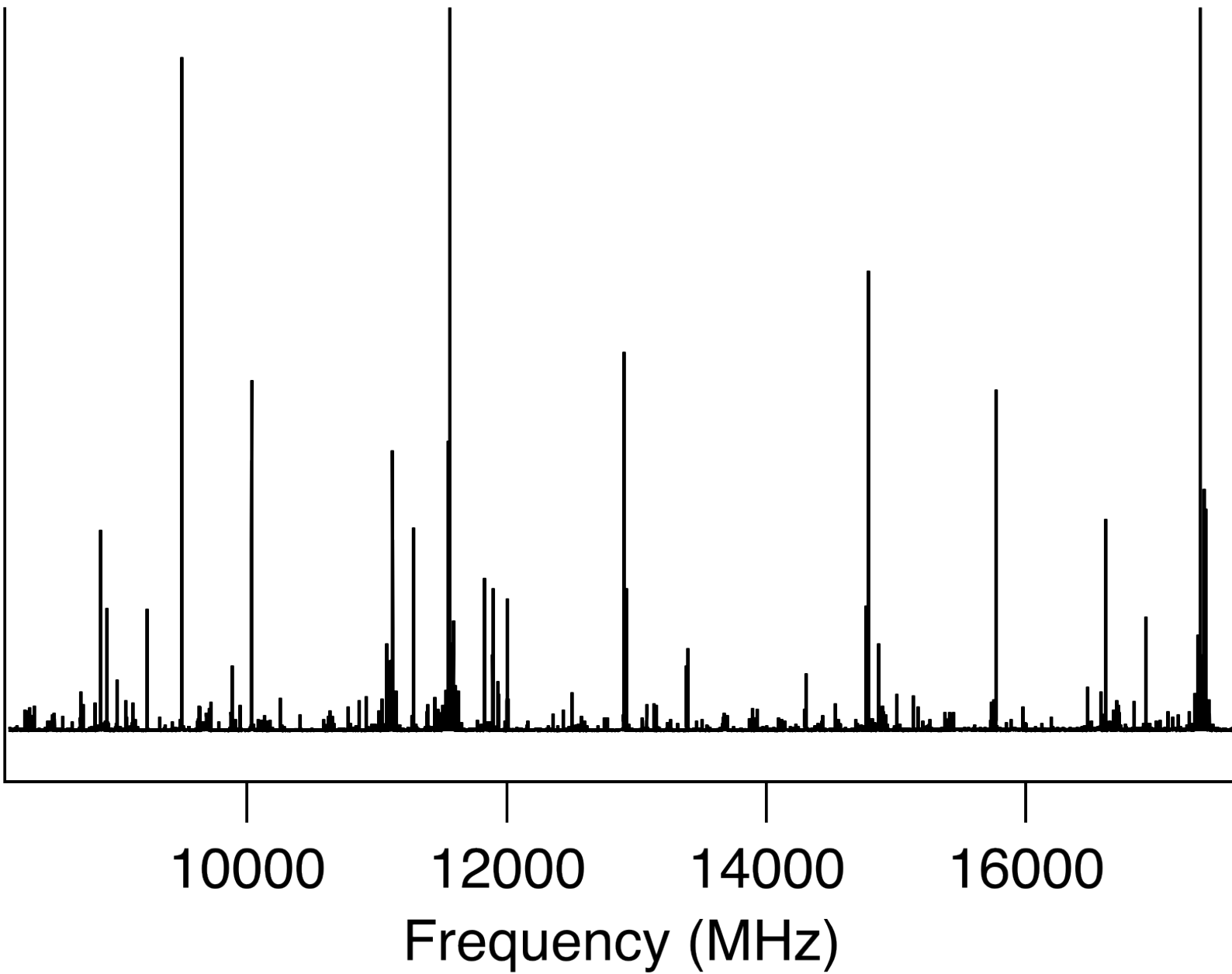


Chemical Intuition Calculation

Scale Based on Frequency Match

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

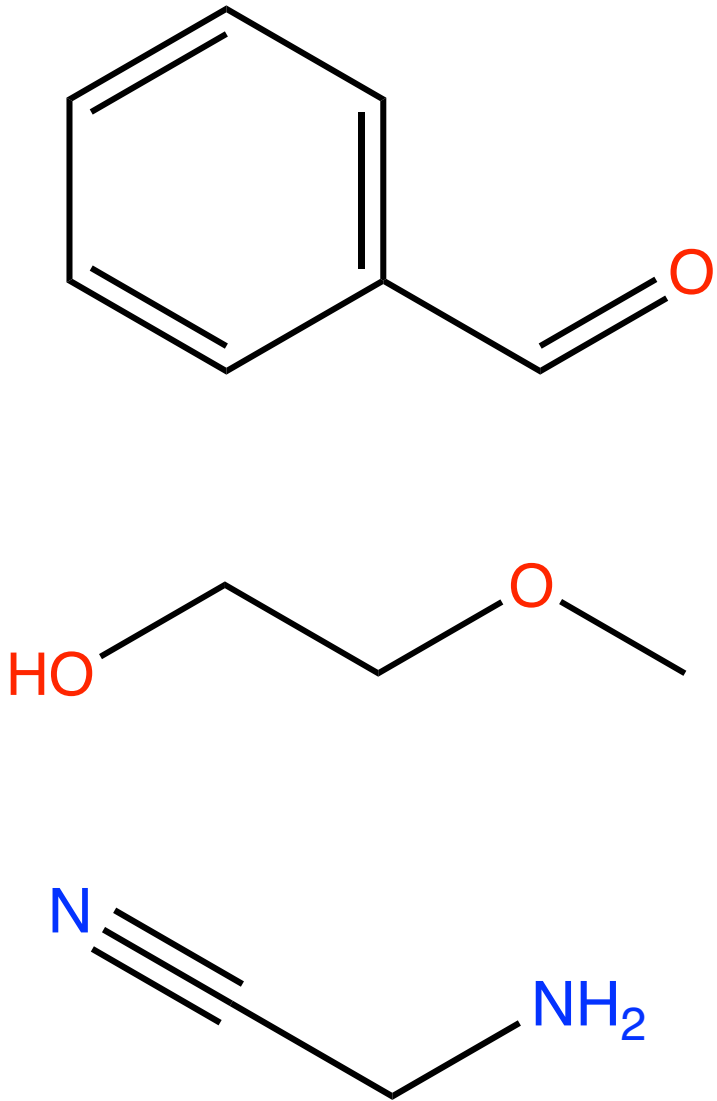


Line Survey

Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

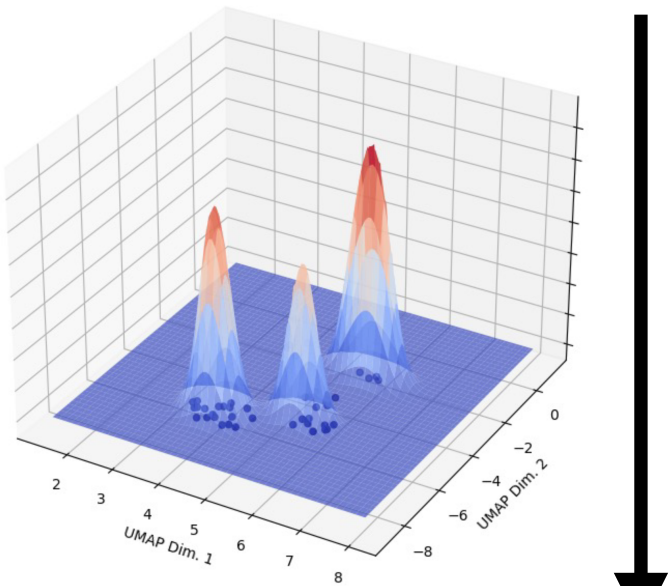
Database Query



8314.92 MHz

8135.07 MHz

