Machine-Learned Chemical Intuition













Brett A. McGuire

Disclaimers

These slides were not designed to be fully interpretable on their own - they were merely support for the oral presentation.

Many animations will not have properly transferred to PDF, and there may be some slides where animations have been rendered in such a way that some other information is obscured.

The specific examples and use cases here are drawn exclusively from work in my group and from my colleagues that I collaborate with, simply for my own ease of preparation. Many other groups and teams have thought and worked in this space, and have arrived at similar, complementary, and better approaches than we have.

The individuals on the title slide, and especially K.L.K. Lee, are the ones that actually have done all this work. I am merely summarizing and condensing (perhaps poorly) what they've done.

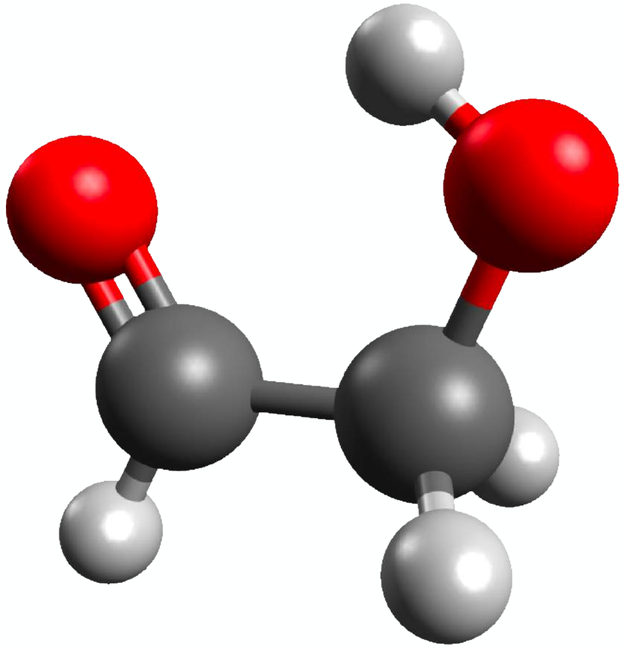
I am not an ML expert - I have done my best to condense what we've done, and the underlying principles, into an easily digestible form. I am confident there are errors at times, or places where I have made simplifications that elide more nuanced underpinnings. Buyer beware.

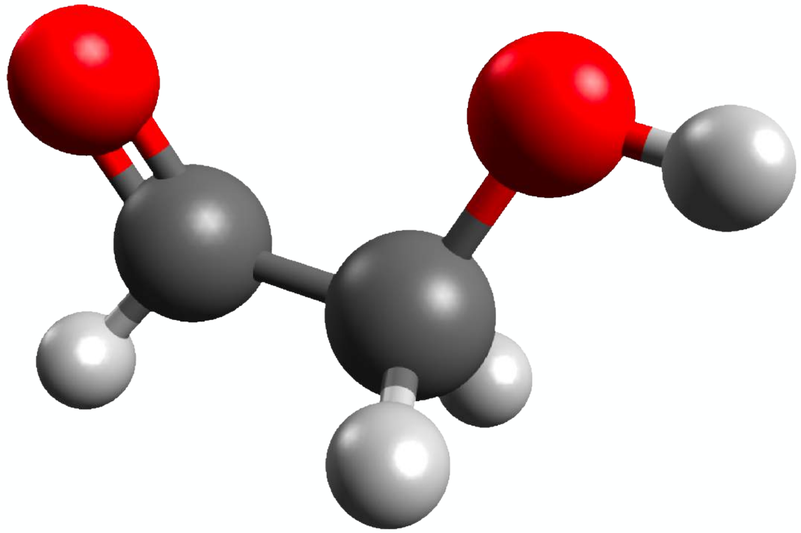
What is Chemical Intuition?

The ability to look at a molecule or set of molecules and quickly assess (with reasonable accuracy) their properties and how they will interact with specific other species or their environment

Let's Try

Glycolaldehyde





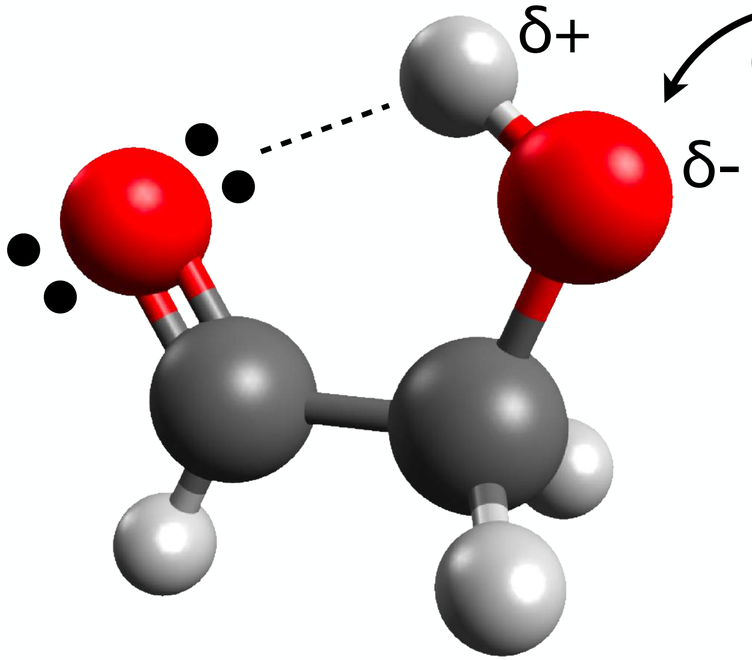
Which of these conformations of glycolaldehyde is more stable (lower in energy, happier)?

0

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Let's Try

Glycolaldehyde

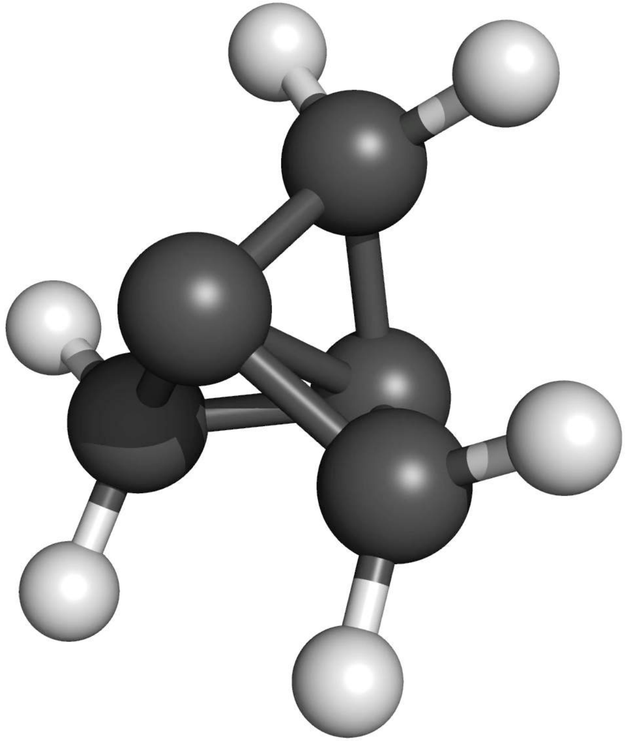


Electronegative
(Hogs electron density)

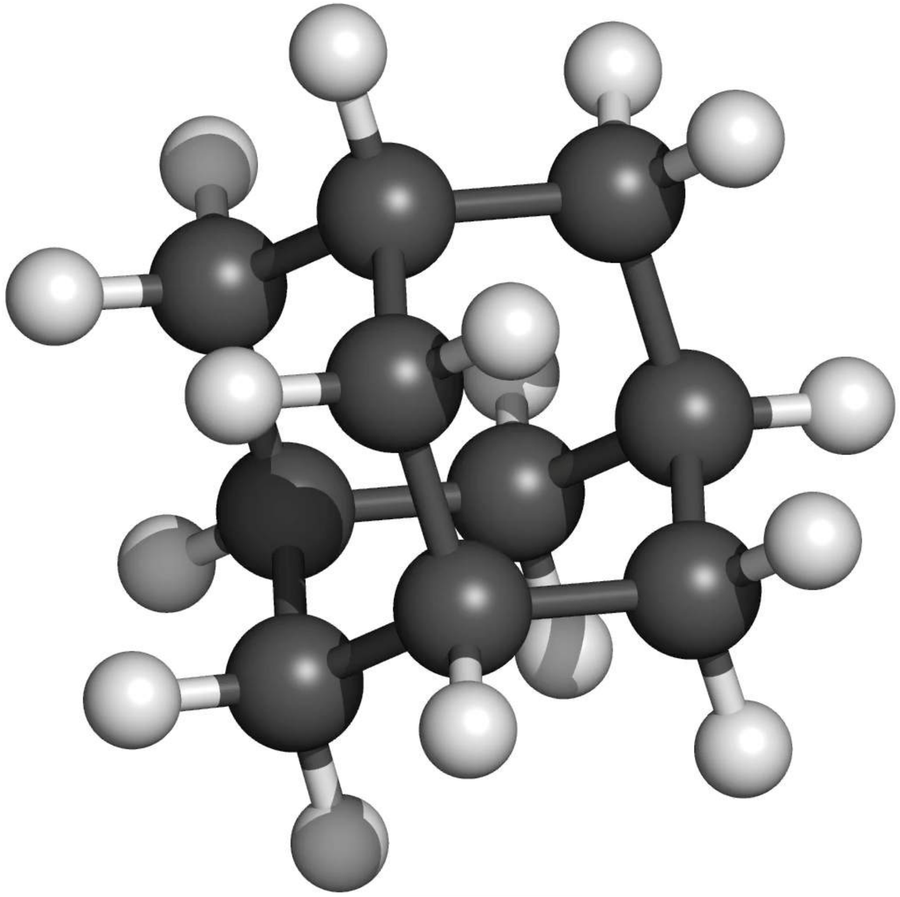




Let's Try (Again)



Propellane



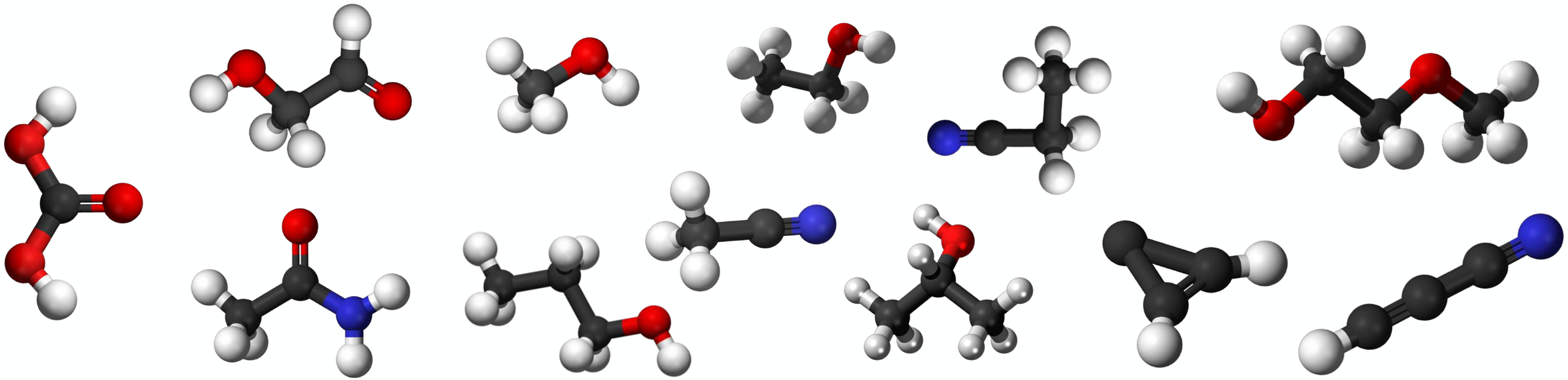
Adamantane

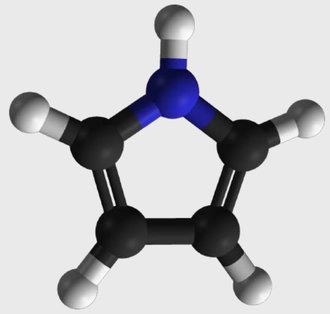
Which is more stable?

0

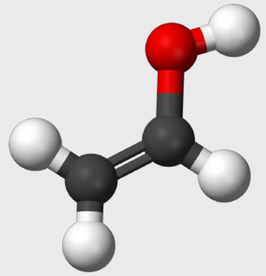
Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

One More …

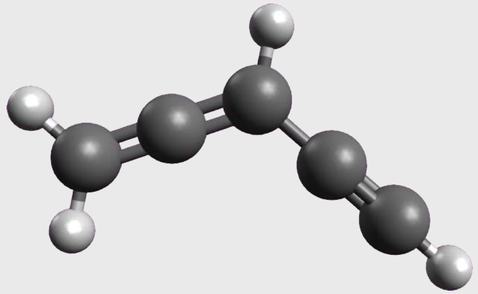




pyrrole



vinyl alcohol



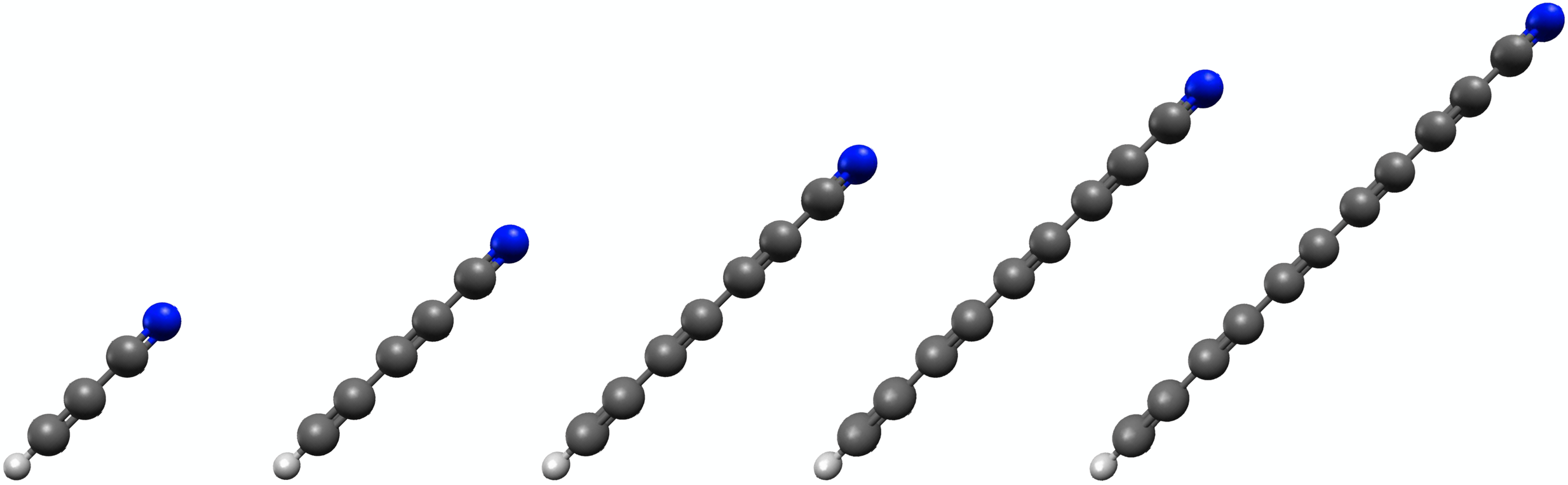
allenylacetylene

Rank in order of most likely (top) to be in this mixture to least likely (bottom)

0

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

JUST A SHOW OF HANDS

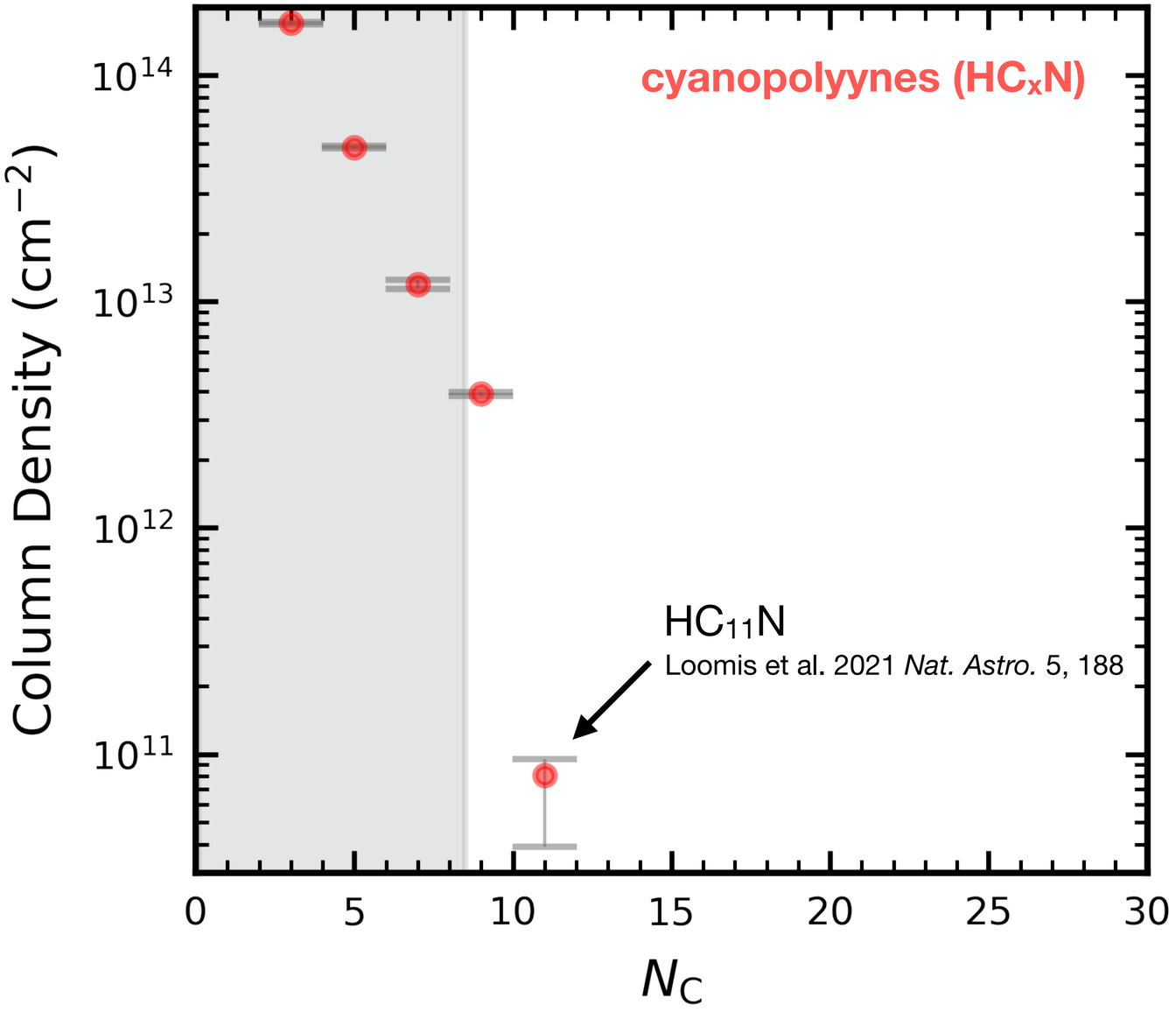


HC_3N HC_5N HC_7N HC_9N $HC_{11}N$

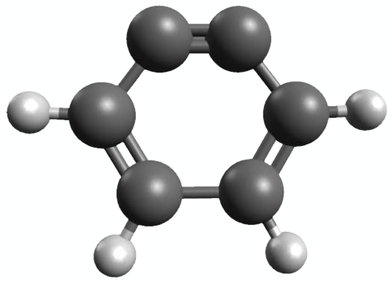
Higher / Lower Column Density

JUST A SHOW OF HANDS

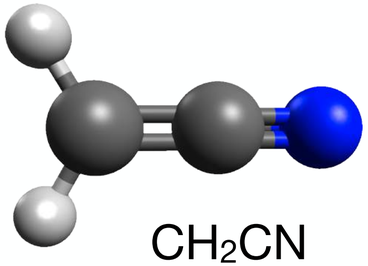
Xue et al. 2025 *ApJS*, in revision
McCarthy & McGuire 2025 *Ann. Rev. Phys. Chem.*, submitted