8135.14 MHz



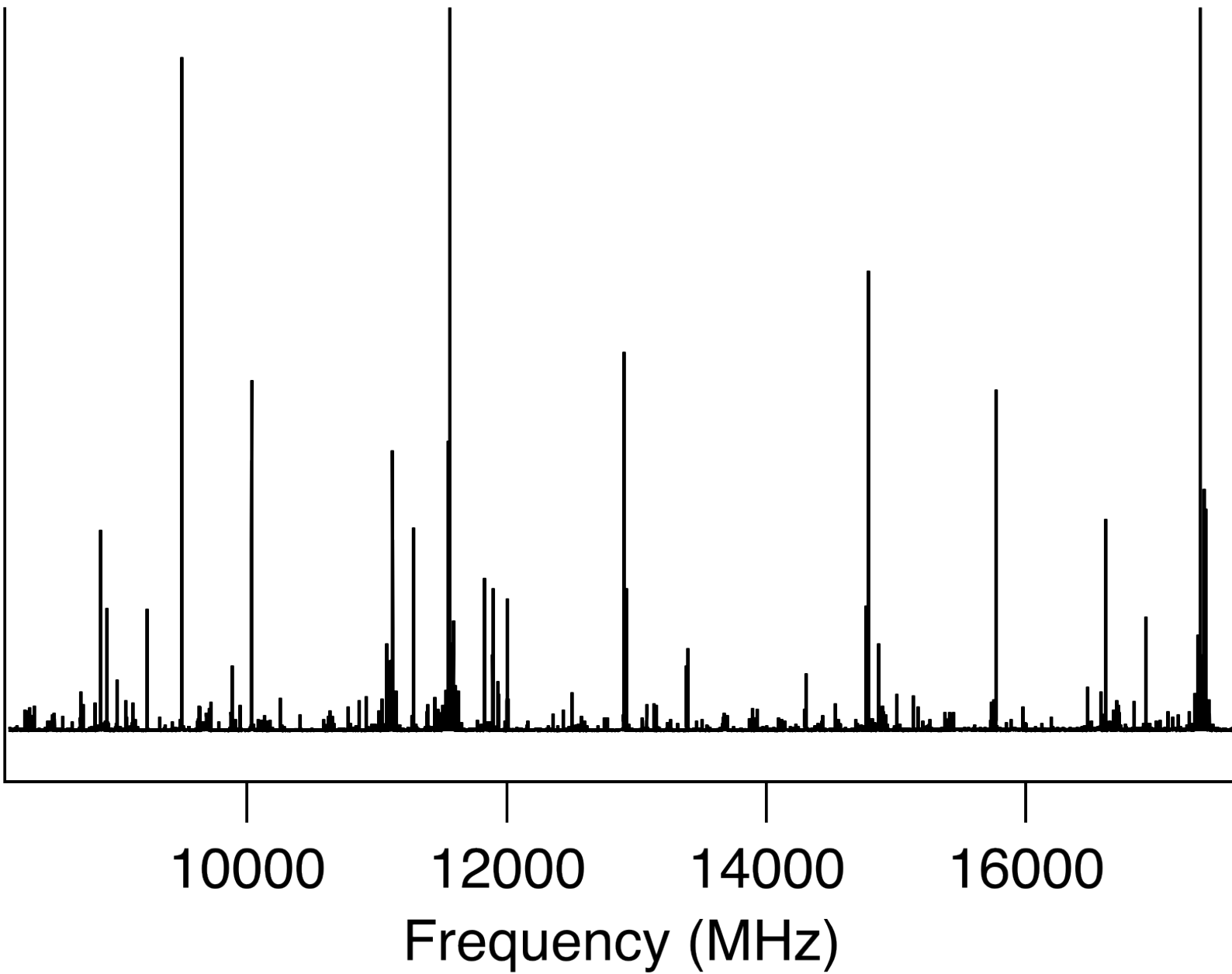
Chemical Intuition Calculation

Scale Based on Intensity Match

Scale Based on Frequency Match

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

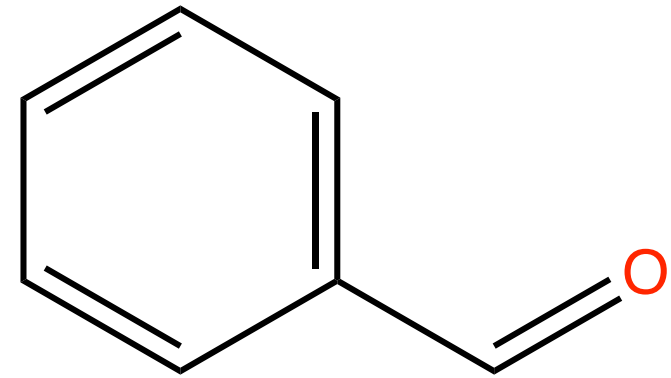


Line Survey

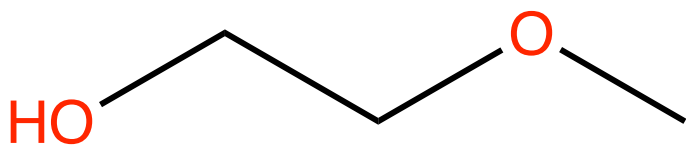
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

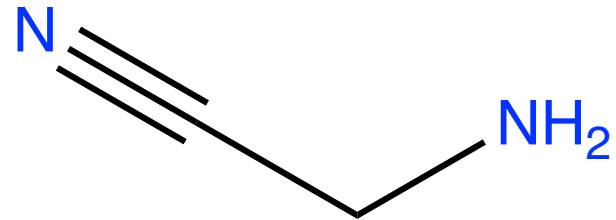
Database Query



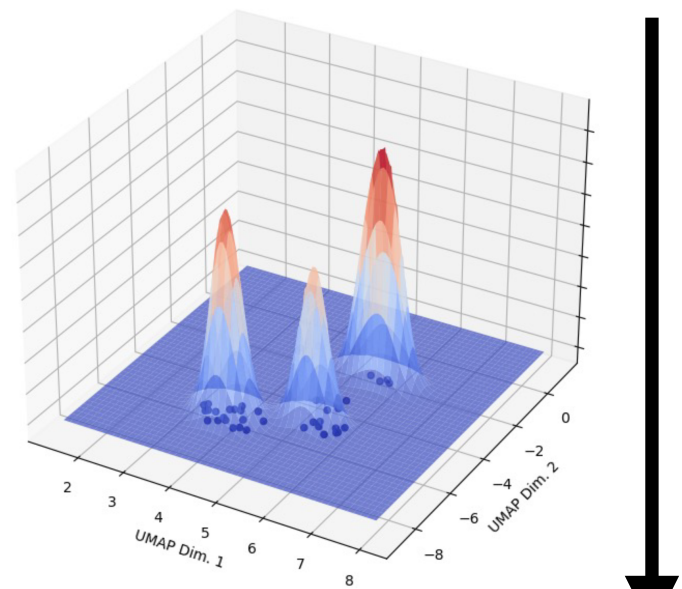
8314.92 MHz



8135.07 MHz

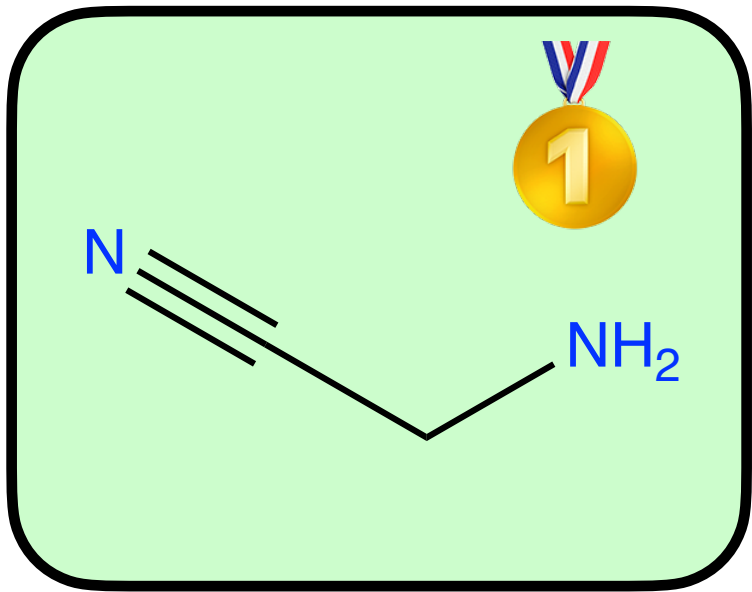


8135.14 MHz



Chemical Intuition Calculation

Assign Molecule with Top Score

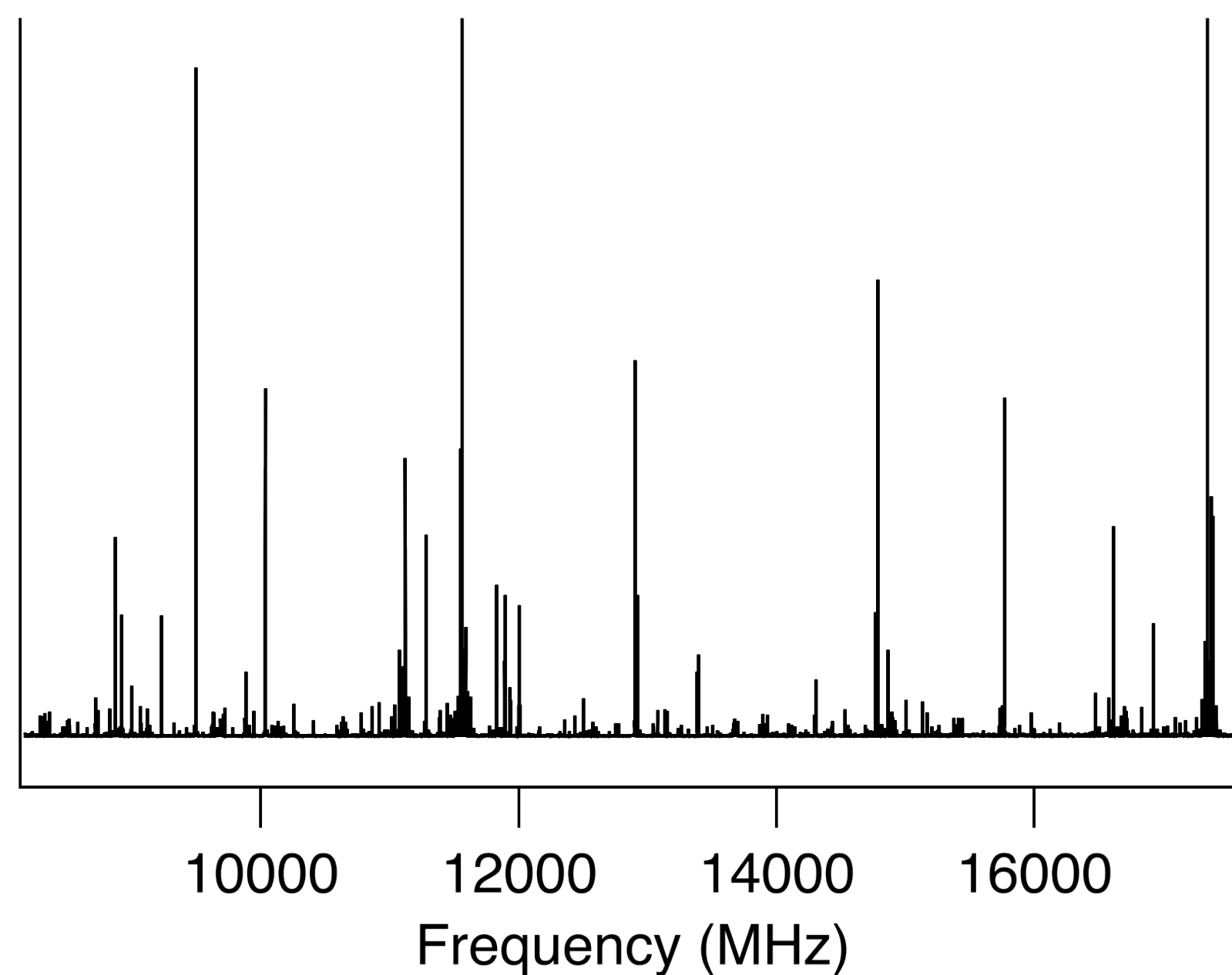


Scale Based on Intensity Match

Scale Based on Frequency Match

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

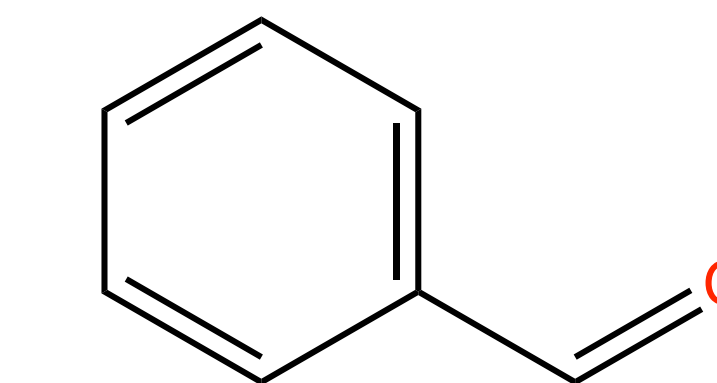


Line Survey

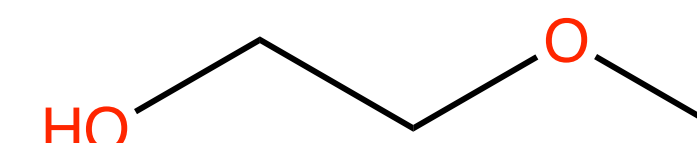
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