Online Again …



o-benzyne



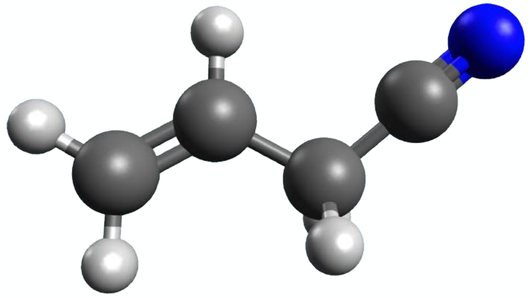
CH_2CN

HC_5N

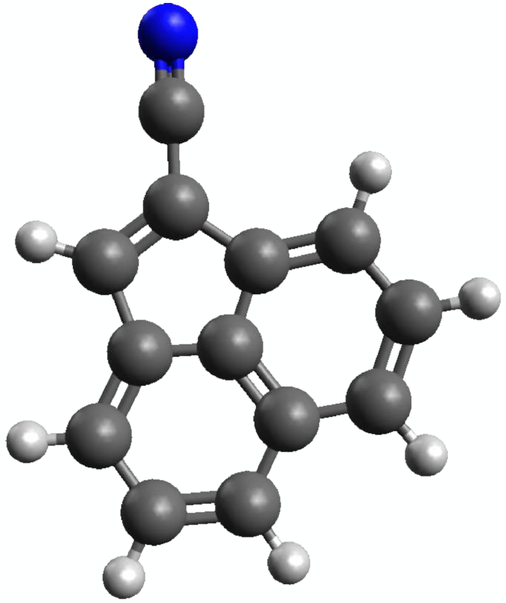




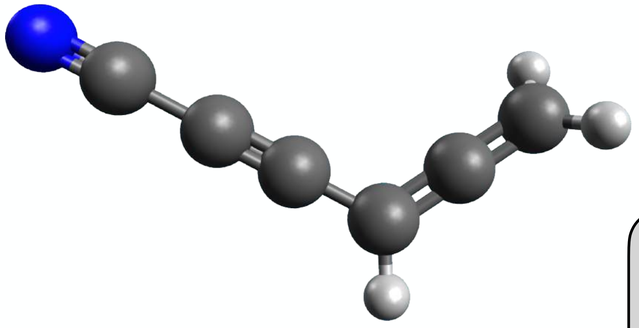
HC_5O



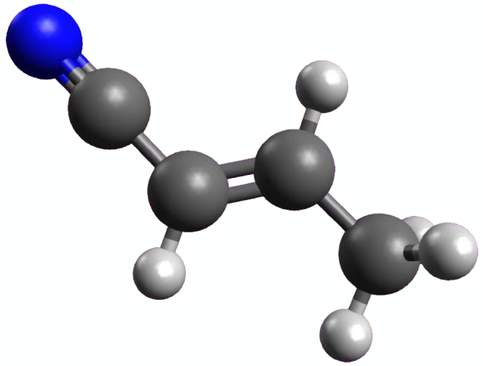
CH_2CHCH_2CN



1-cyanoacenaphthylene



$CH_2CCHCCCN$



$CH_3CHCHCN$

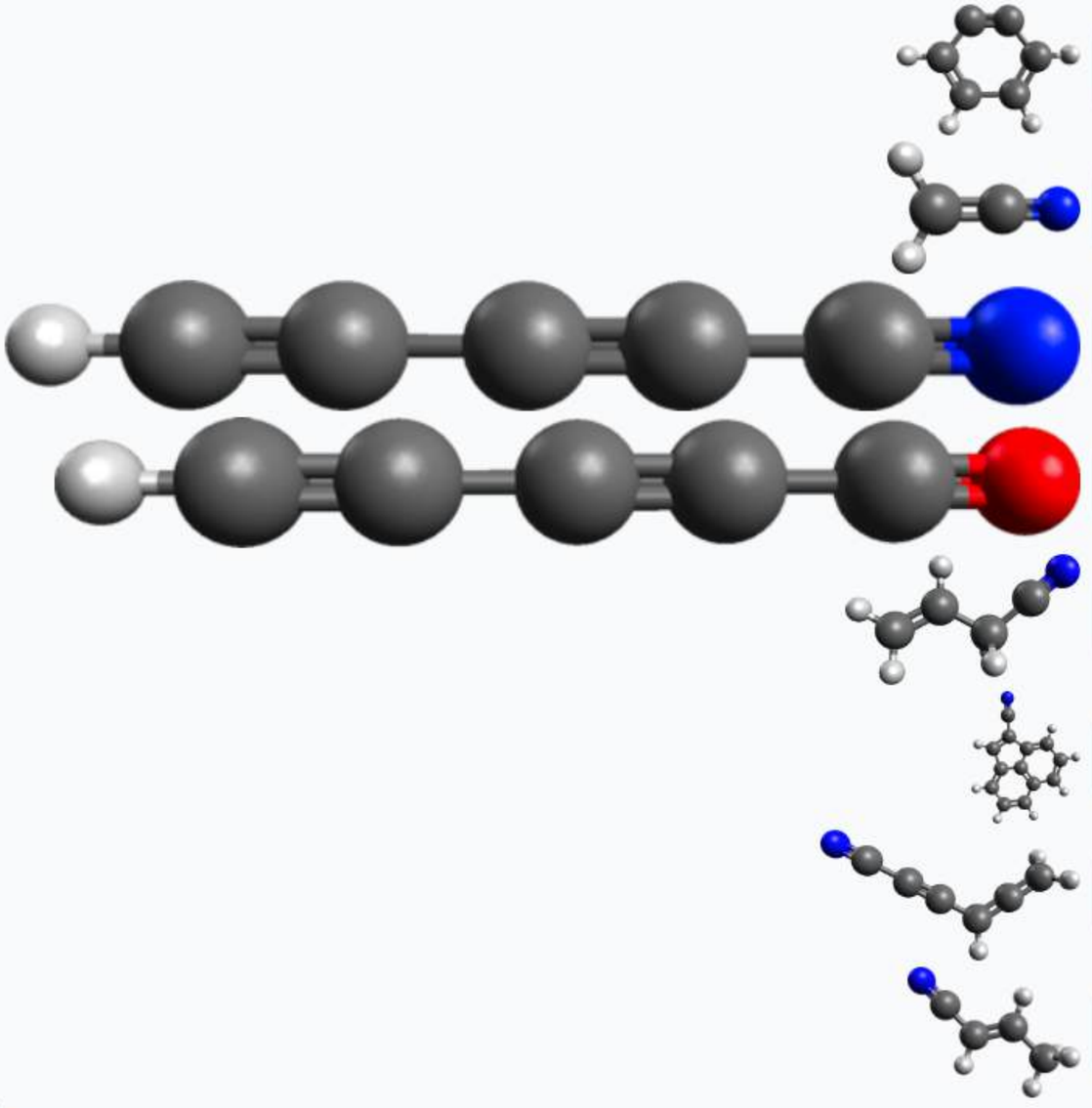
TMC-1

Cold, Dark, Starless Cloud

T_{ex} ~8 K

Rank these from highest column density (top) to lowest (bottom)

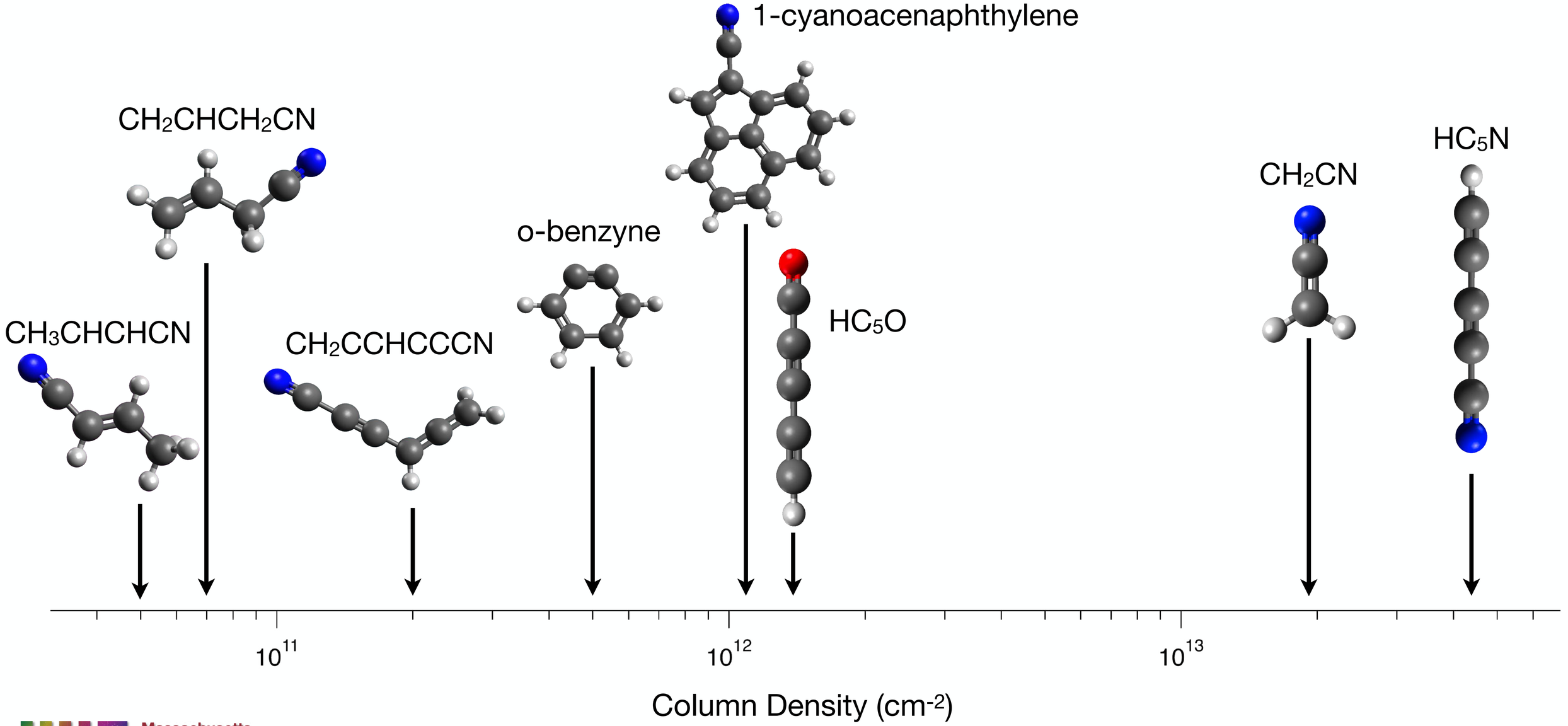
0



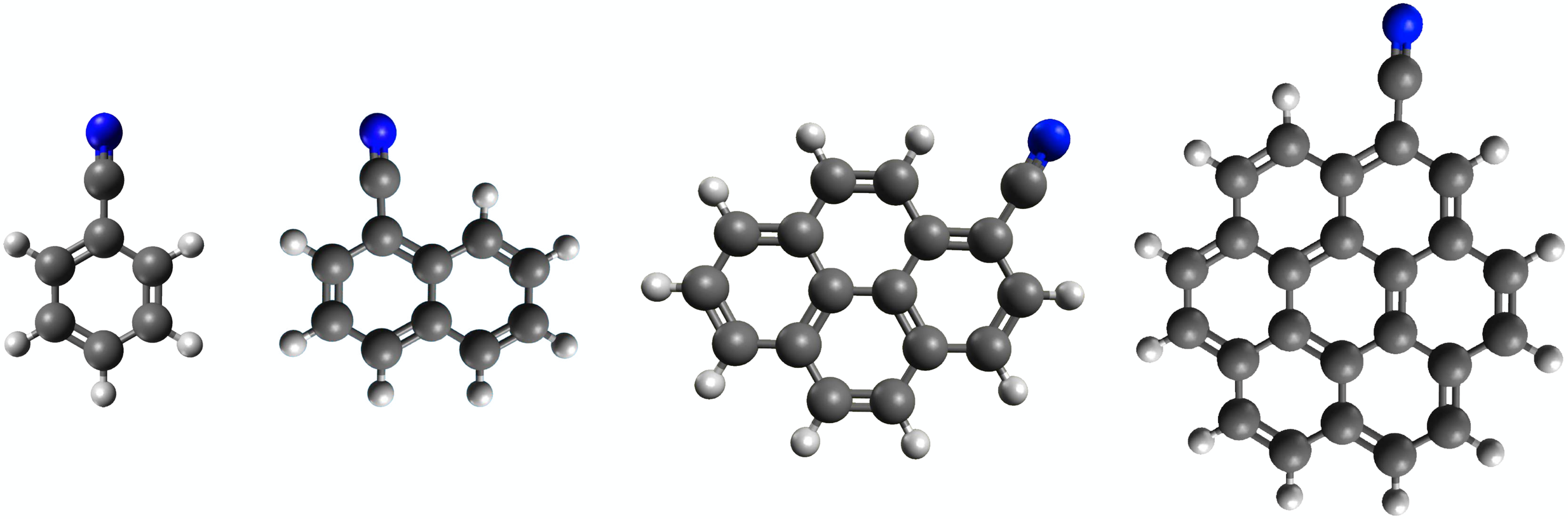
Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

ONLINE AGAIN …

Xue et al. 2025 *ApJS*, in revision
Cernicharo et al. 2021 *A&A* L9
Cernicharo et al. 2022 *A&A* L5
Shingledecker et al. 2021 *A&A* L12