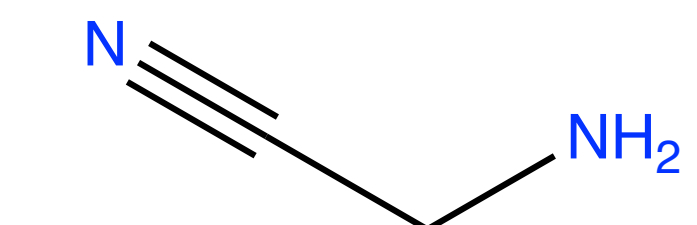
Database Query



8314.92 MHz

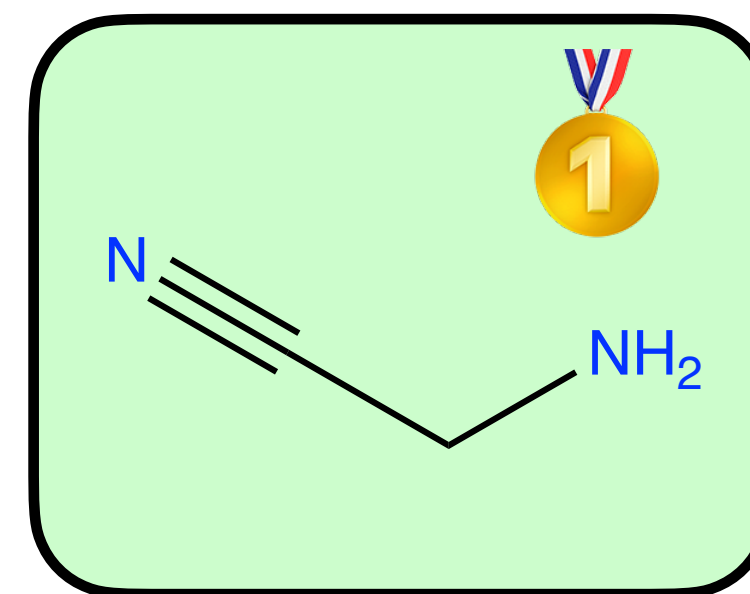


8135.07 MHz

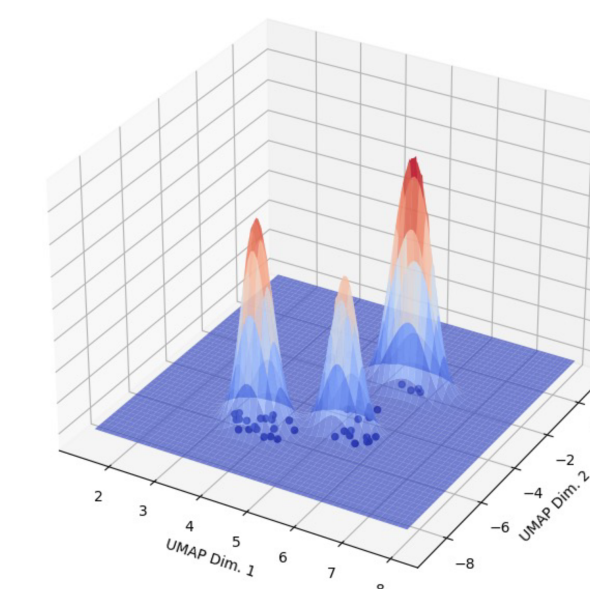


8135.14 MHz

Assign Molecule
with Top Score



Update Surface



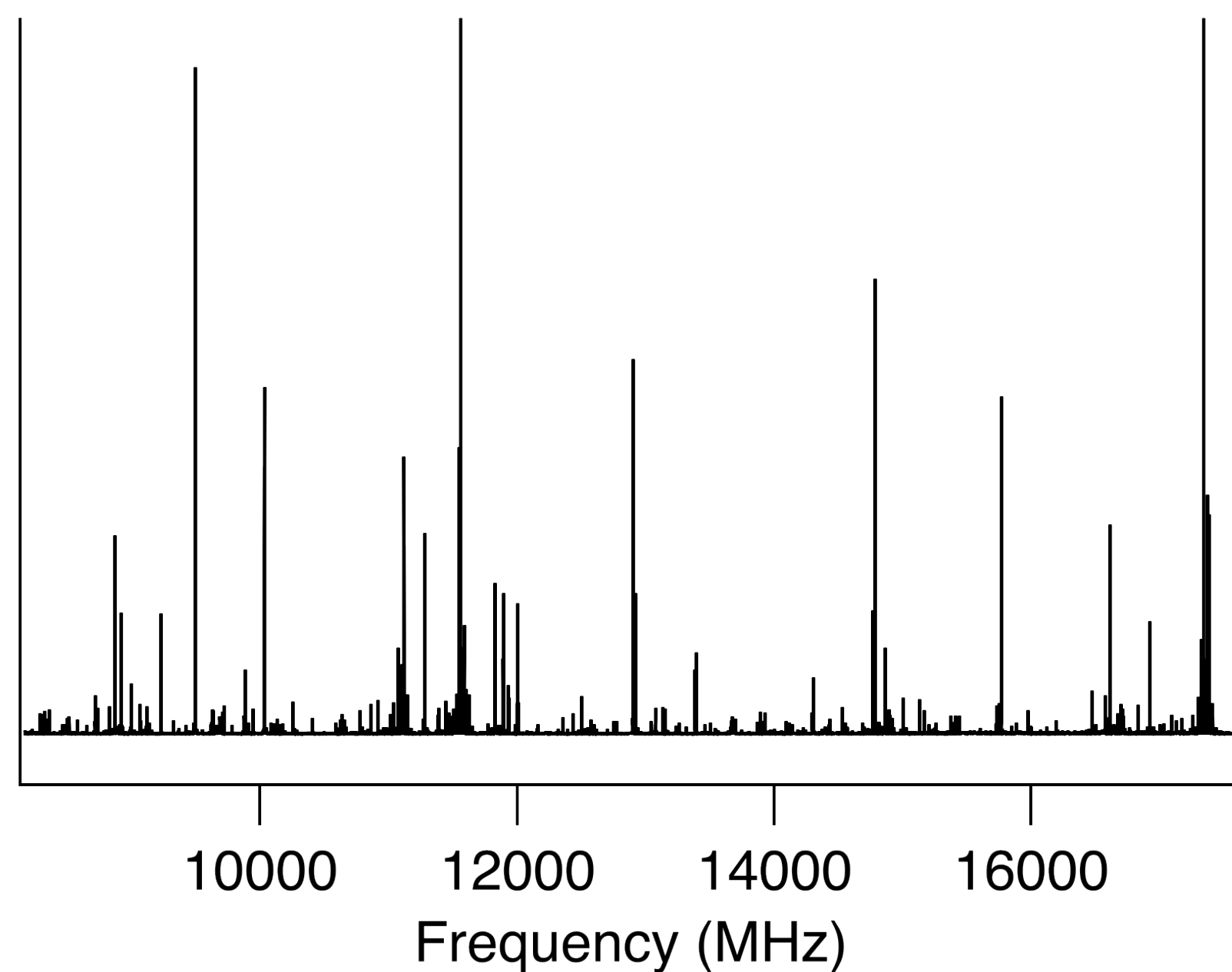
Chemical
Intuition
Calculation

Scale Based on
Intensity Match

Scale Based on
Frequency
Match

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

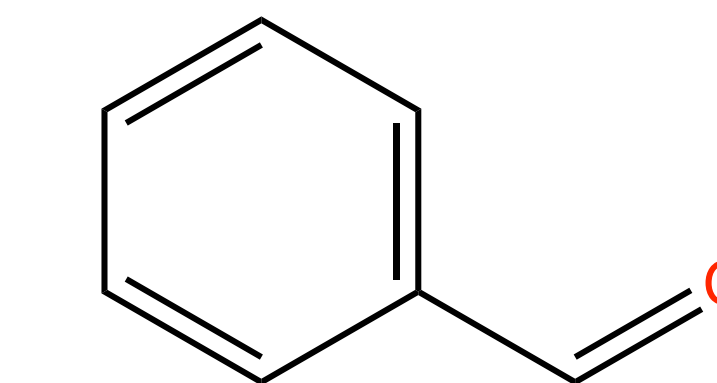


Line Survey

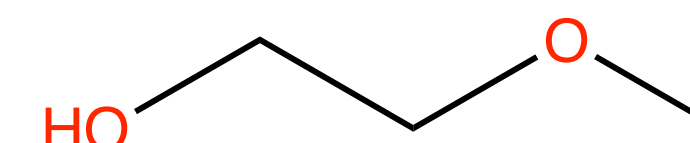
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

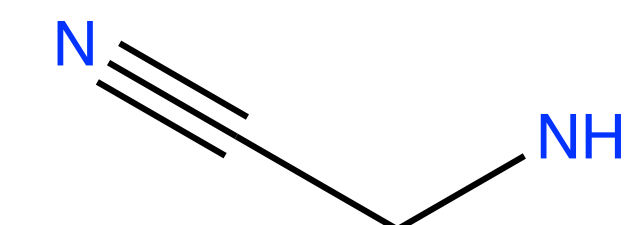
Database Query



8314.92 MHz

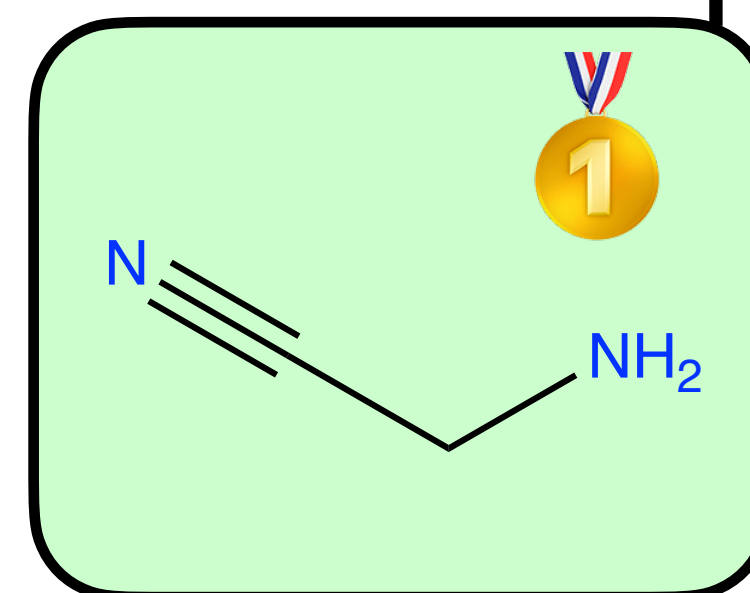


8135.07 MHz

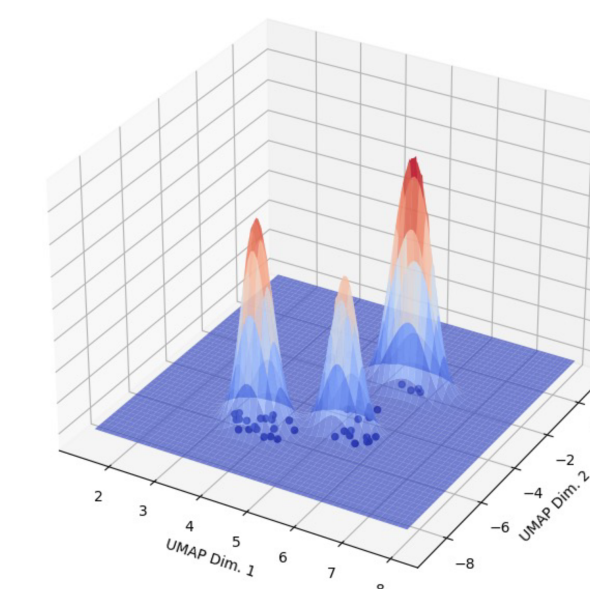


8135.14 MHz

Assign Molecule
with Top Score



Update Surface



Chemical
Intuition
Calculation

Scale Based on
Intensity Match

Scale Based on
Frequency
Match

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2026 *ApJ*, in revision

Testing on laboratory data ...

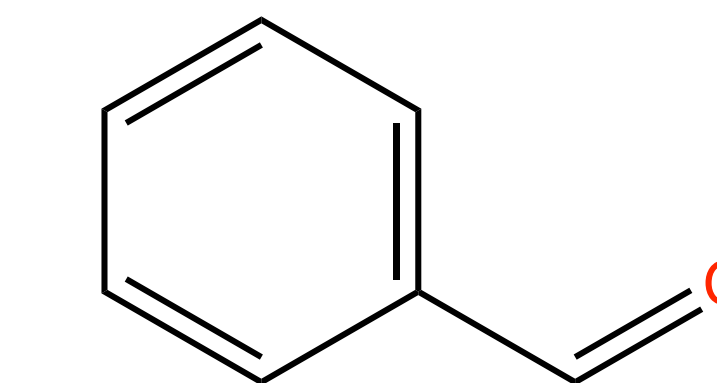
~98% Accuracy vs Human

~15 minutes vs many days

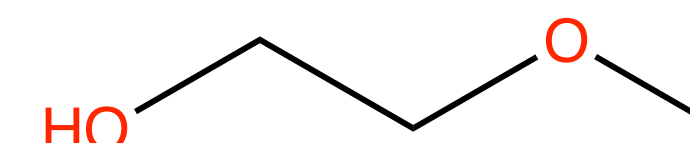
Peak Finder

Freq.	Int.
8131	0.25
8315	0.93
9273	0.18
...	...