JUST A SHOW OF HANDS



benzonitrile

1-cyanonaphthalene

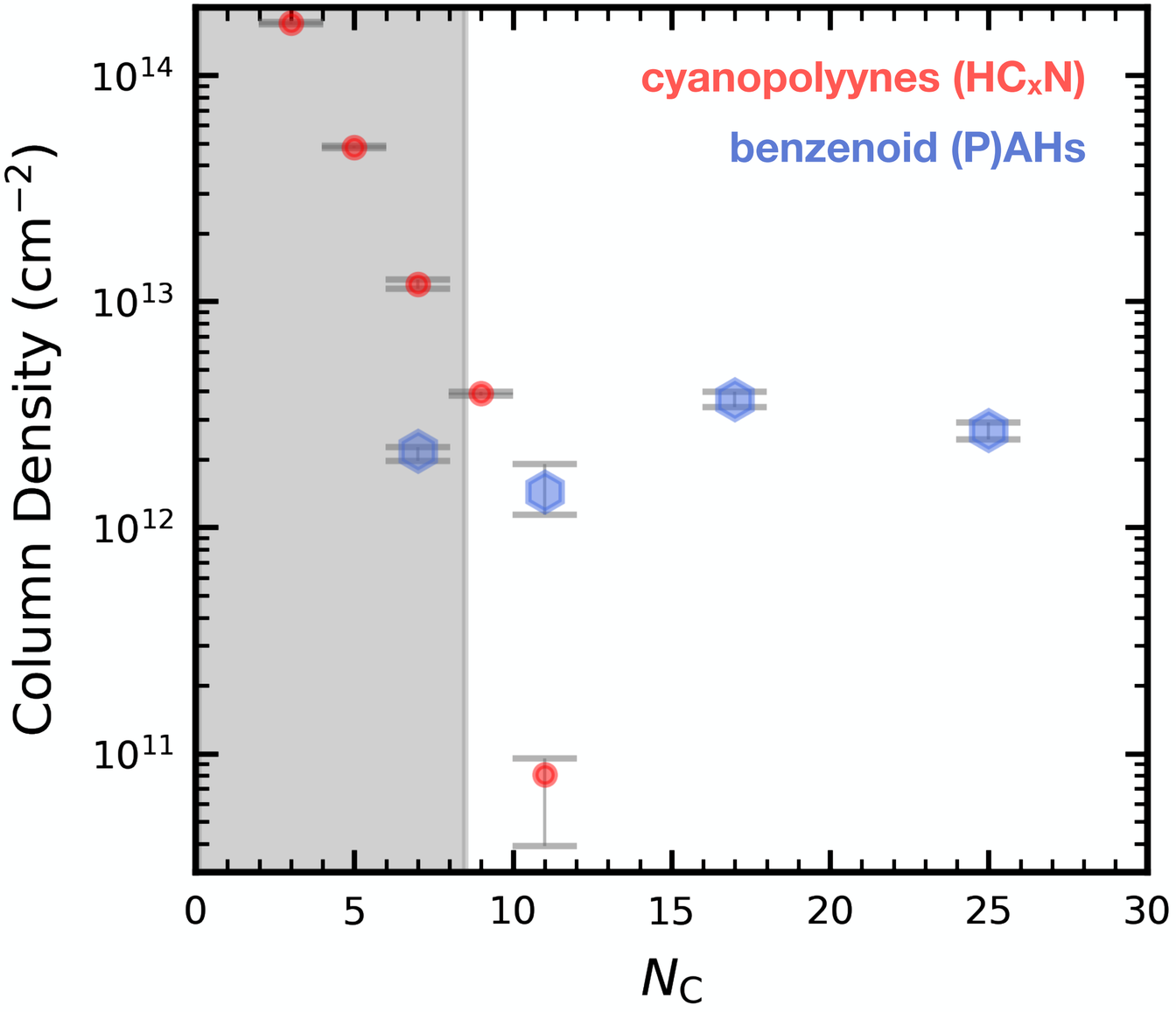
1-cyanopyrene

cyanocoronene

Higher / Lower Column Density

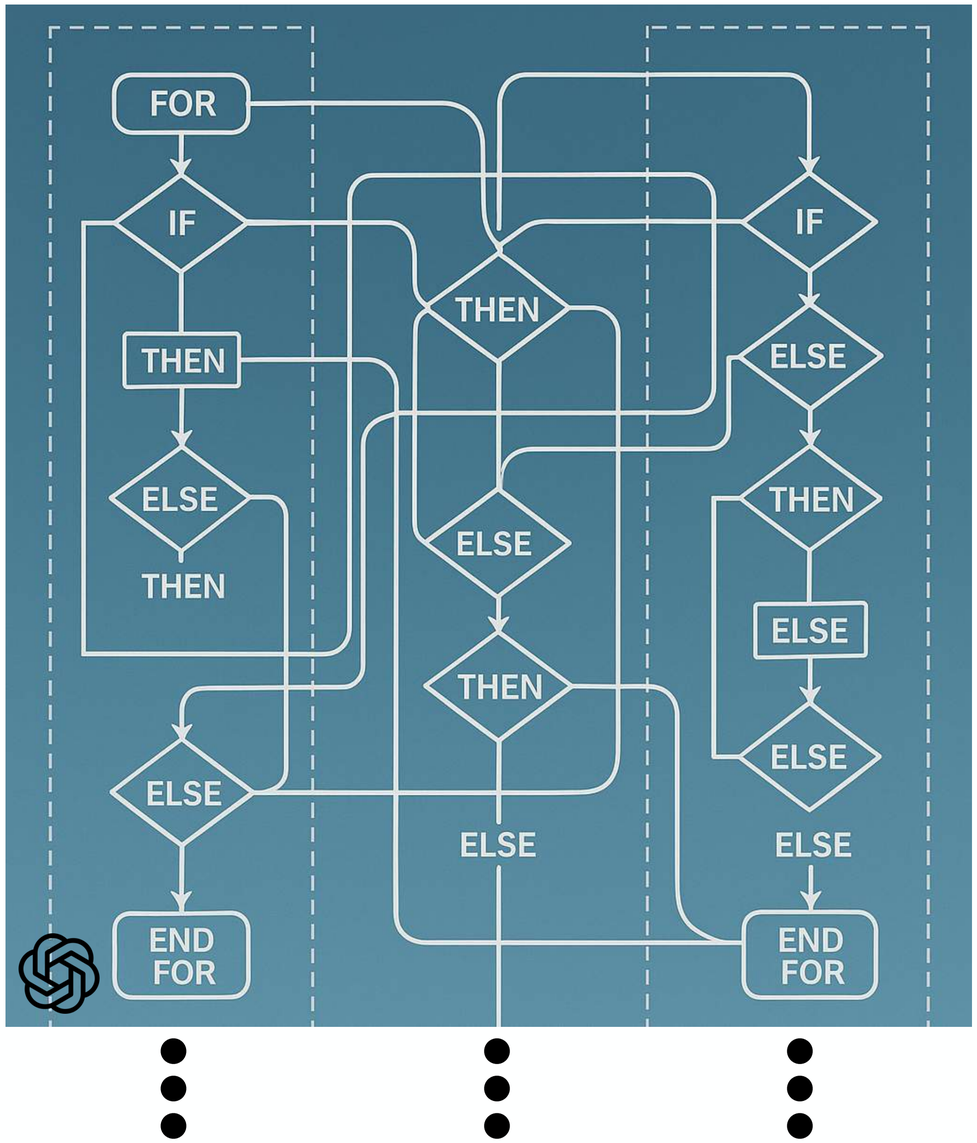
JUST A SHOW OF HANDS

Xue et al. 2025 *ApJS*, in revision
Loomis et al. 2021 Nat. Astro. 5, 188
McCarthy & McGuire 2025 *Ann. Rev. Phys. Chem.*, submitted



Massachusetts Institute of Technology

Artificial Chemical Intuition



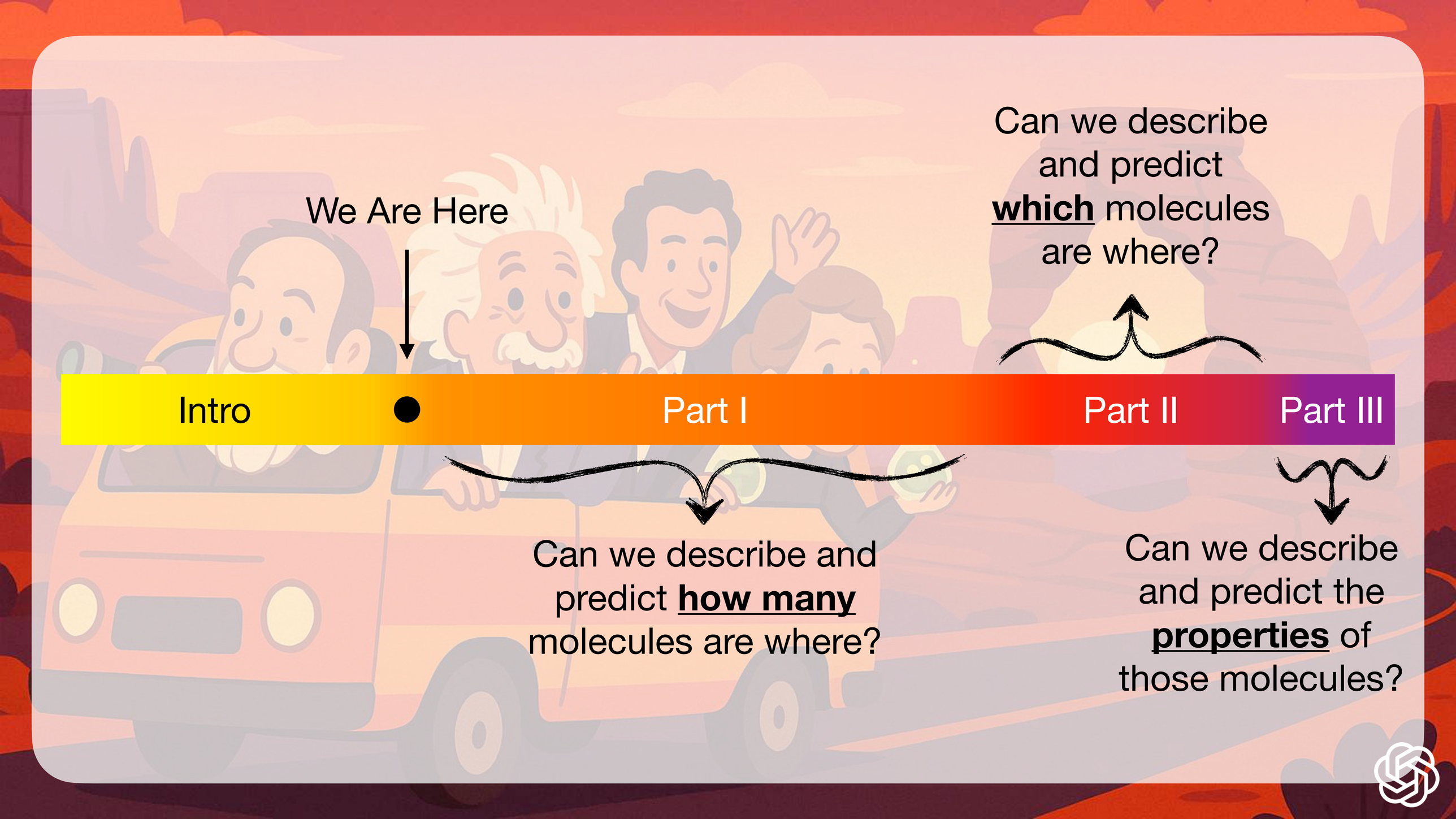
Goal

Use ML to quickly and efficiently get an ‘intuition’ of how a molecular identity maps to a desired property

Abundance

Chemical Similarity

Physical Properties


We Are Here
Can we describe and predict **which** molecules are where?
Intro
Part I
Part II
Part III
Can we describe and predict **how many** molecules are where?
Can we describe and predict the **properties** of those molecules?

Part I

Abundances









Machine Learning of Chemical Inventories

Lee et al. 2021 *ApJL* 917, L6

Implementation of Rare Isotopologues into Machine Learning of the Chemical Inventory of the Solar-type Protostellar Source IRAS 16293-2422

Fried et al. 2023 *Digital Discovery* 2, 952

Explaining the Chemical Inventory of Orion KL through Machine Learning

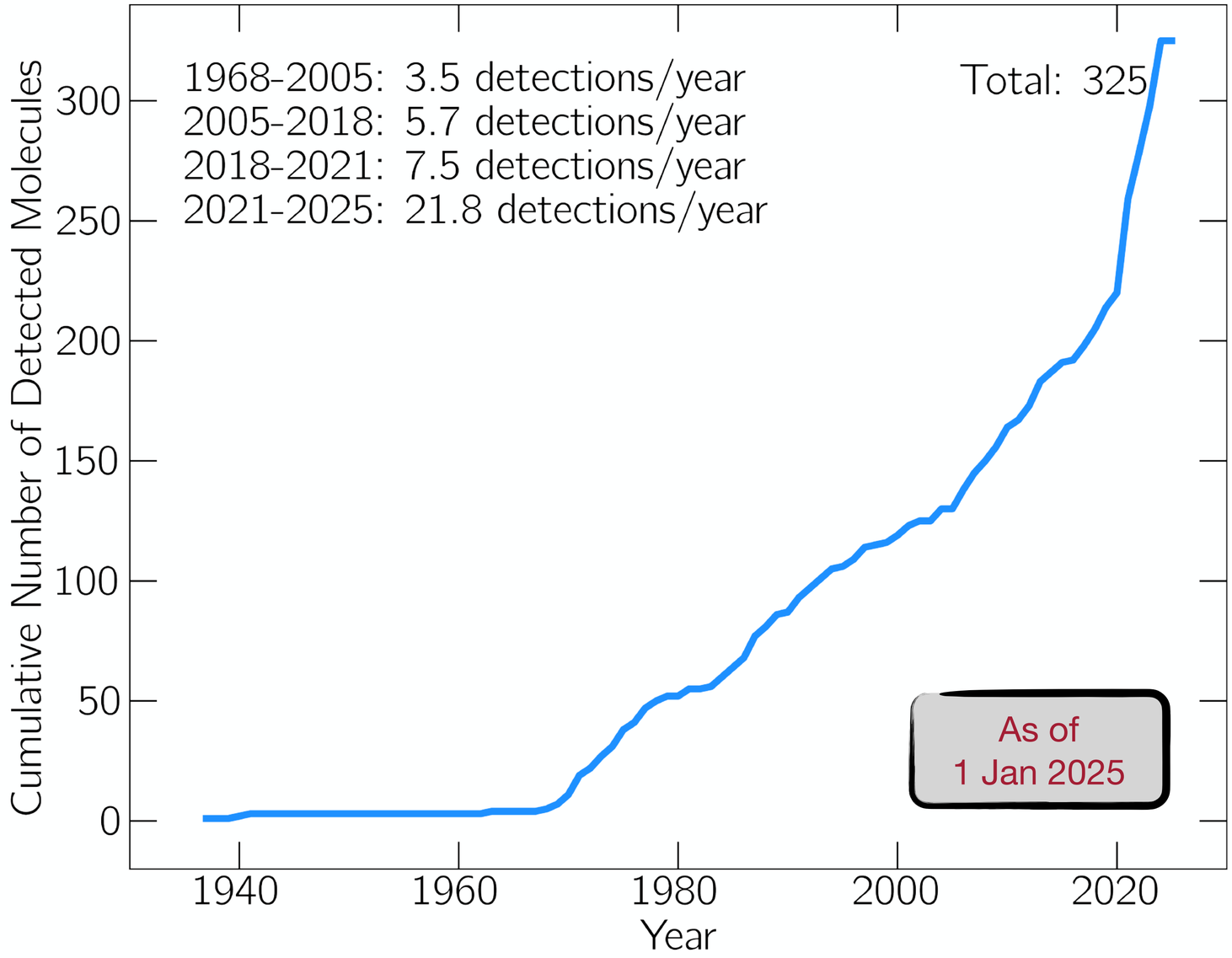
Scolati et al. 2023 *ApJ* 959, 108

Exploring Effects of Modified Machine Learning Pipelines of Astrochemical Inventories

Toru Shay et al. 2025 *ApJ* 985, 123

Motivation - What's the Problem We're Solving?

McGuire 2022 *ApJS* 259, 30



MOTIVATION - WHAT'S THE PROBLEM WE'RE SOLVING?

The millimeter-wave spectrum of methyl ketene and the astronomical search for it★

C. Bermúdez[1], B. Tercero[2], R. A. Motiyenko[1], L. Margulès[1], J. Cernicharo[3], Y. Ellinger[4], and J.-C. Guillemin[5]

Millimeter-wave spectroscopy and modeling of 1,2-butanediol★

Laboratory spectrum in the 59.6–103.6 GHz region and comparison with the ALMA archived observations

A. Vigorito[1], C. Calabrese[1,2,3,**], S. Melandri[1], A. Caracciolo[1,***], S. Mariotti[4], A. Giannetti[4], M. Massardi[4], and A. Maris[1]

Submillimeter spectroscopy and astronomical searches of vinyl mercaptan, C_2H_3SH★

M.-A. Martin-Drumel[1,2,**], K. L. K. Lee[2], A. Belloche[3], O. Zingsheim[1], S. Thorwirth[1], H. S. P. Müller[1], F. Lewen[1], R. T. Garrod[4], K. M. Menten[3], M. C. McCarthy[2], and S. Schlemmer[1]

Laboratory rotational spectrum and astronomical search for methoxyacetaldehyde★

L. Kolesniková[1], I. Peña[1], E. R. Alonso[1], B. Tercero[2], J. Cernicharo[3], S. Mata[1], and J. L. Alonso[1]

RADIOASTRONOMICAL OBSERVATIONS IN ORION KL AND THE SgrB2 LINE SURVEYS

Submillimeter wave spectroscopy of ethyl isocyanide and its searches in Orion★

L. Margulès[1], B. Tercero[2], J. C. Guillemin[3], R. A. Motiyenko[1], and J. Cernicharo

Laboratory Measurements and Astronomical Search for Thioacetamide

Assimo Maris,[†] Camilla Calabrese,[‡,§] Laura B. Favero,[∥] Luca Evangelisti,[†] Imanol Usabiaga,[†] Sergio Mariotti,[⊥] Claudio Codella,[#,▽] Linda Podio,[#] Nadia Balucani,[▽,○] Cecilia Ceccarelli,[▽] Bertrand LeFloch,[▽] and Sonia Melandri*,[†]

an Astronomical Search in Sgr B2(N), Orion-KL, and NGC 6334I

Elena R. Alonso[1,2], Brett A. McGuire[3,4,6], Lucie Kolesniková[1], P. Brandon Carroll[4], Iker León[1], Crystal L. Brogan[3], Todd R. Hunter[3], Jean-Claude Guillemin[5], and Jose L. Alonso[1]

Motivation - What's the Problem We're Solving?

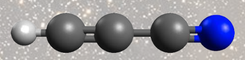
A Problem

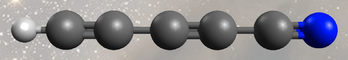
There are an ~infinite number of molecular targets, but finite human laboratory resources to study them.

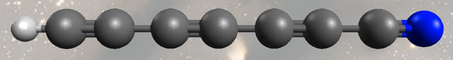
We clearly do a suboptimal job choosing targets.

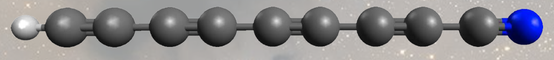
How are we doing this now?