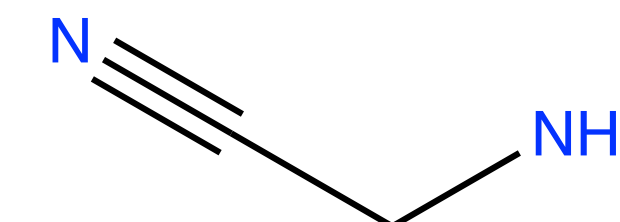
Database Query



8314.92 MHz



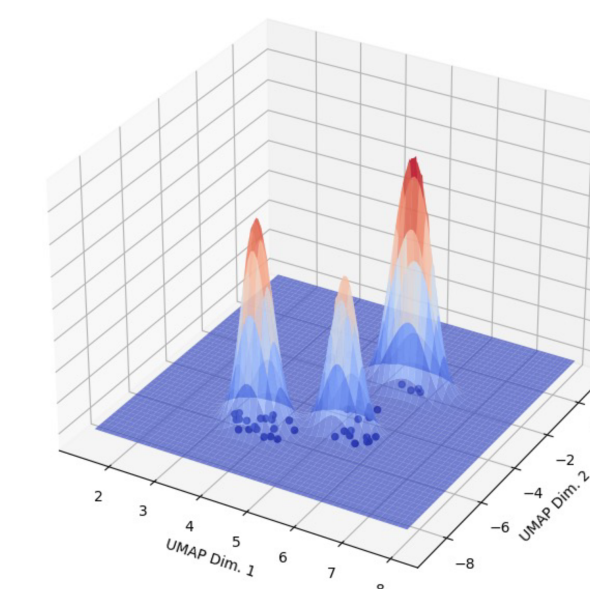
8135.07 MHz



8135.14 MHz

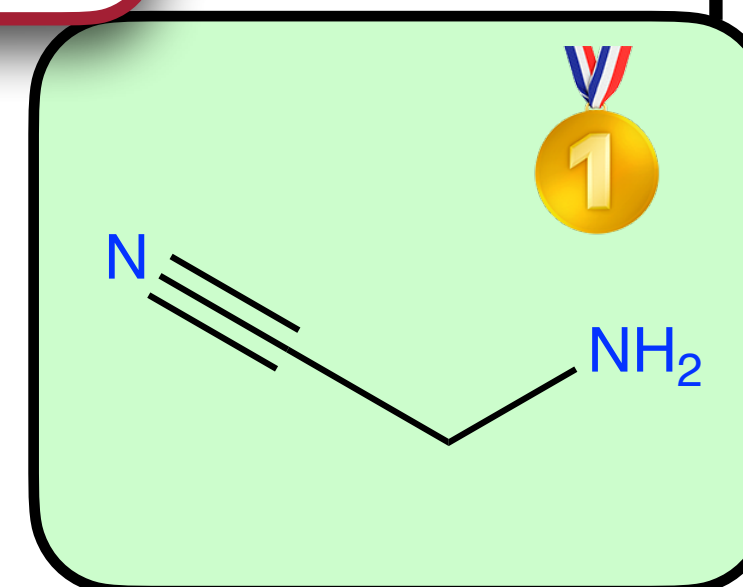
Next Line

Update Surface



Chemical Intuition Calculation

Assign Molecule with Top Score



Scale Based on Intensity Match

Scale Based on Frequency Match

... automate ΔV determination

... automate v_{lsr} determination

... try a few T_{ex} and N_{T} for common species

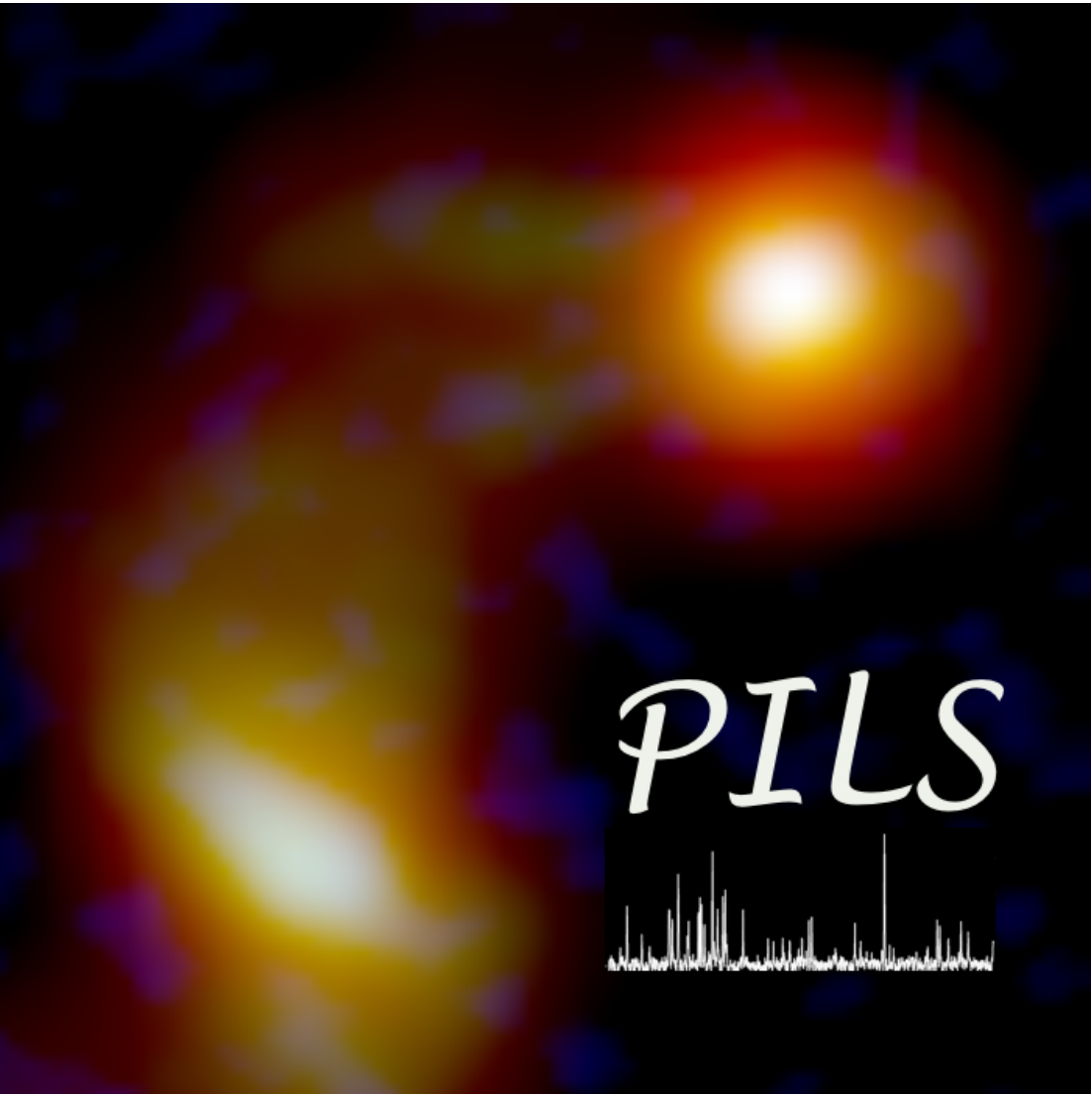


TMC-1

Green Bank Telescope

~4 - 36 GHz