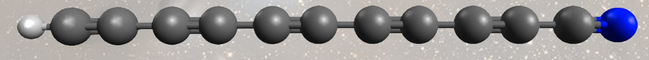
Method 1: Logical Next Steps

HC_3N 

HC_5N 

HC_7N 

HC_9N 

$HC_{11}N$ 

Limited to very similar species and incremental shifts

Non-detection of $HC_{11}N$ towards TMC-1: constraining the chemistry of large carbon-chain molecules

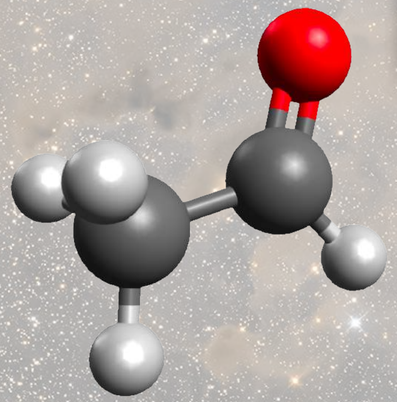
Ryan A. Loomis,[1★] (
Brett A. McGuire,[4,5]
Shawn T. Booth,[6‡] P

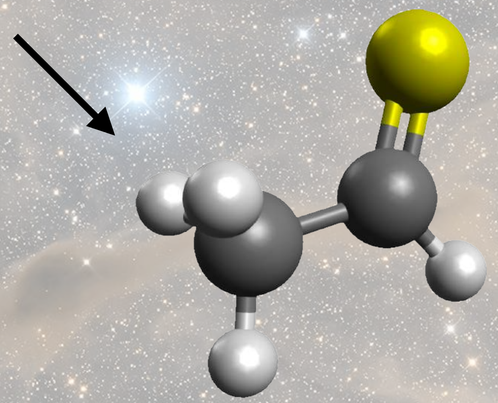
An investigation of spectral line stacking techniques and application to the detection of $HC_{11}N$

Ryan A. Loomis[1✉], Andrew M. Burkhardt[2], Christopher N. Shingledecker[3,4,5], Steven B. Charnley[6], Martin A. Cordiner[6,7], Eric Herbst[8,9], Sergei Kalenskii[10], Kin Long Kelvin Lee[2], Eric R. Willis[8], Ci Xue[8], Anthony J. Remijan[2], Michael C. McCarthy[2] and Brett A. McGuire[1,2,11✉]

Microwave Emission of the $2_{11} \rightarrow 2_{12}$ Rotational Transition in Interstellar Acetaldehyde

N. Fourikis,[A] M. W. Sinclair,[A] B. J. Robinson,[A] P. D. Godfrey[B] and R. D. Brown[B]

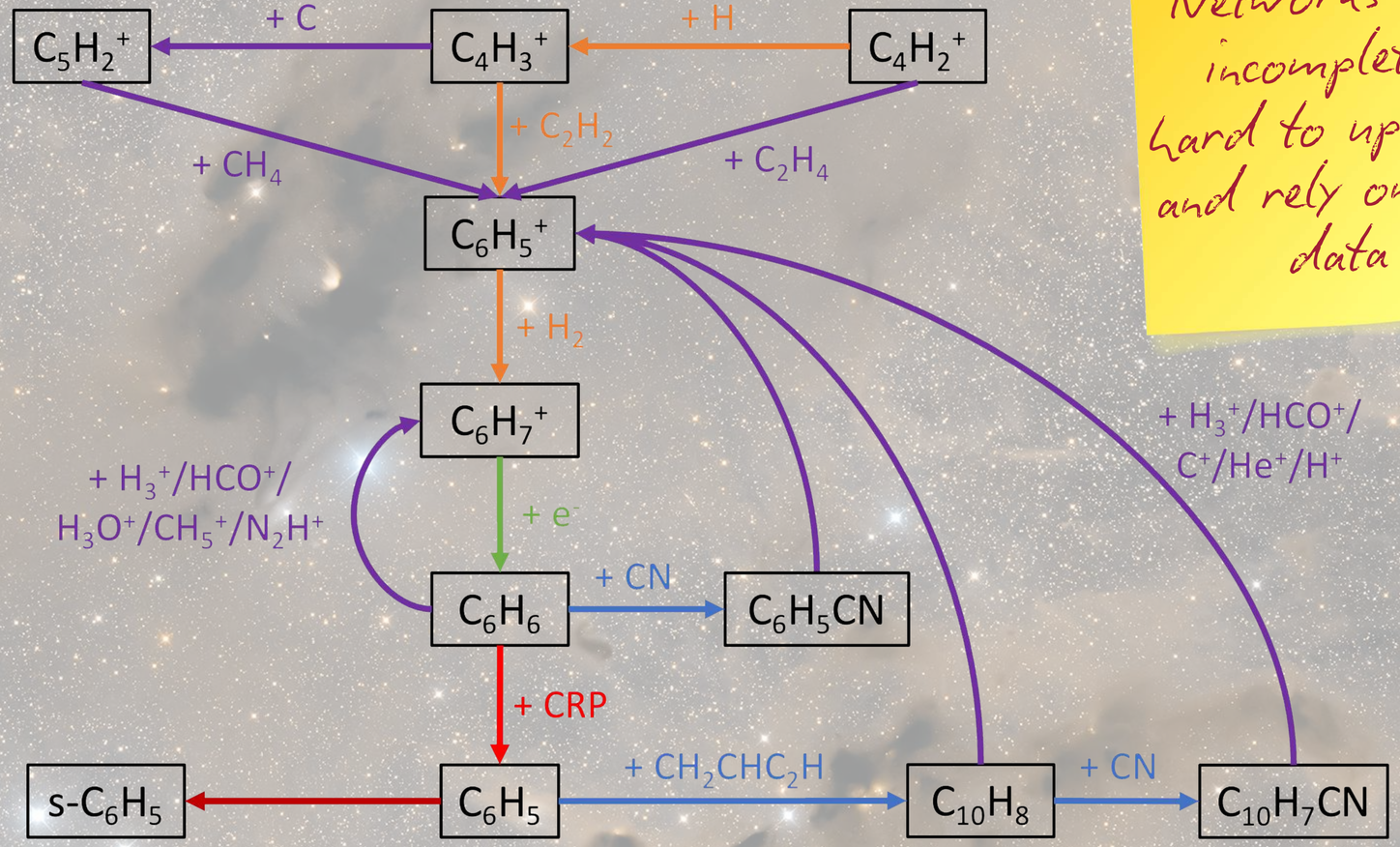




Detection of thioacetaldehyde (CH_3CHS) in TMC-1: Sulfur-oxygen differentiation along the hydrogenation sequence★

M. Agúndez[1,★★], G. Molpeceres[1], C. Cabezas[1], N. Marcelino[2,3], B. Tercero[2,3], R. Fuentetaja[1], P. de Vicente[3], and J. Cernicharo[1,★★]

Method 2: Use Chemical Models



Networks are incomplete, hard to update, and rely on iffy data

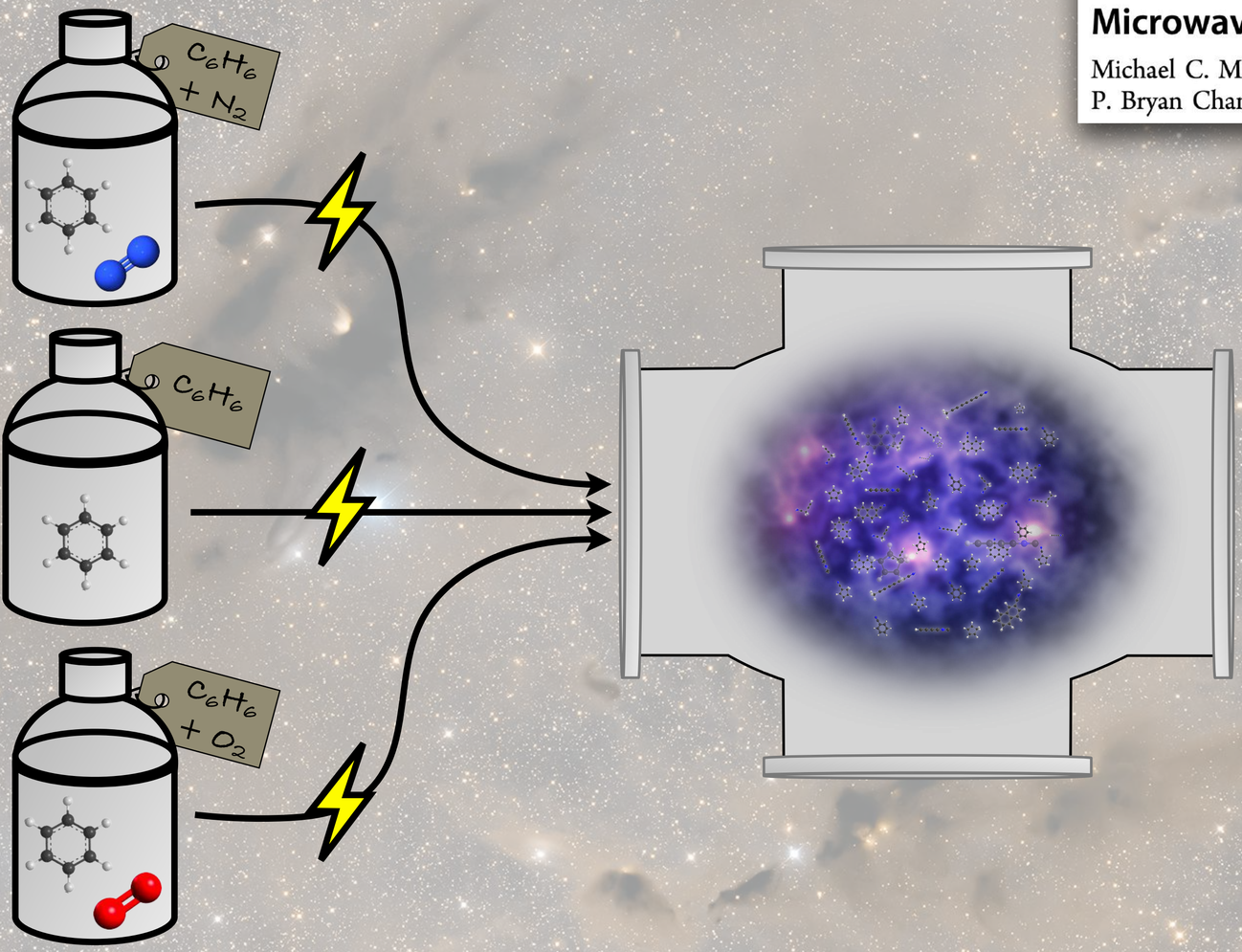
Byrne et al. 2024 *PCCP* 26, 26734

Method 3: Blow Sh*t Up

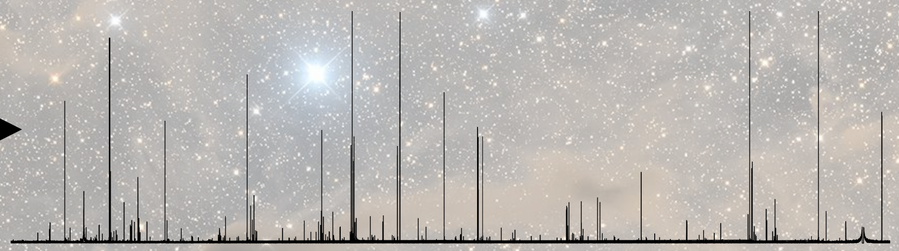
Exhaustive Product Analysis of Three Benzene Discharges by Microwave Spectroscopy

Michael C. McCarthy,* Kin Long Kelvin Lee, P. Brandon Carroll, Jessica P. Porterfield, P. Bryan Changala, James H. Thorpe, and John F. Stanton

2020 *JPCA* 124, 5170

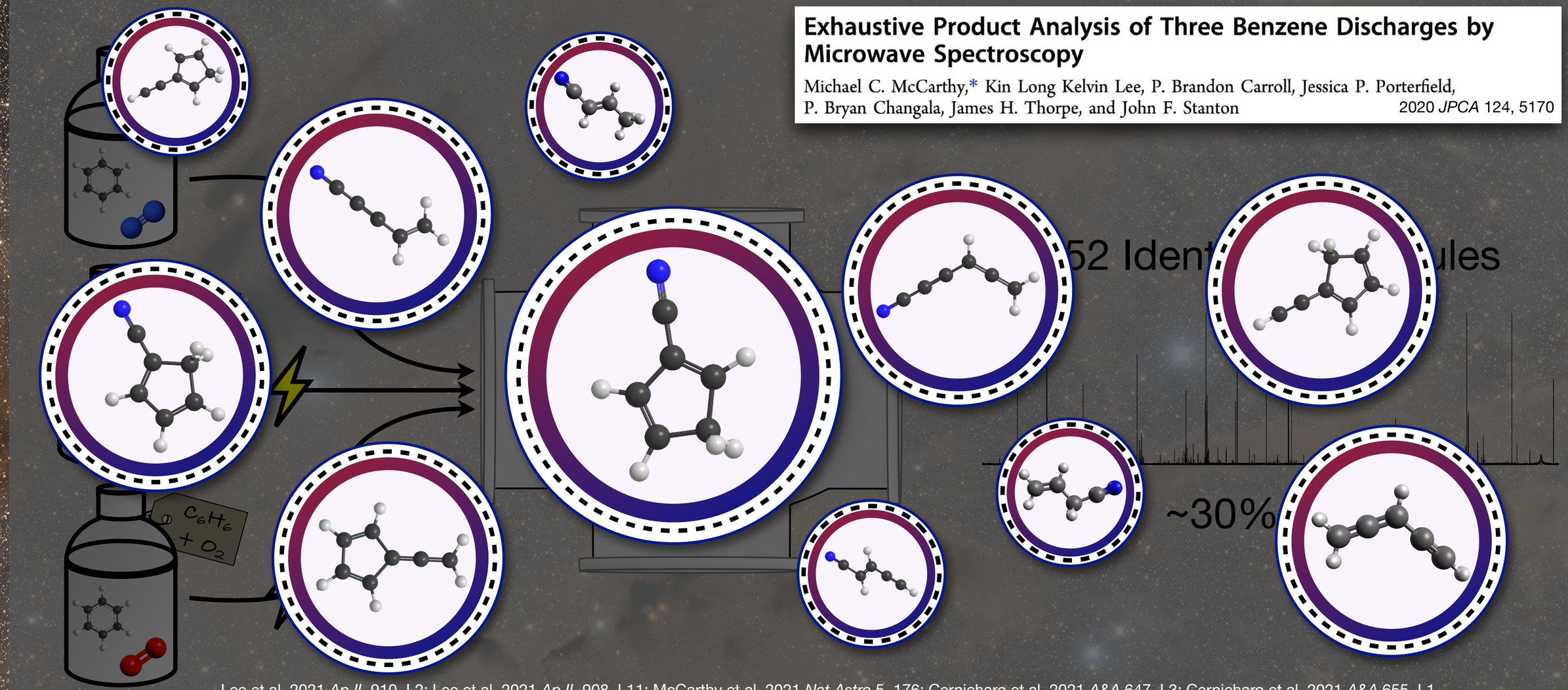


152 Identified Molecules



~30% New

Method 3: Blow Sh*t Up



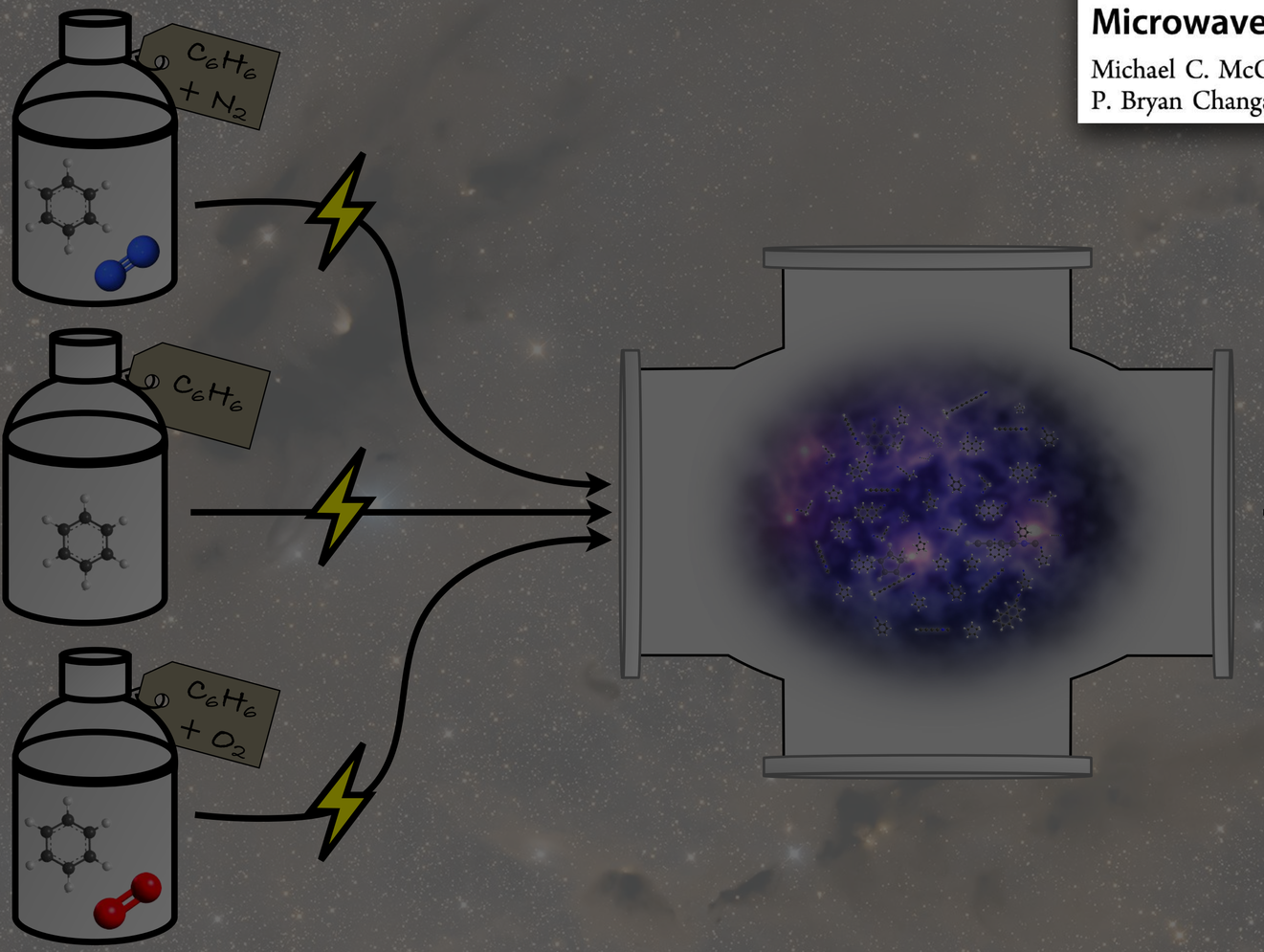
Lee et al. 2021 *ApJL* 910, L2; Lee et al. 2021 *ApJL* 908, L11; McCarthy et al. 2021 *Nat Astro* 5, 176; Cernicharo et al. 2021 *A&A* 647, L3; Cernicharo et al. 2021 *A&A* 655, L1
Cernicharo et al. 2022 *A&A* 663, L5; Cernicharo et al. 2022 *A&A* 663, L9; Shingledecker et al. 2021 *A&A* 652, L12

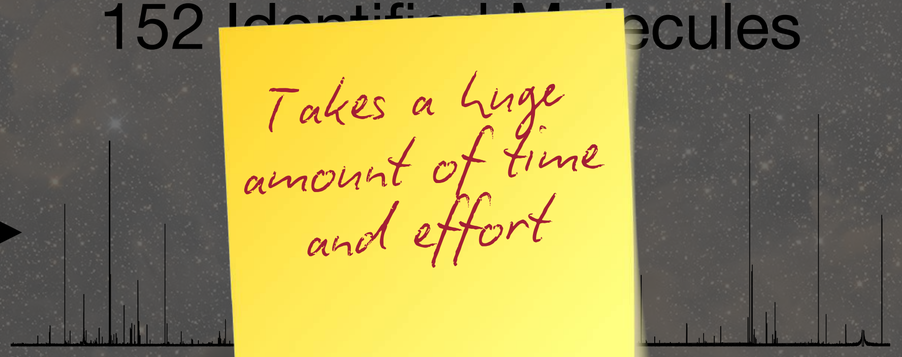
Method 3: Blow Sh*t Up

Exhaustive Product Analysis of Three Benzene Discharges by Microwave Spectroscopy

Michael C. McCarthy,* Kin Long Kelvin Lee, P. Brandon Carroll, Jessica P. Porterfield, P. Bryan Changala, James H. Thorpe, and John F. Stanton

2020 *JPCA* 124, 5170





152 Identified Molecules

Takes a huge amount of time and effort

30% New

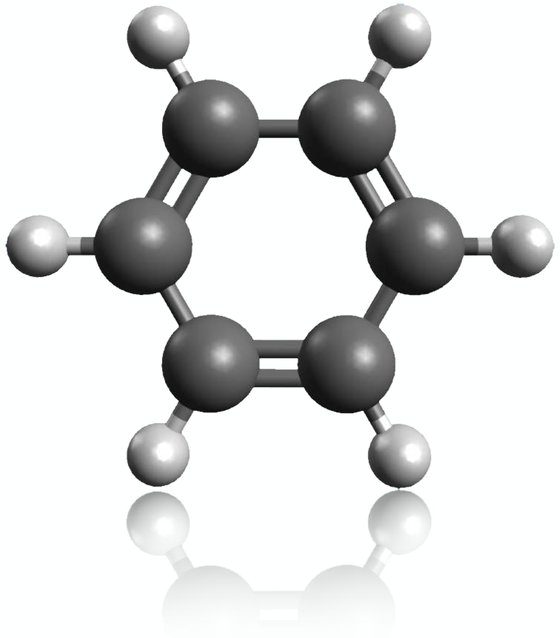
The Goal

Use ML + automation to combine the best aspects of each method, while minimizing their downsides

Molecular Feature Vectors

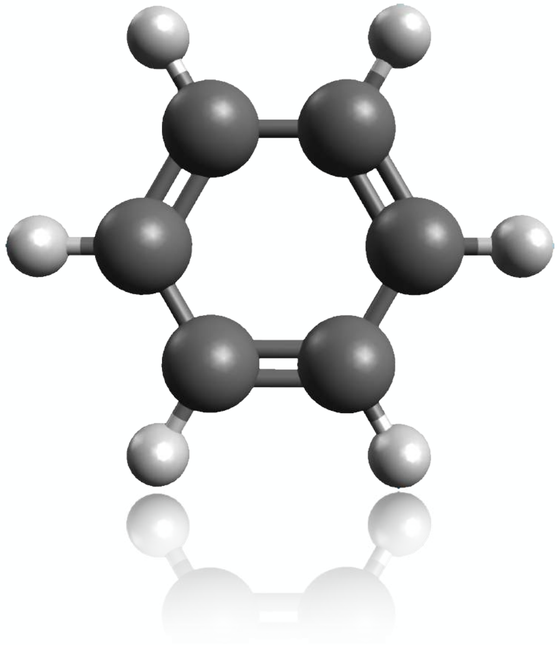
Human Readable

Machine Readable
+ useful(?)



$$\begin{bmatrix} 0.23 \\ 1.24 \\ -36.1 \\ \\ \\ \dots \\ \\ \\ \\ 42.0 \end{bmatrix}$$

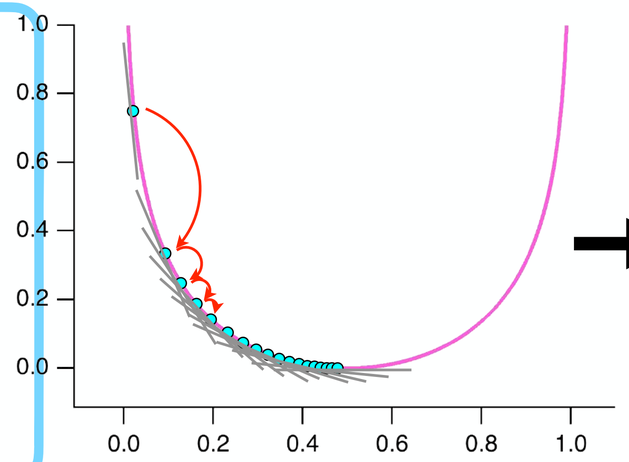
Molecular Feature Vectors



c1ccccc1

C_6H_6

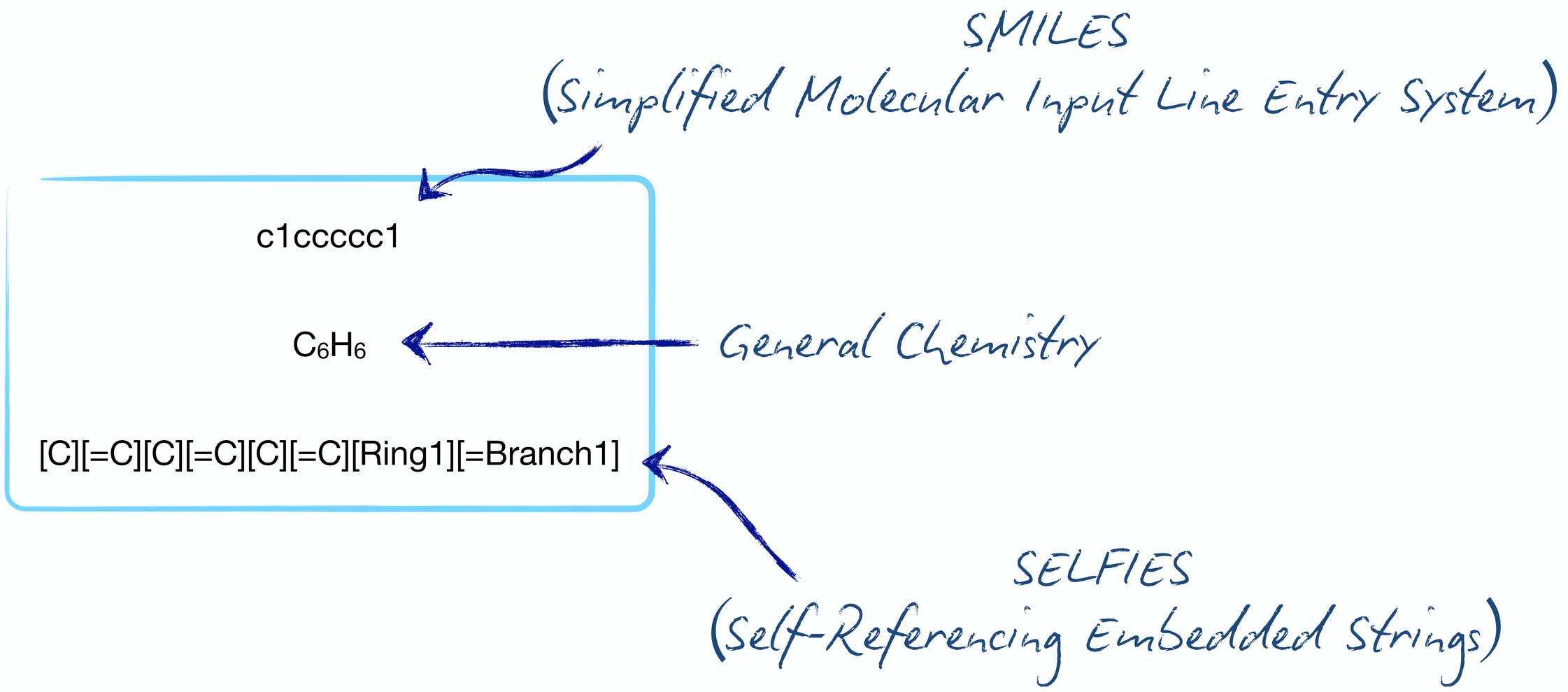
[C][=C][C][=C][C][=C][Ring1][=Branch1]



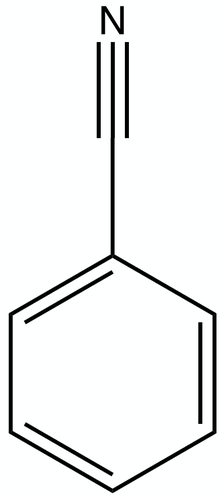
0.23
1.24
-36.1
...
42.0

Molecular Feature Vectors

Weininger 1988 *J. Chem. Inf. Comp. Sci.* 28, 31
Krenn et al. 2020 *Mach. Learn. Sci. Tech.* 1, 045024



Molecular Feature Vectors



Benzonitrile
N#CC1=CC=CC=C1

Q: How many unique carbon atoms in benzonitrile?

Possible Tokens

C
=C (or C=)
C1
N#

How many unique carbon atoms in benzonitrile (C6H5CN)?

0

1 0%

2 0%

3 0%

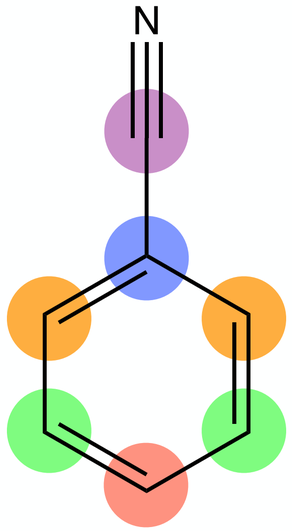
4 0%

5 0%

6 0%

Start the presentation to see live content. For screen share software, share the entire screen. Get help at **pollev.com/app**

Molecular Feature Vectors



Benzonitrile
N#CC1=CC=CC=C1

SMILES characters alone are insufficient. They lack context.

We need better tokens, and better context.

Possible Tokens

C
=C (or C=)
C1
N#


Is four a lot?
Depends on the context.
Dollars? No.
Incorrectwhoquotes
Murders? Yes.

Context is Key

I wish I could have a picnic on the bank of the river Seine.

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

I wish I had more money in my bank account.

I once made a wish that I was a fish.

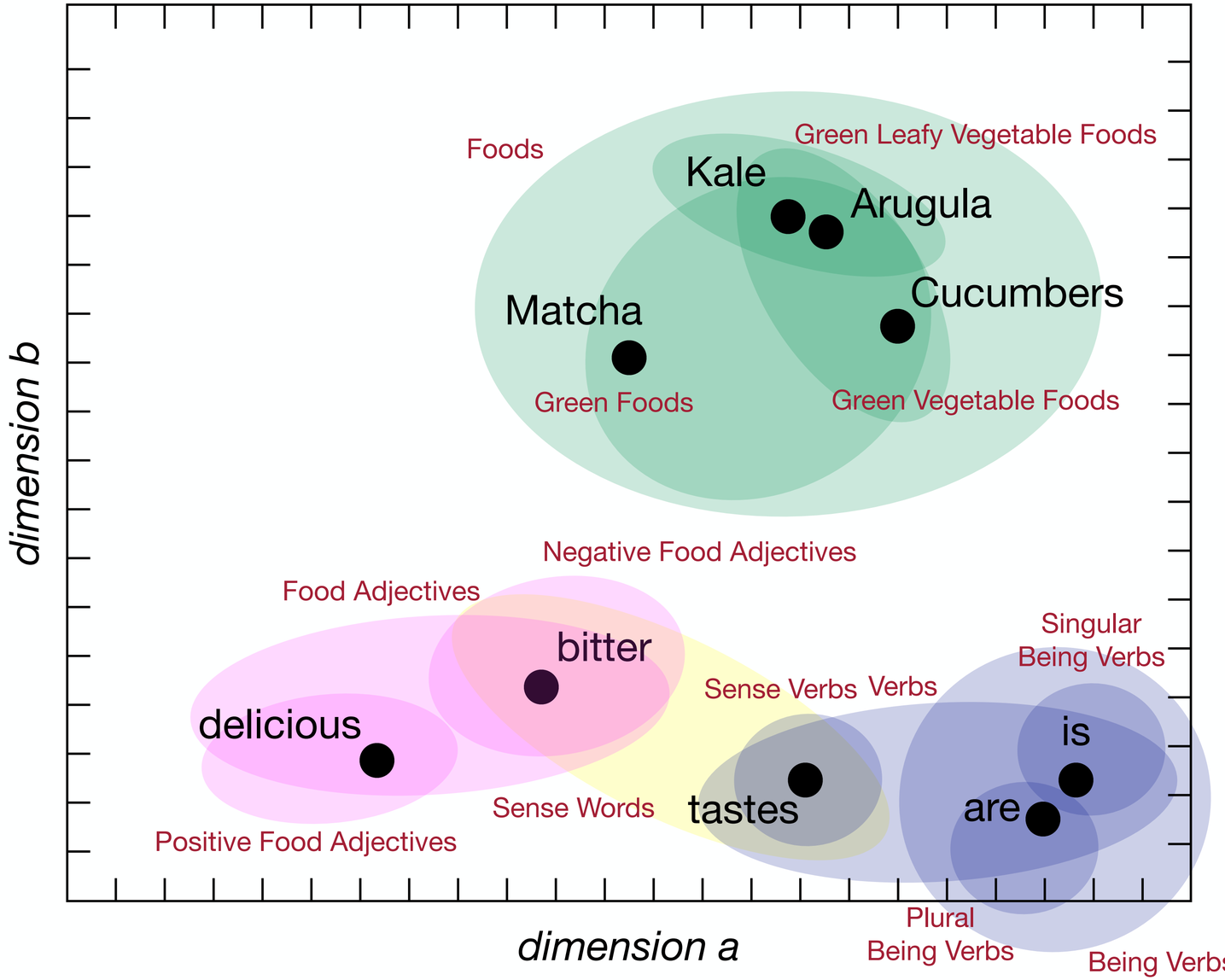
WORD2VEC

"Corpus" of context sentences.
e.g. /r/food on Reddit

"Kale is bitter."
"Arugula tastes bitter."
"Cucumbers are delicious."
"Matcha is delicious."
…

Goal is to create meaningful mathematical vector representations of words that carry information about their use and context

Similar words should be nearby one another in vector space



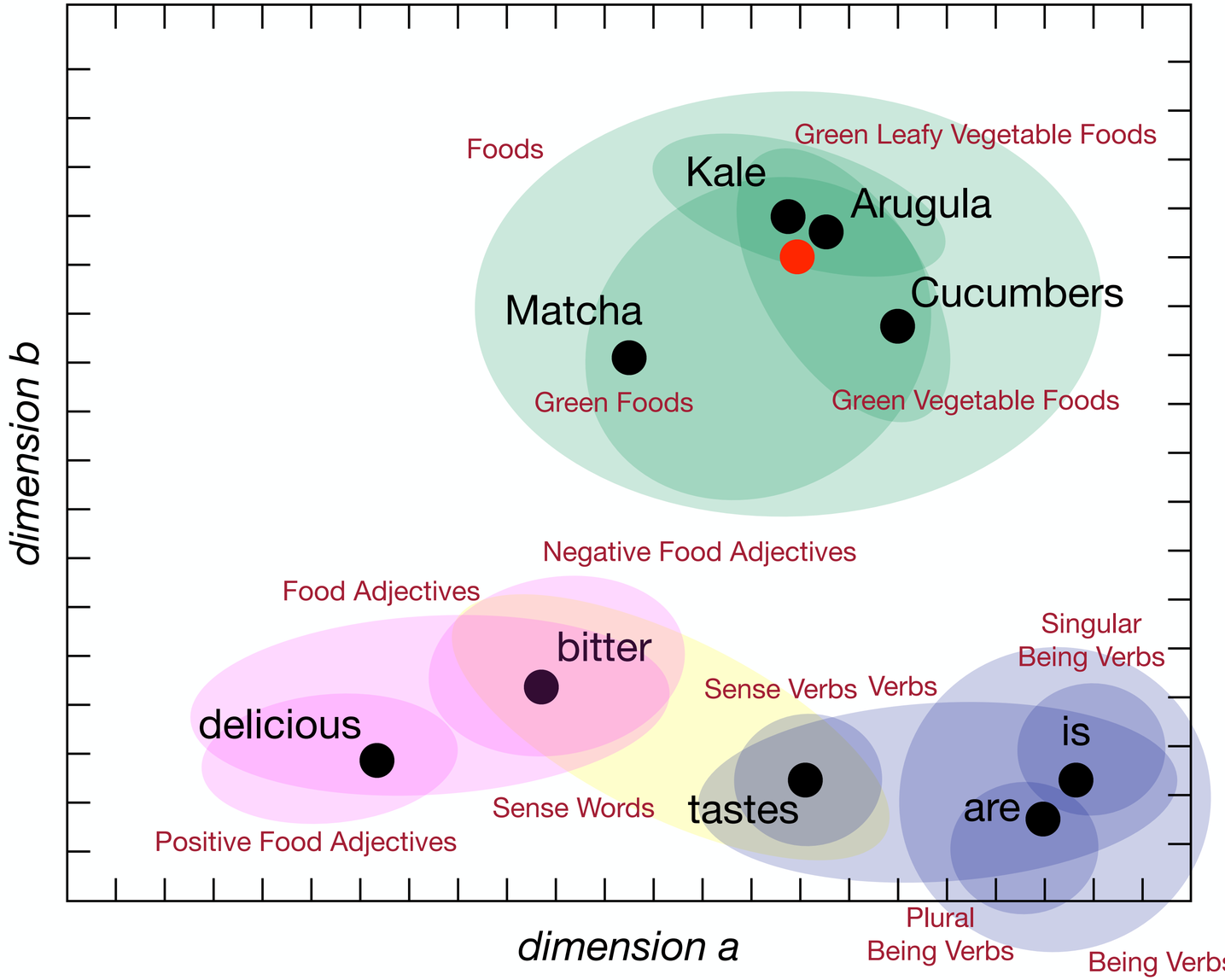
WORD2VEC

Do you want some mizuna?