568 Lines $>5\sigma$



IRAS 16293

ALMA

~329 - 363 GHz

2138 Lines $>5\sigma$



TMC-1

Green Bank Telescope

~4 - 36 GHz

568 Lines $>5\sigma$

Assigned 525/568 Lines
($>99\%$ of Total Intensity)
To 67 Molecules

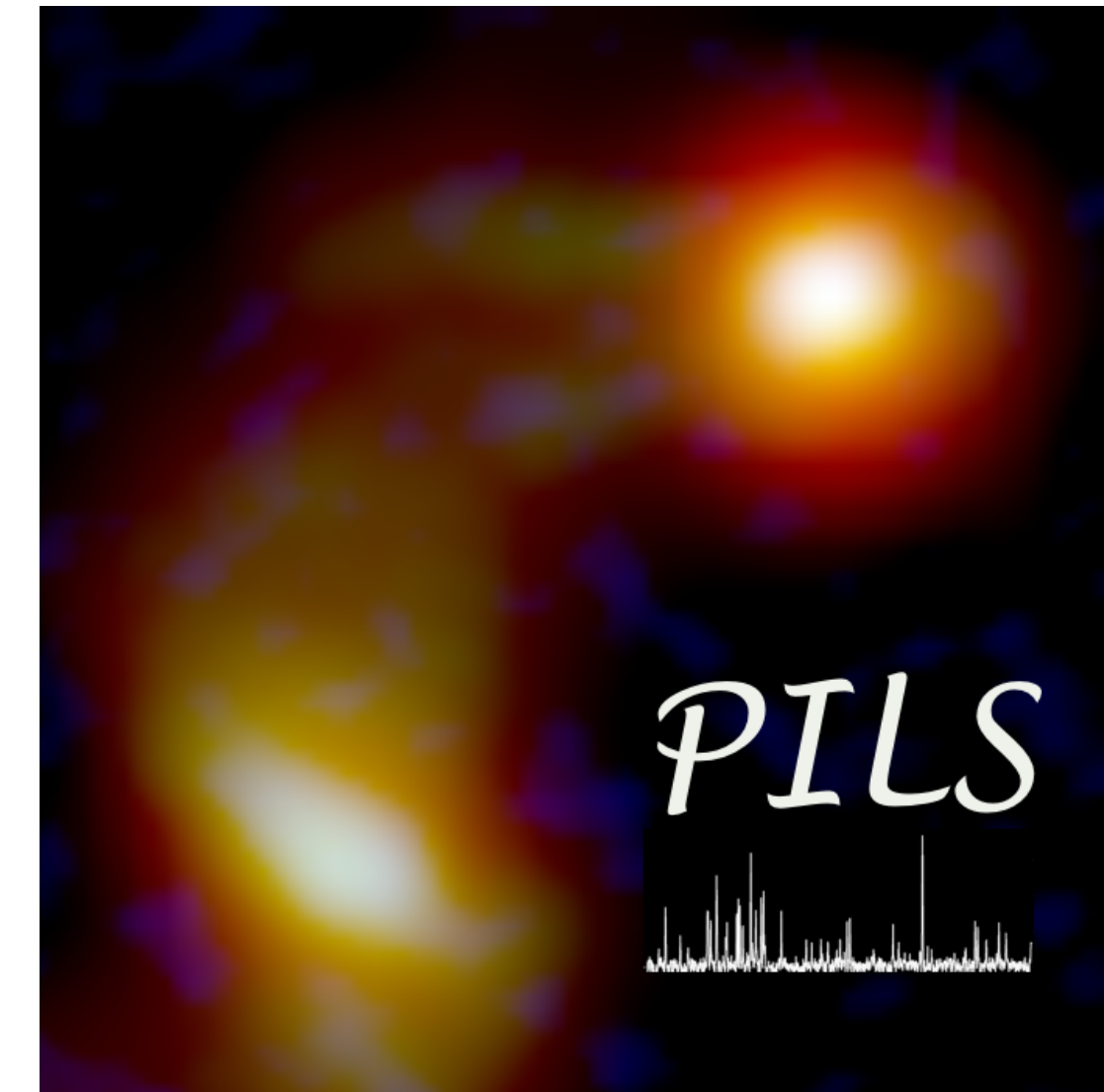
Zero False Positives

Total Time: 17 Minutes

Assigned 1866/2138 Lines
(>90% of Total Intensity)
to 80 Molecules

One False Positive

Total Time: 6 Minutes

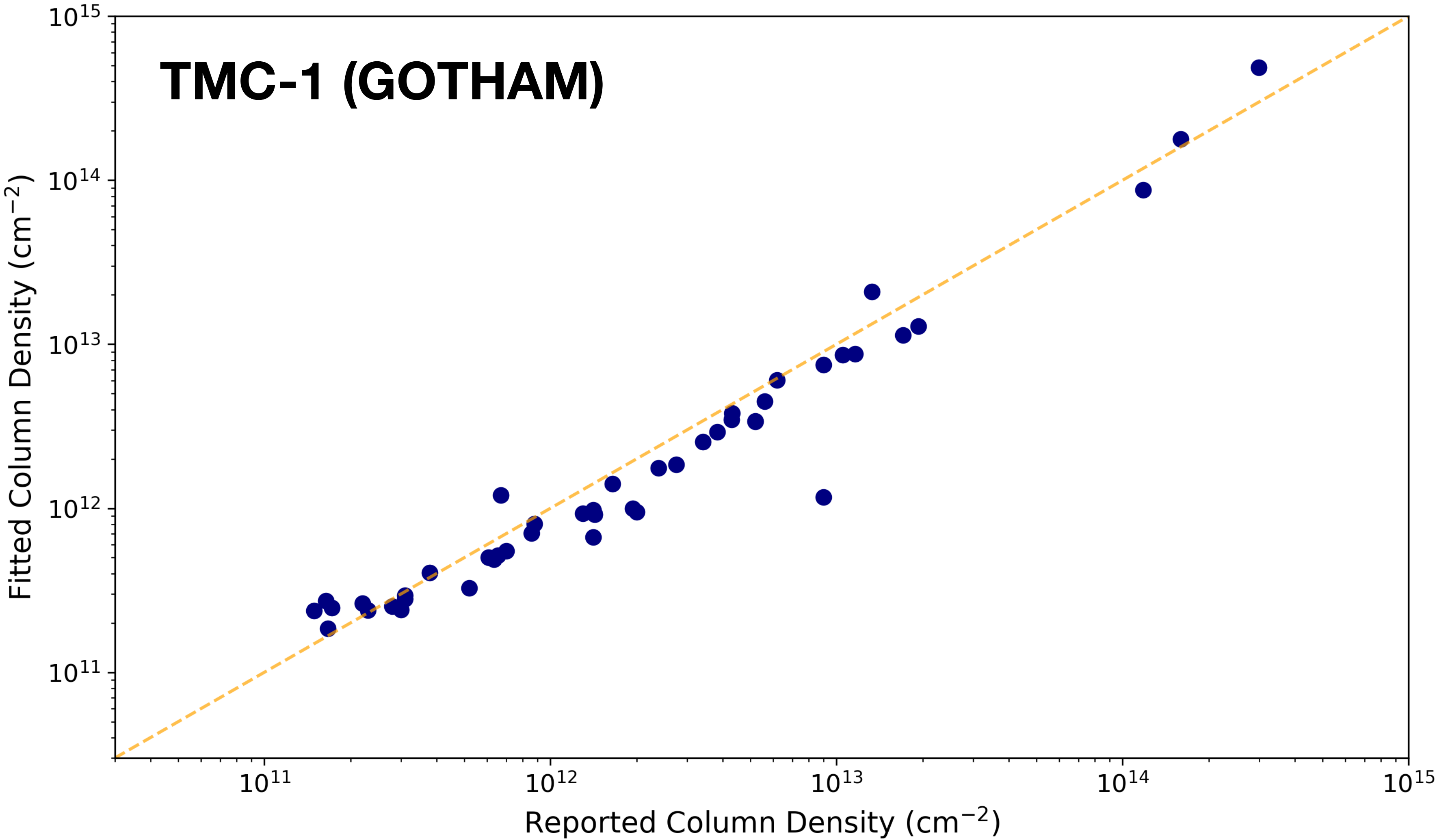


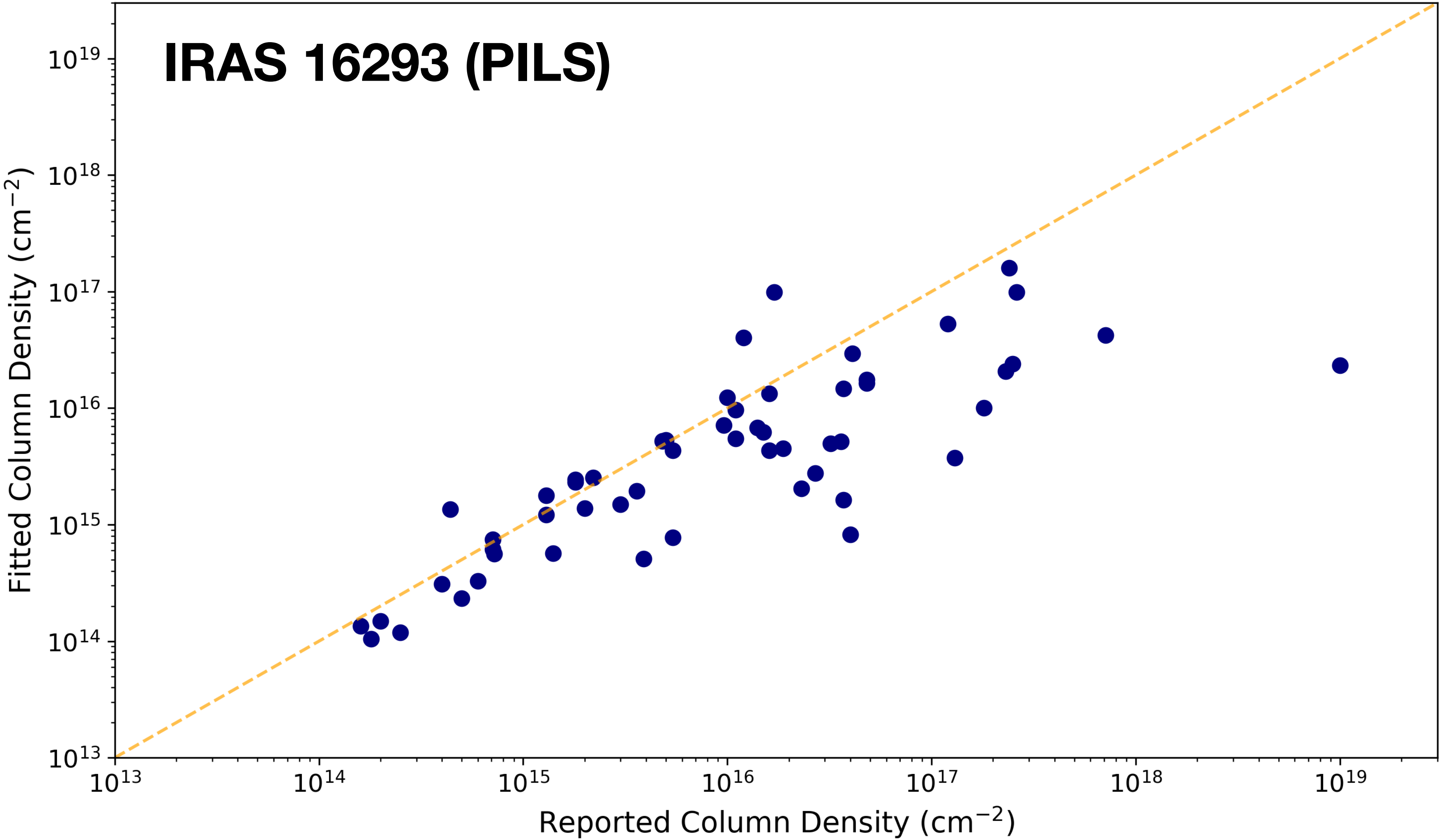
IRAS 16293

ALMA

~329 - 363 GHz

2138 Lines $>5\sigma$





Speed