Goal is to create meaningful mathematical vector representations of words that carry information about their use and context

Similar words should be nearby one another in vector space



WORD2VEC

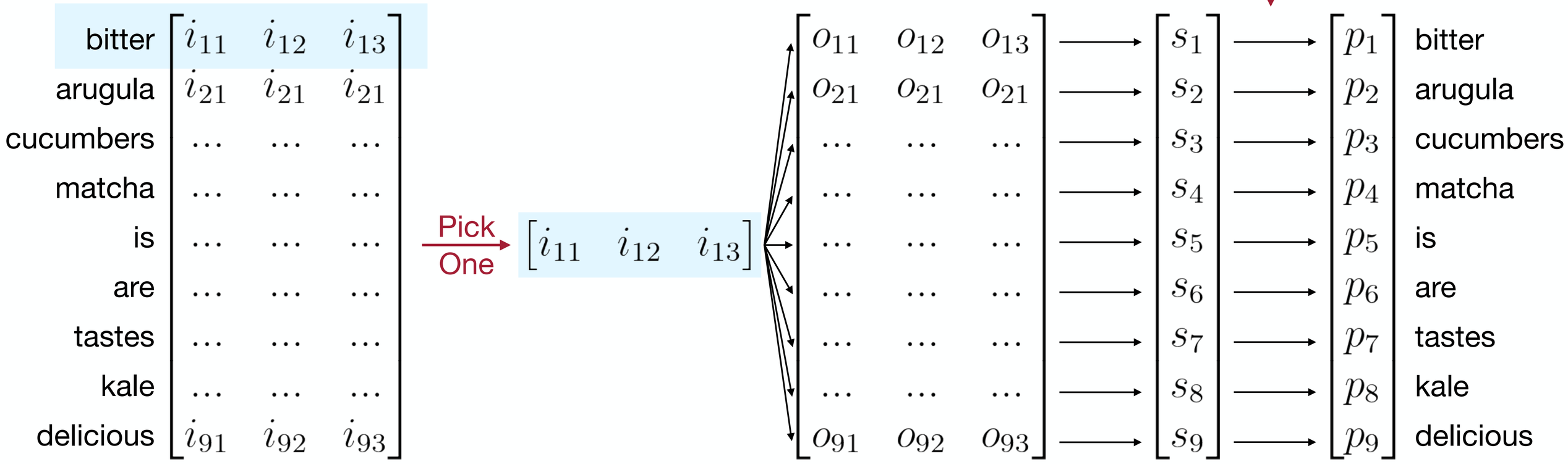
The closer two vectors, the larger the dot product

→ Assume 3-dimension vector representation

(Actually 300-d typically)

Dot Product

Softmax

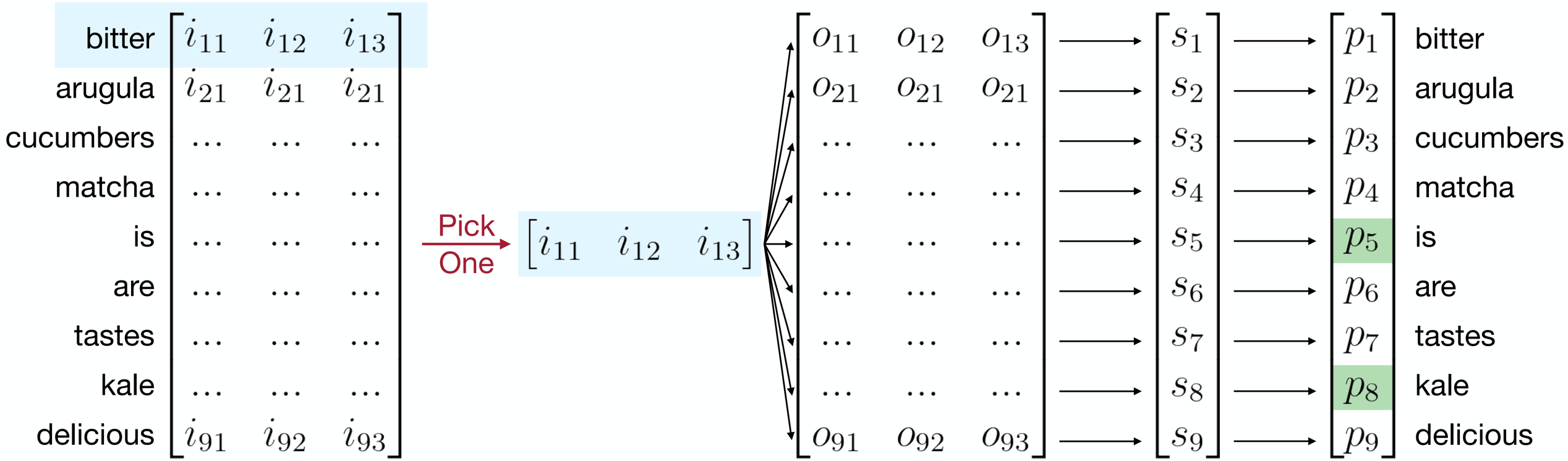


"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

WORD2VEC

"Kale is bitter." "Arugula tastes bitter." "Cucumbers are delicious." "Matcha is delicious."

For each word → For each context with word → Take dot products → Gradient descent to maximize p_x for context words by changing input and output vectors

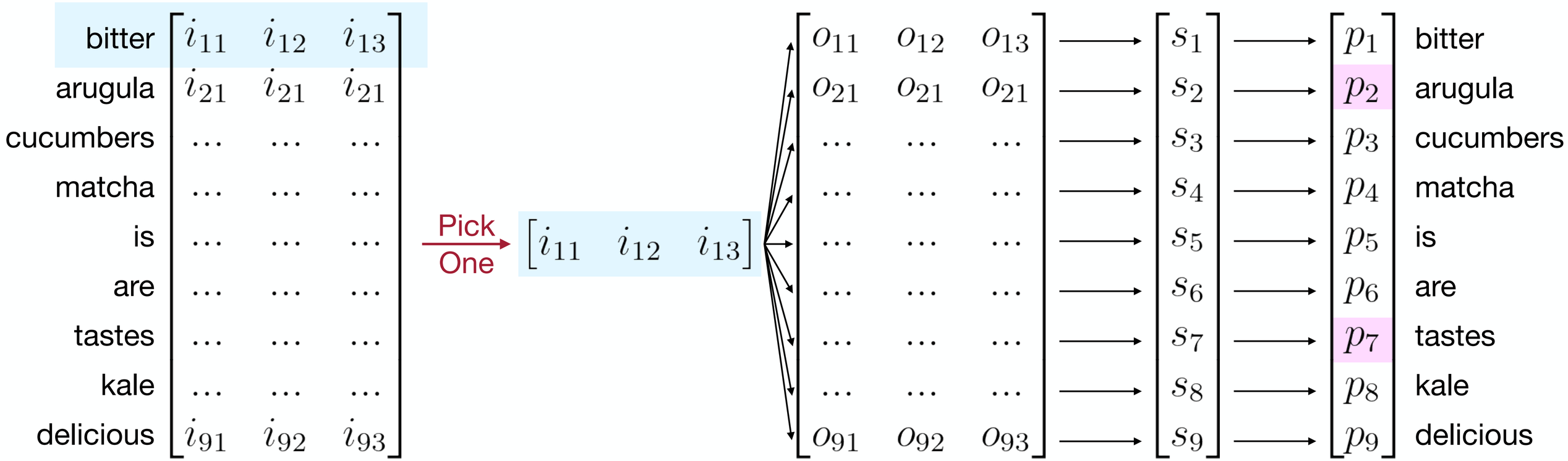


"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

WORD2VEC

"Kale is bitter." "Arugula tastes bitter." "Cucumbers are delicious." "Matcha is delicious."

For each word → For each context with word → Take dot products → Gradient descent to maximize p_x for context words by changing input and output vectors



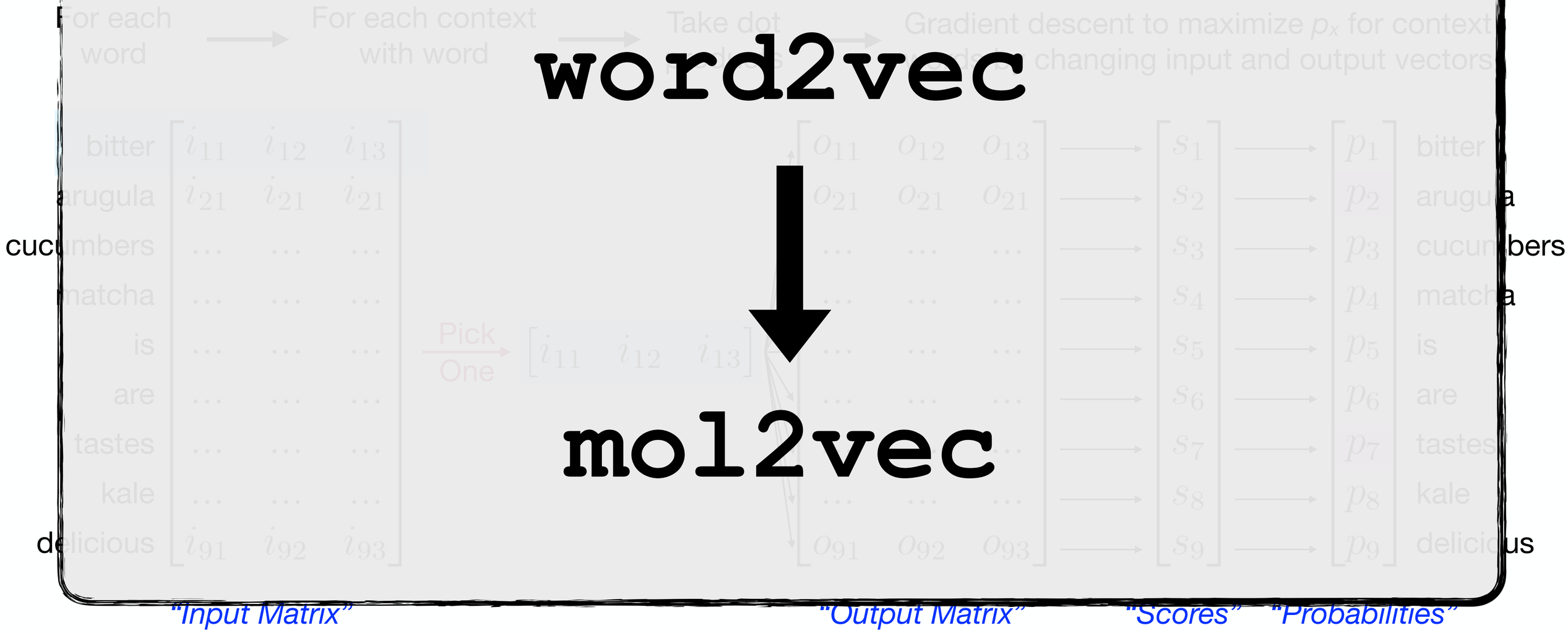
"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

WORD2VEC

“Kale is bitter.” “Arugula tastes bitter.” “Cucumbers are delicious.” “Matcha is delicious.”

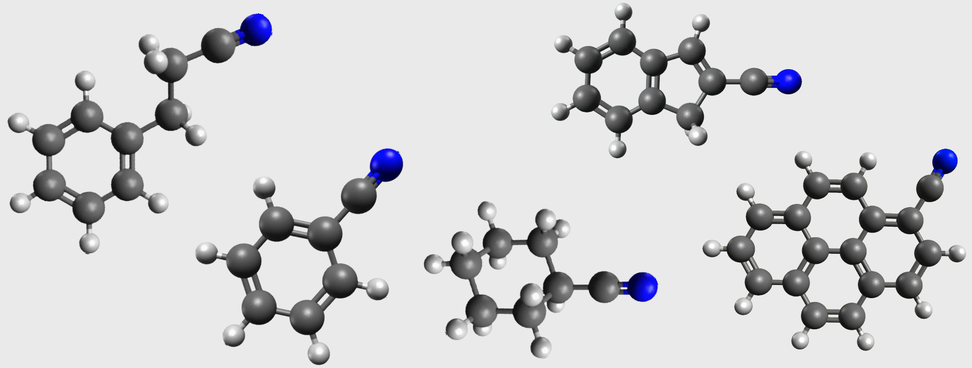
word2vec

mol2vec



WORD2VEC

Corpus of molecules (context sentences).
e.g. -CN tagged (P)AHs



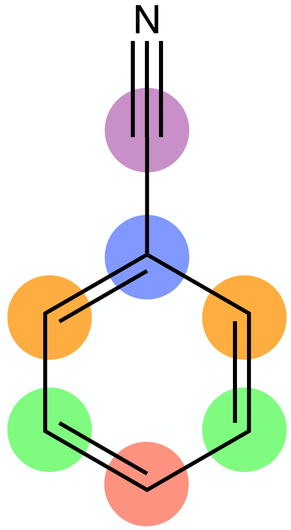
Goal is to create meaningful mathematical vector representations of “words” that carry information about their use and context

Similar “words” should be nearby one another in vector space

Question

If the molecule is a sentence, what are the words?

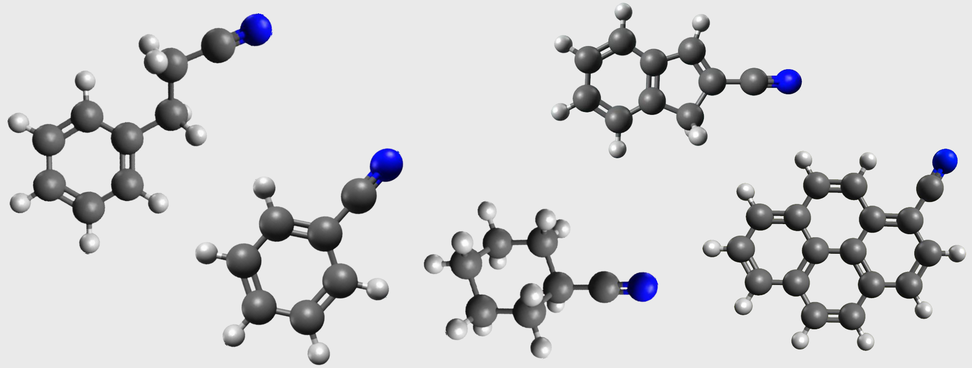
Context matters even when deciding on tokens (“words”) for molecules.



Benzonitrile
N#CC1=CC=CC=C1

WORD2VEC

Corpus of molecules (context sentences).
e.g. -CN tagged (P)AHs



Goal is to create meaningful mathematical vector representations of “words” that carry information about their use and context

Similar “words” should be nearby one another in vector space

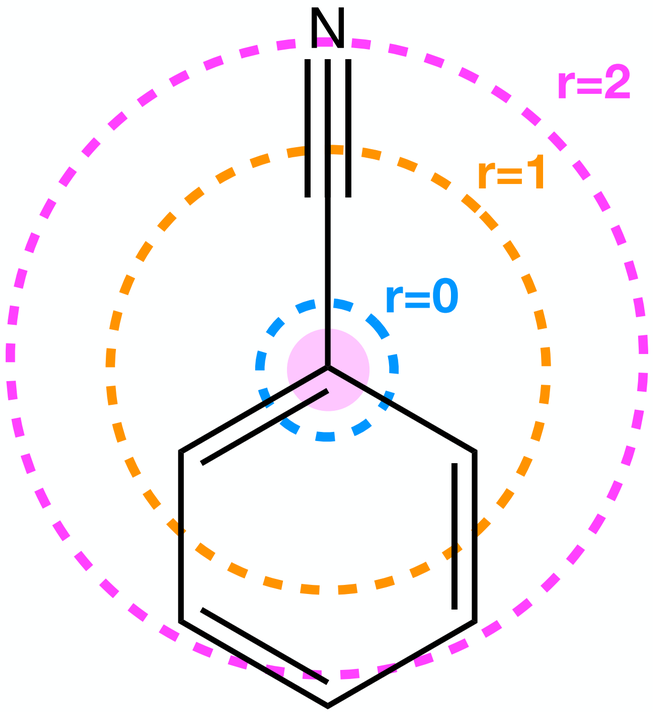
Aalto University is amazing.

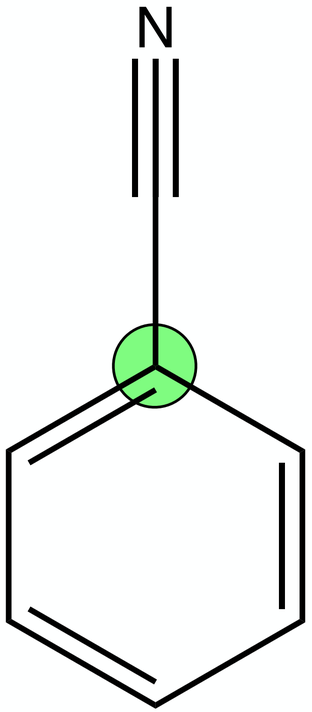
Aalto University is amazing.

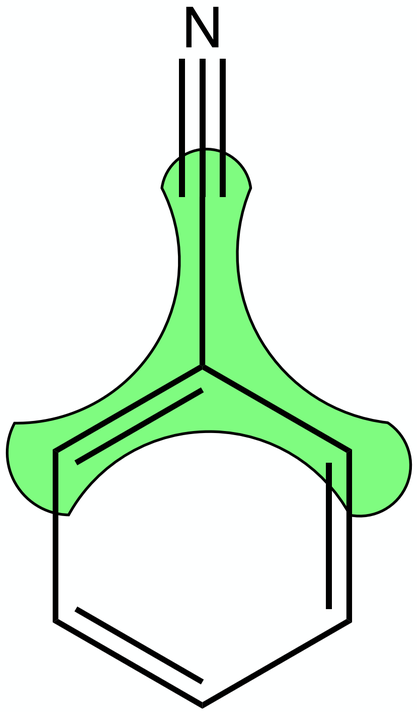
[Aalto, University, Aalto University, is, amazing]

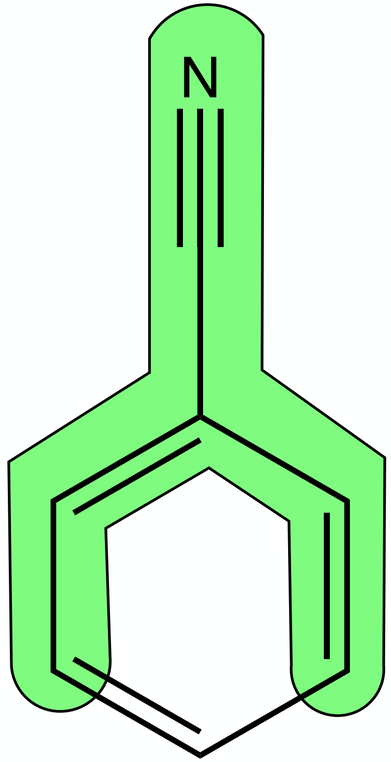
Context: Morgan Fingerprints

Morgan 1965 *J. Chem. Doc.* 5, 107
Jaeger et al. 2018 *J. Chem. Inf. Mod.* 58, 27



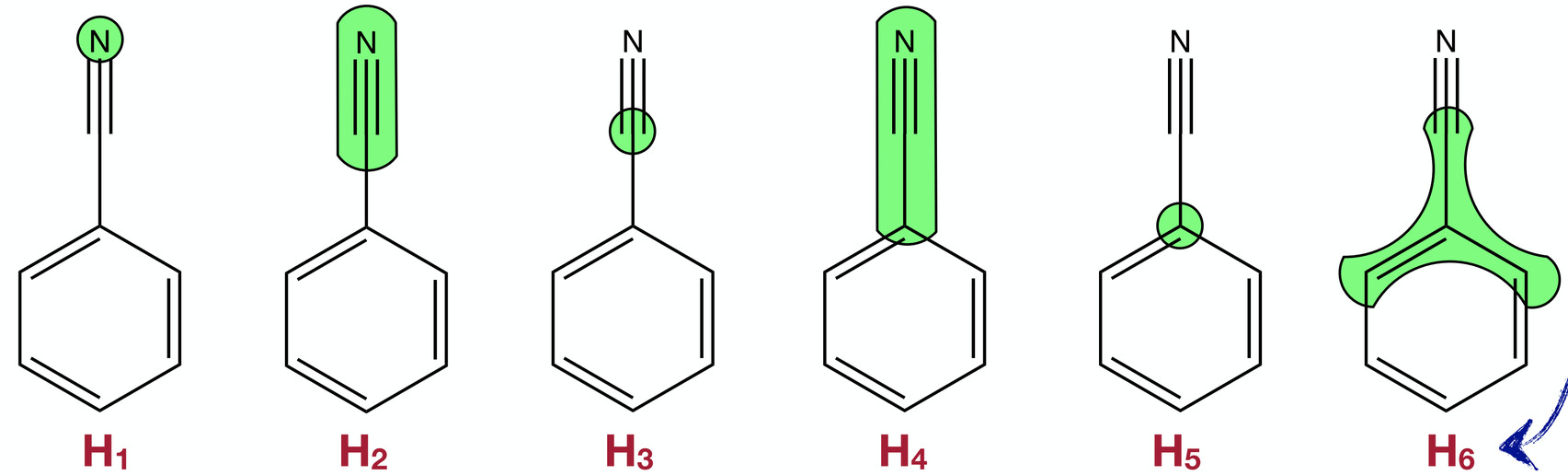






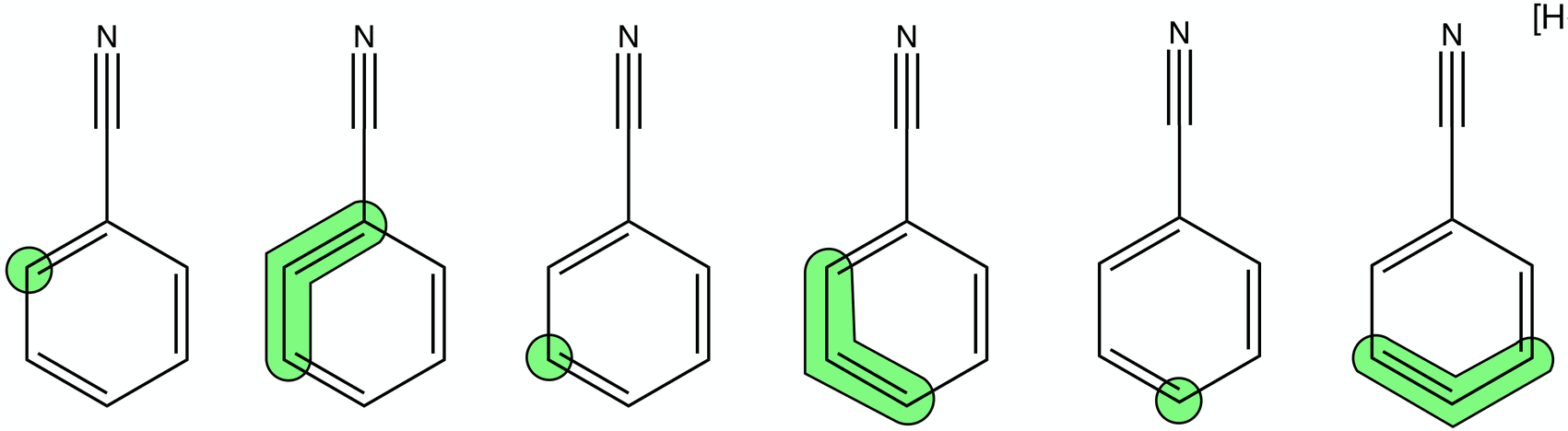
CONTEXT: MORGAN FINGERPRINTS

Morgan 1965 *J. Chem. Doc.* 5, 107
Jaeger et al. 2018 *J. Chem. Inf. Mod.* 58, 27



H_1 H_2 H_3 H_4 H_5 H_6

"Words" in the "sentence" that is benzonitrile



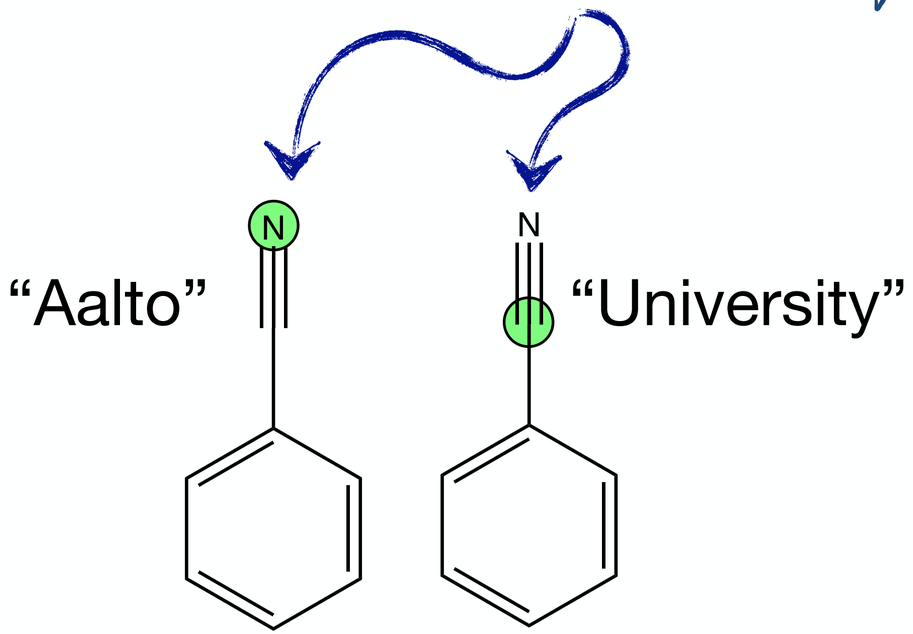
H_7 H_8 H_9 H_{10} H_{11} H_{12}

$[H_1, H_2, H_3, H_4, H_5, H_6, H_7, H_8, H_9, H_{10}, H_{11}, H_{12}]$

CONTEXT: MORGAN FINGERPRINTS

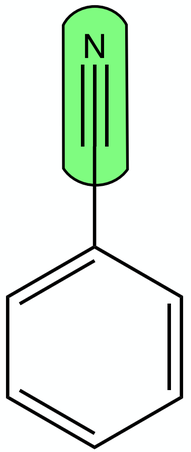
Morgan 1965 *J. Chem. Doc.* 5, 107
Jaeger et al. 2018 *J. Chem. Inf. Mod.* 58, 27

We want to have tokens for this N and C so we can encode their relationship to each other



"Aalto"

"University"



"Aalto University"

We also want a token for "C#N" so we can encode how the relationship of the -CN subgroup.

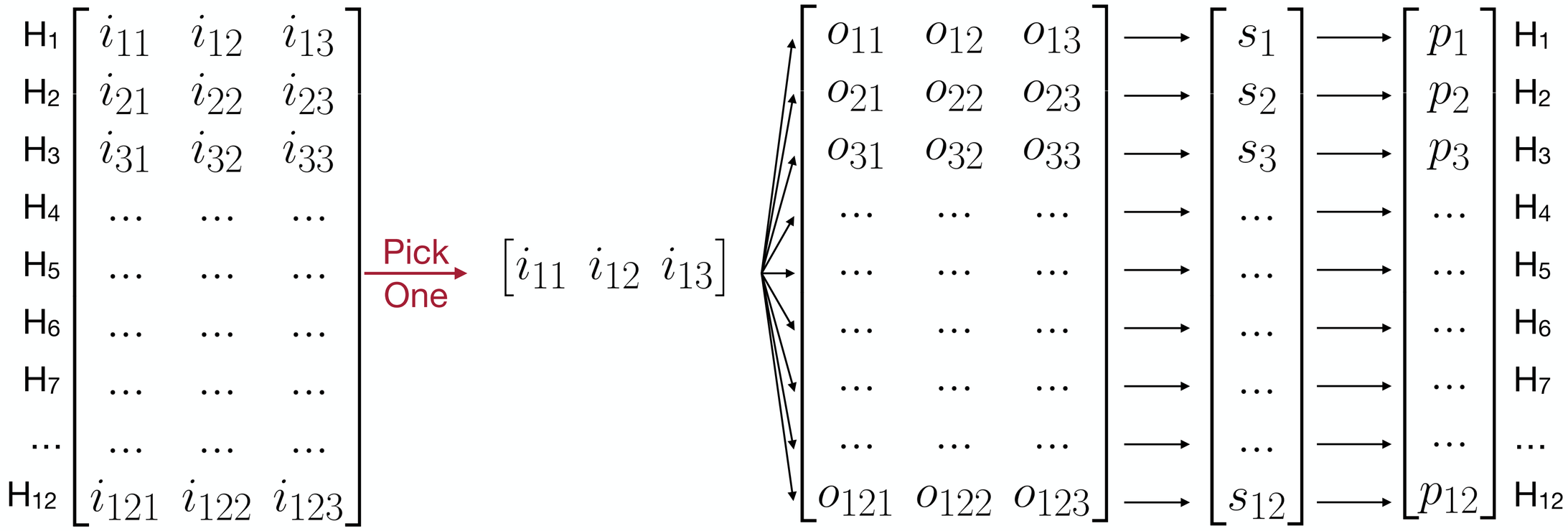
MOL2VEC

(word)
For each hash, go through all molecules in corpus
(sentences)

Benzonitrile

[H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12]

(word) (sentence)
If hash in molecule, train on context given a window



"Input Matrix"

"Output Matrix"

"Scores"

"Probabilities"

MOL2VEC

(word)
For each hash, go through all molecules in corpus
(sentences)

Benzonitrile

$[H_1, H_2, H_3, H_4, H_5, H_6, H_7, H_8, H_9, H_{10}, H_{11}, H_{12}]$

(word) (sentence)
If hash in molecule, train on context given a window

H_1
H_2
H_3
H_4
H_5
H_6
H_7
...
H_{12}

For simplicity, we're assuming that all molecules (sentences) in our corpus are made of the benzonitrile hashes (words), just ordered differently.

In reality, there will be thousands, just like there are thousands of words in a real word2vec corpus.

"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

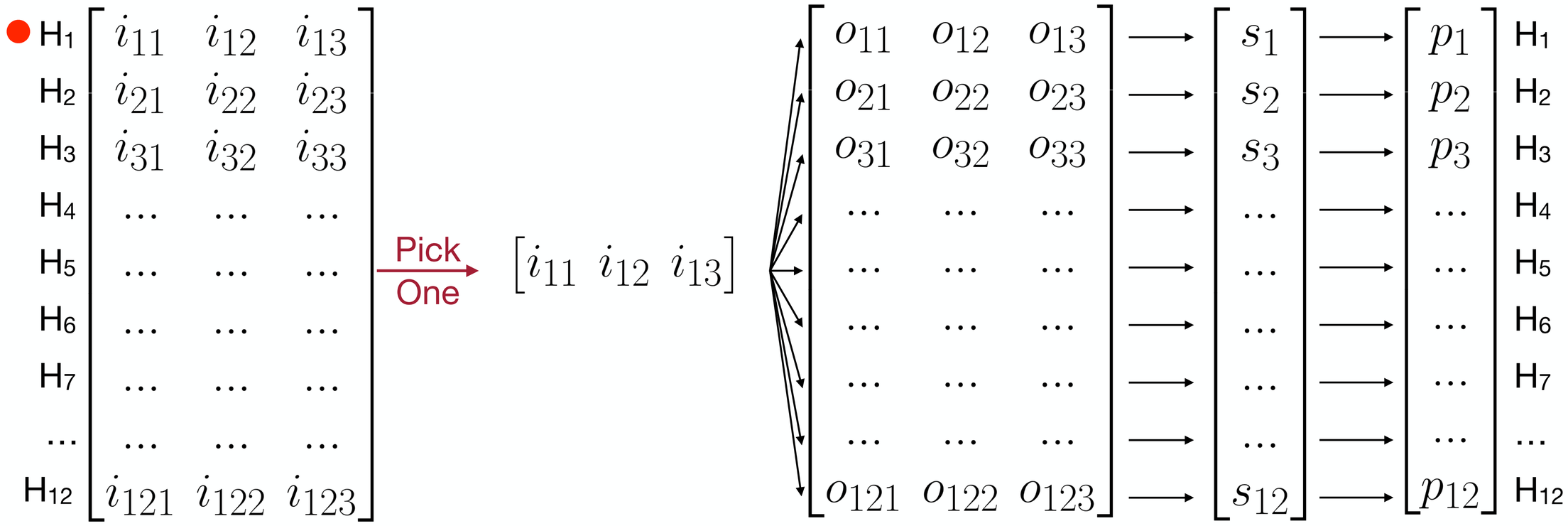
MOL2VEC

(word)
For each hash, go through all molecules in corpus
(sentences)

Benzonitrile

[H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12]

(word) (sentence)
If hash in molecule, train on context given a window

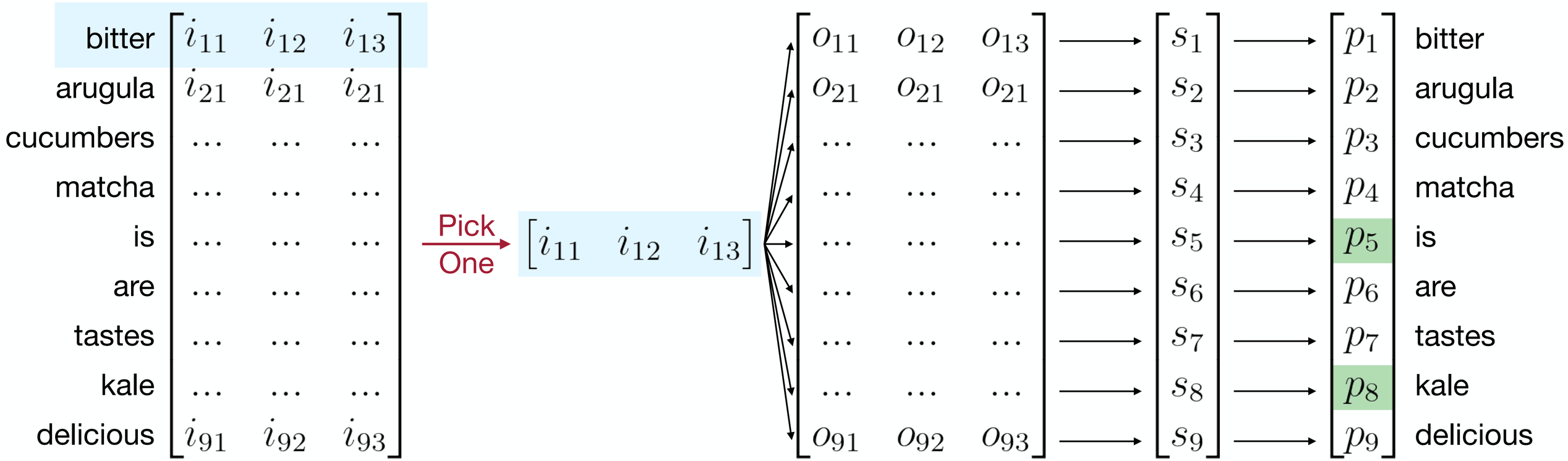


"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

WORD2VEC

"Kale is bitter." "Arugula tastes bitter." "Cucumbers are delicious." "Matcha is delicious."

For each word → For each context with word → Take dot products → Gradient descent to maximize p_x for context words by changing input and output vectors



"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

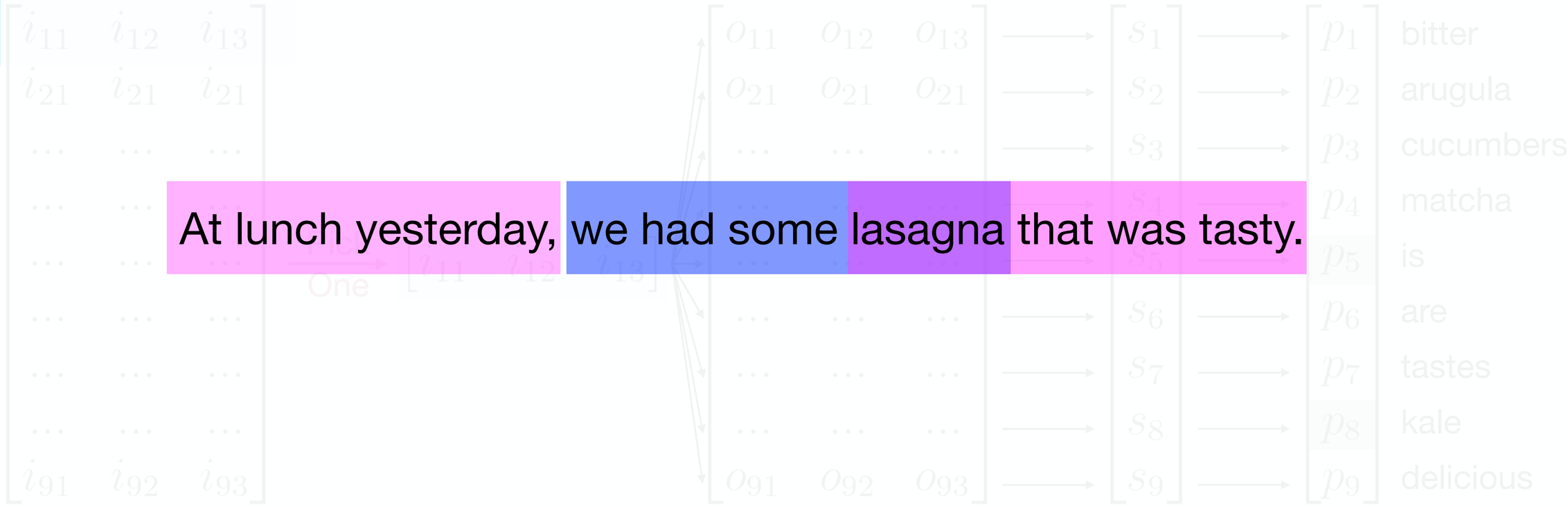
WORD2VEC

"Kale is bitter." "Arugula tastes bitter." "Cucumbers are delicious." "Matcha is delicious."