Fast

Speed

Fast

Line ID Completeness

V. High

Speed

Fast

Line ID Completeness

V. High

Accuracy

Accuracy

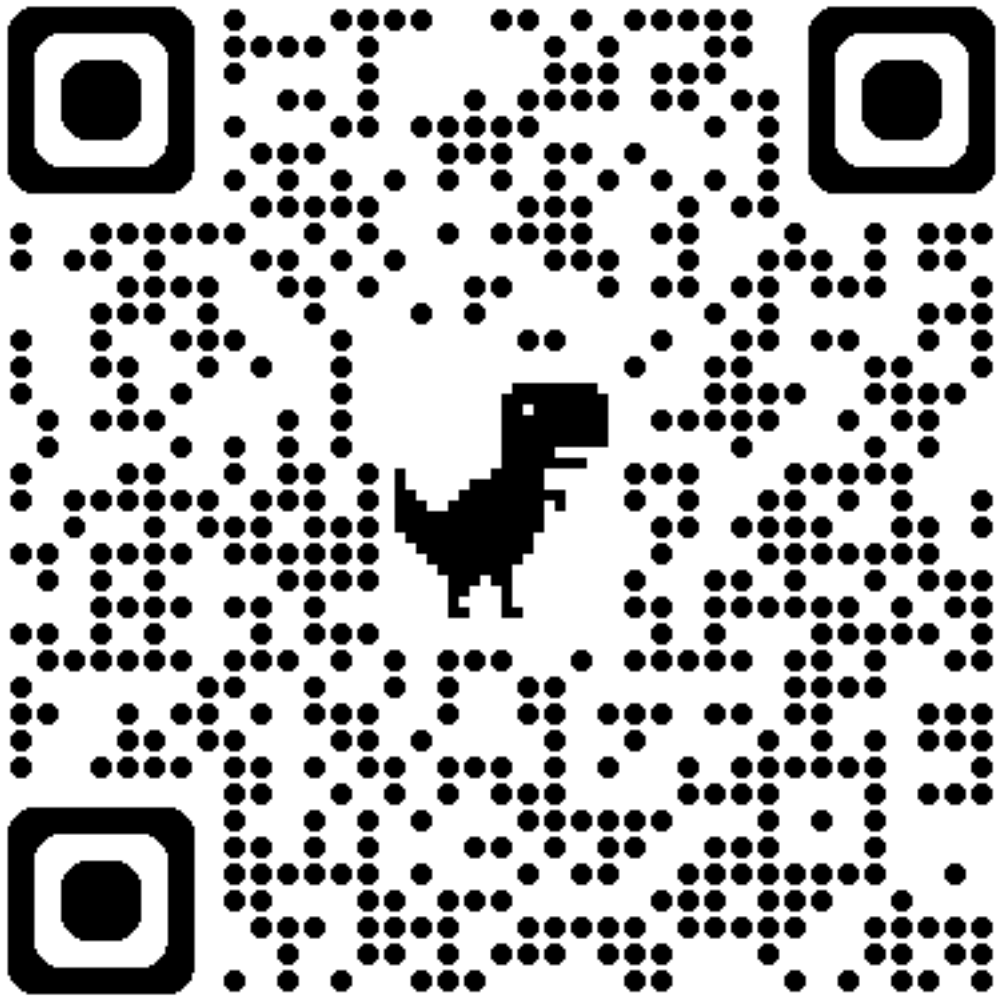
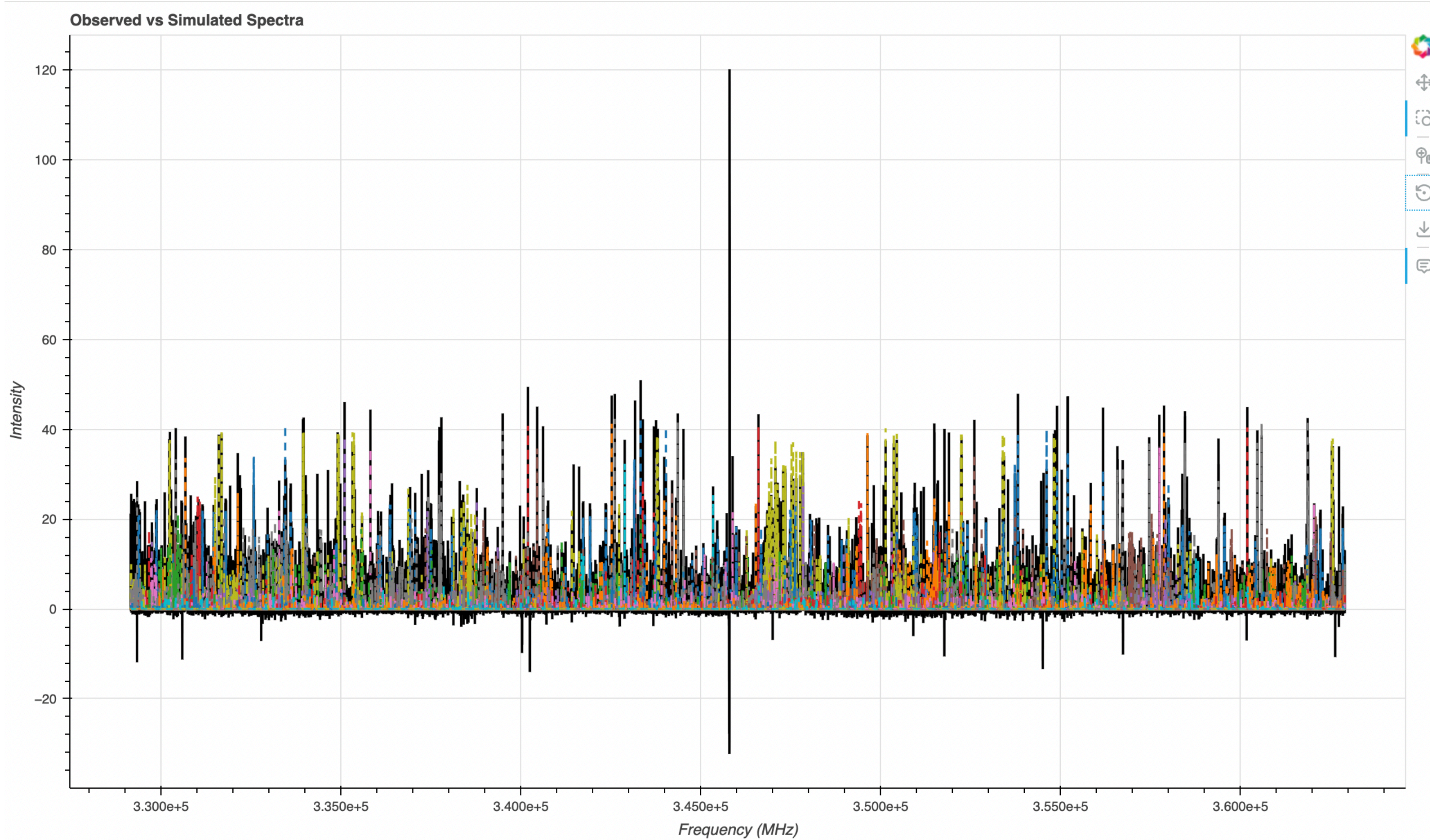
Speed	Fast
Line ID Completeness	V. High
Line ID	V. High
ΔV	V. High
V_{lsr}	V. High

Accuracy

Speed	Fast
Line ID Completeness	V. High
Line ID	V. High
ΔV	V. High
V_{lsr}	V. High
T_{ex}	Pretty Good

Accuracy

Speed	Fast
Line ID Completeness	V. High
Line ID	V. High
ΔV	V. High
V_{lsr}	V. High
T_{ex}	Pretty Good
N_T	Not Bad



https://github.com/zfried/astro_amase