For each word → For each context with word → Take dot products → Gradient descent to maximize p_x for context words by changing input and output vectors

bitter
arugula
cucumbers
matcha
is
are
tastes
kale
delicious

At lunch yesterday, we had some lasagna that was tasty.



"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

MOL2VEC

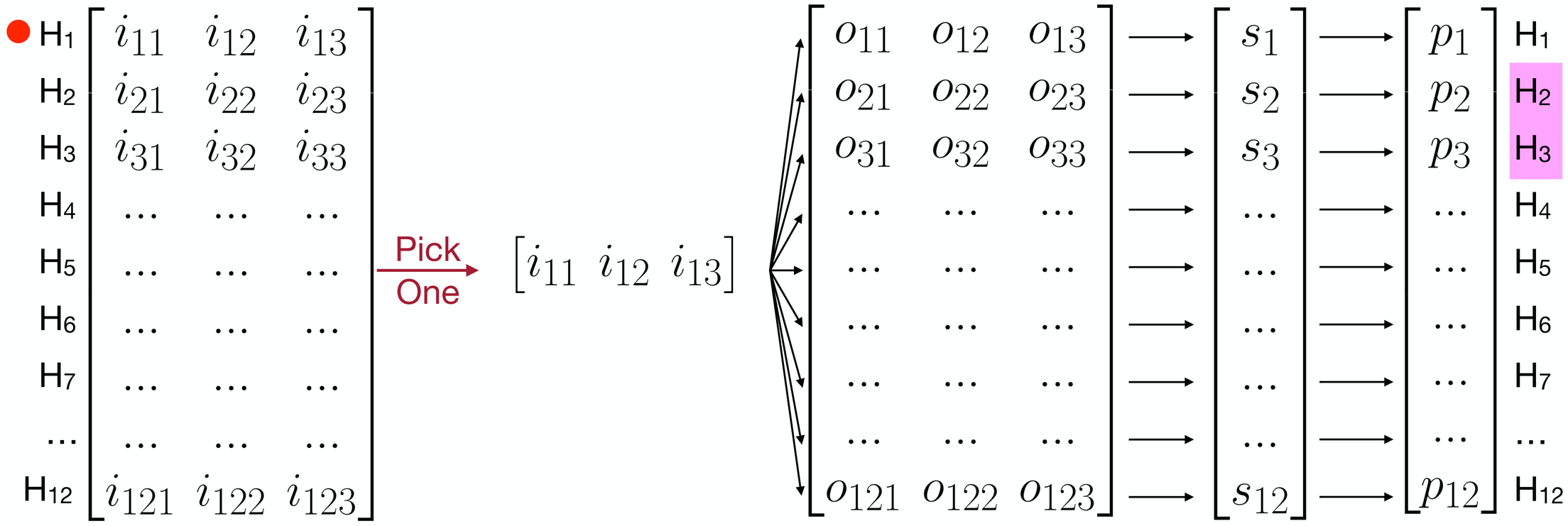
(word)
For each hash, go through all molecules in corpus
(sentences)

window = 2

Benzonitrile

[H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12]

(word) (sentence)
If hash in molecule, train on context given a window



"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

MOL2VEC

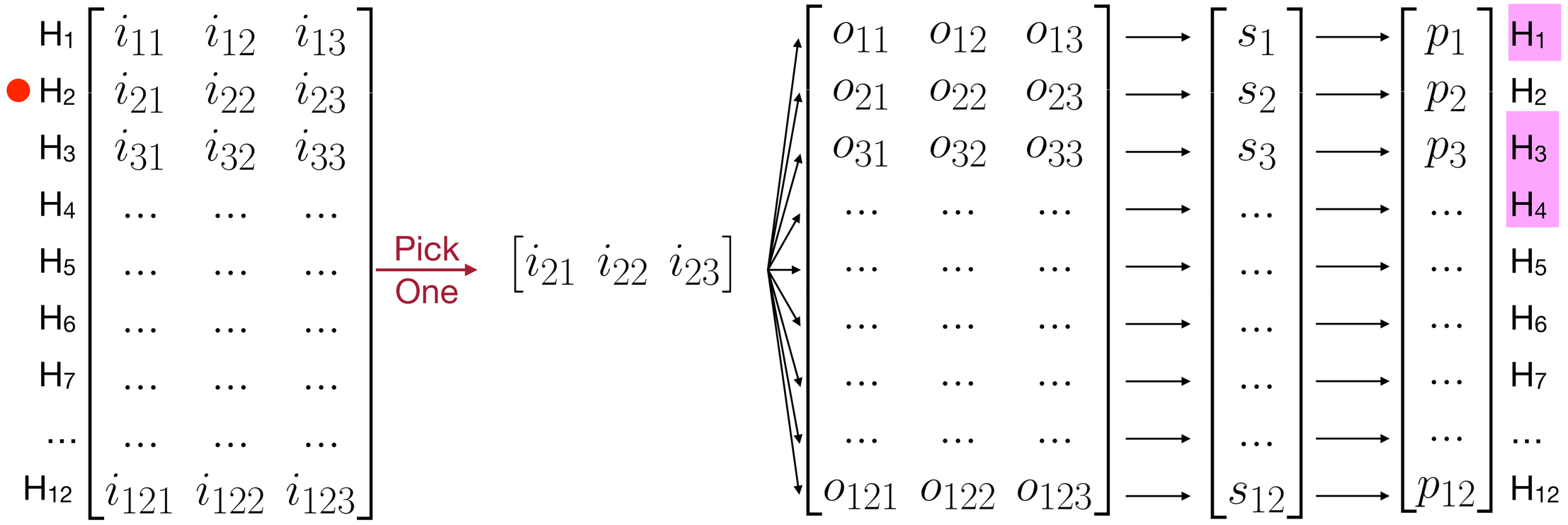
(word)
For each hash, go through all molecules in corpus
(sentences)

window = 2

Benzonitrile

[H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12]

(word) (sentence)
If hash in molecule, train on context given a window



"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

MOL2VEC

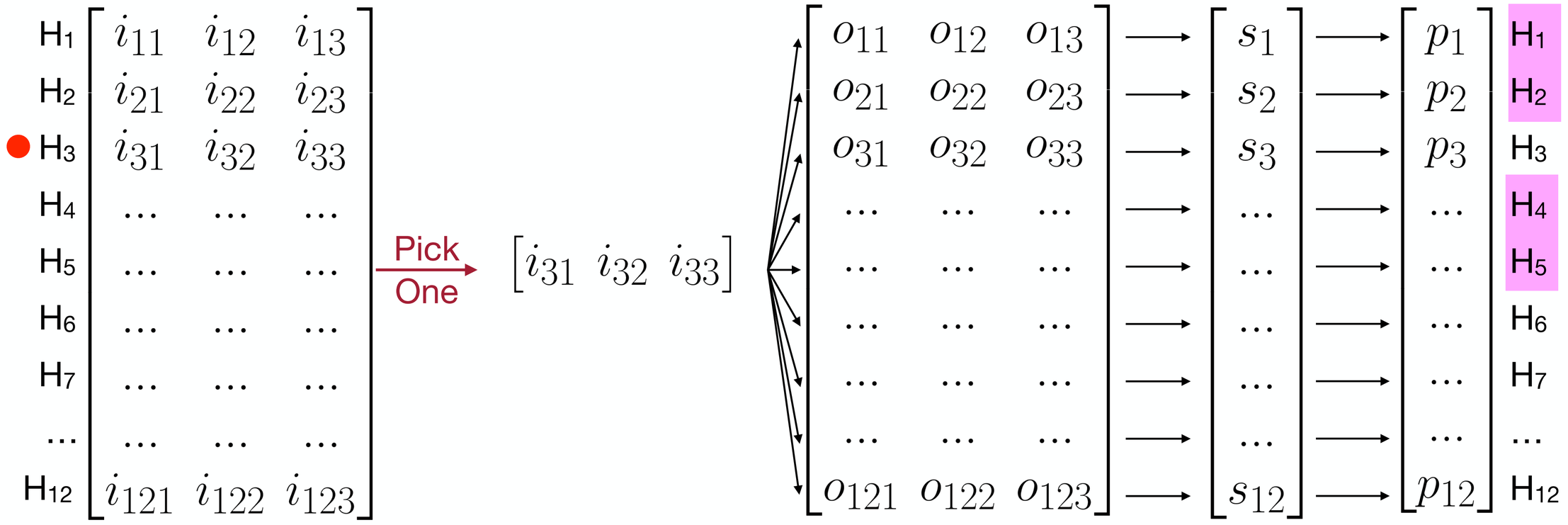
(word)
For each hash, go through all molecules in corpus
(sentences)

window = 2

Benzonitrile

[H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12]

(word) (sentence)
If hash in molecule, train on context given a window



"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

MOL2VEC

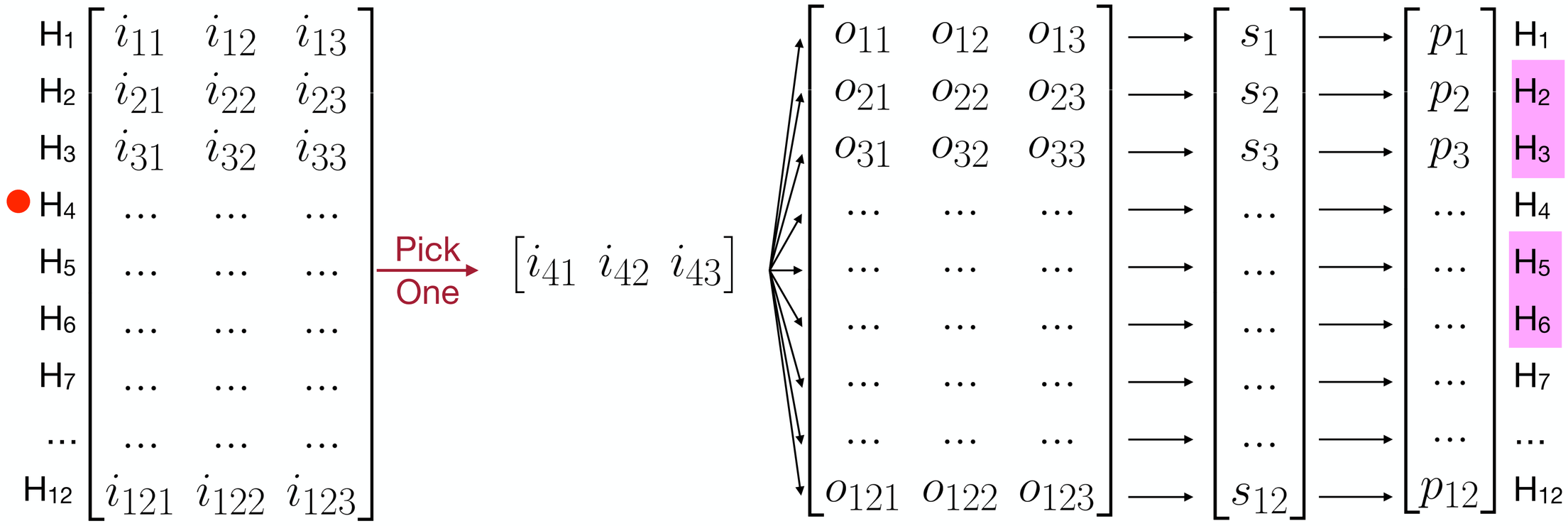
(word)
For each hash, go through all molecules in corpus
(sentences)

window = 2

Benzonitrile

[H1, H2, H3, H4, H5, H6, H7, H8, H9, H10, H11, H12]

(word) (sentence)
If hash in molecule, train on context given a window



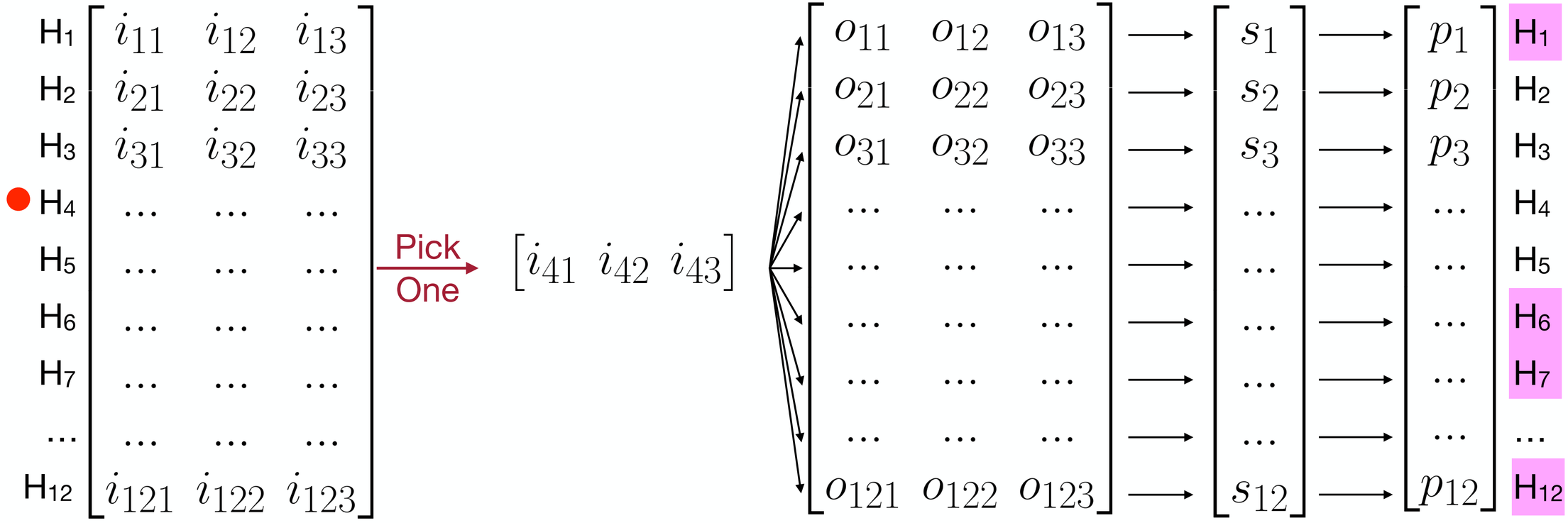
"Input Matrix" *"Output Matrix"* *"Scores"* *"Probabilities"*

MOL2VEC

Random Corpus Molecule #834,892

window = 2

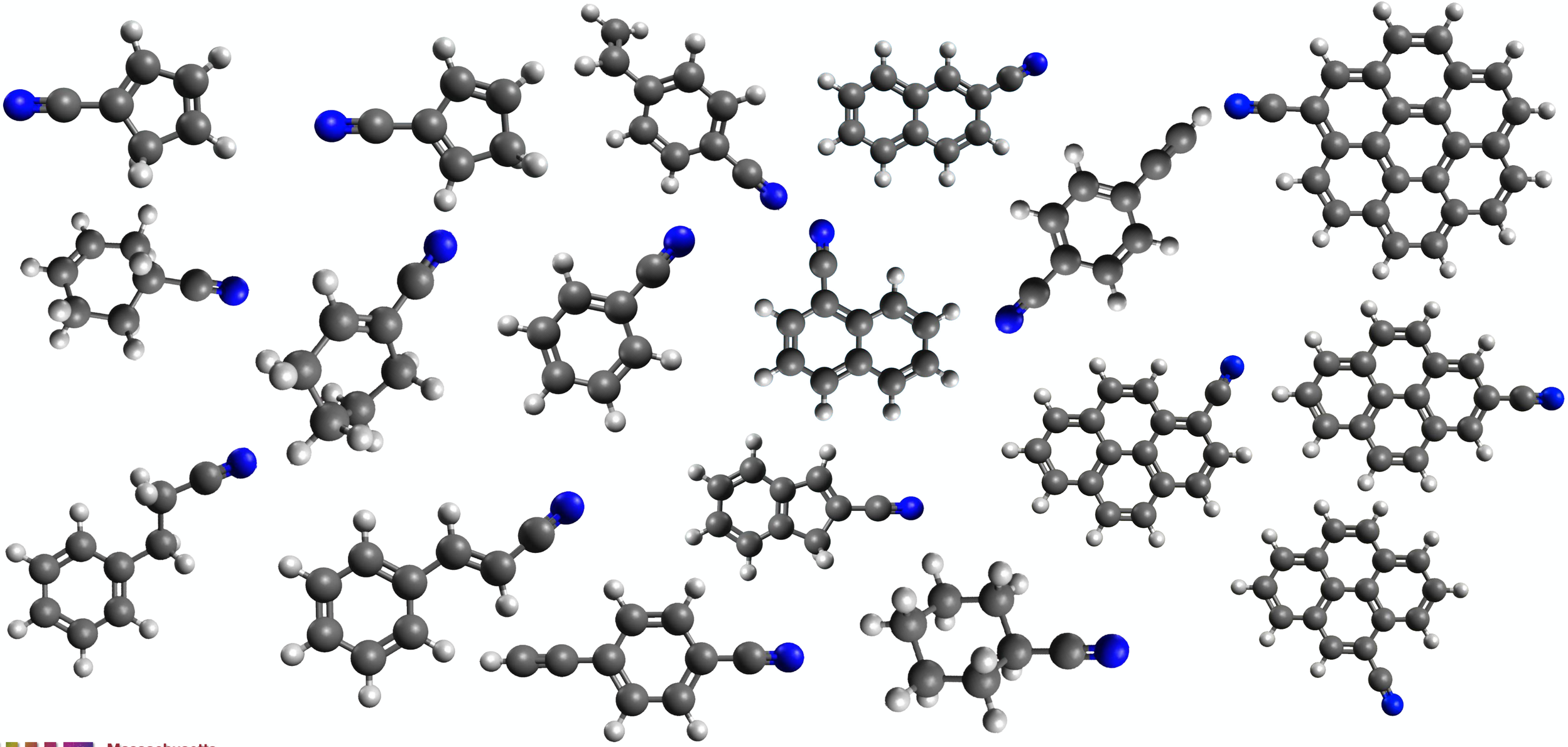
$[H_1, H_6, H_{12}, H_4, H_7, H_1, H_9, H_2, H_{11}, H_4, H_2, H_{12}]$



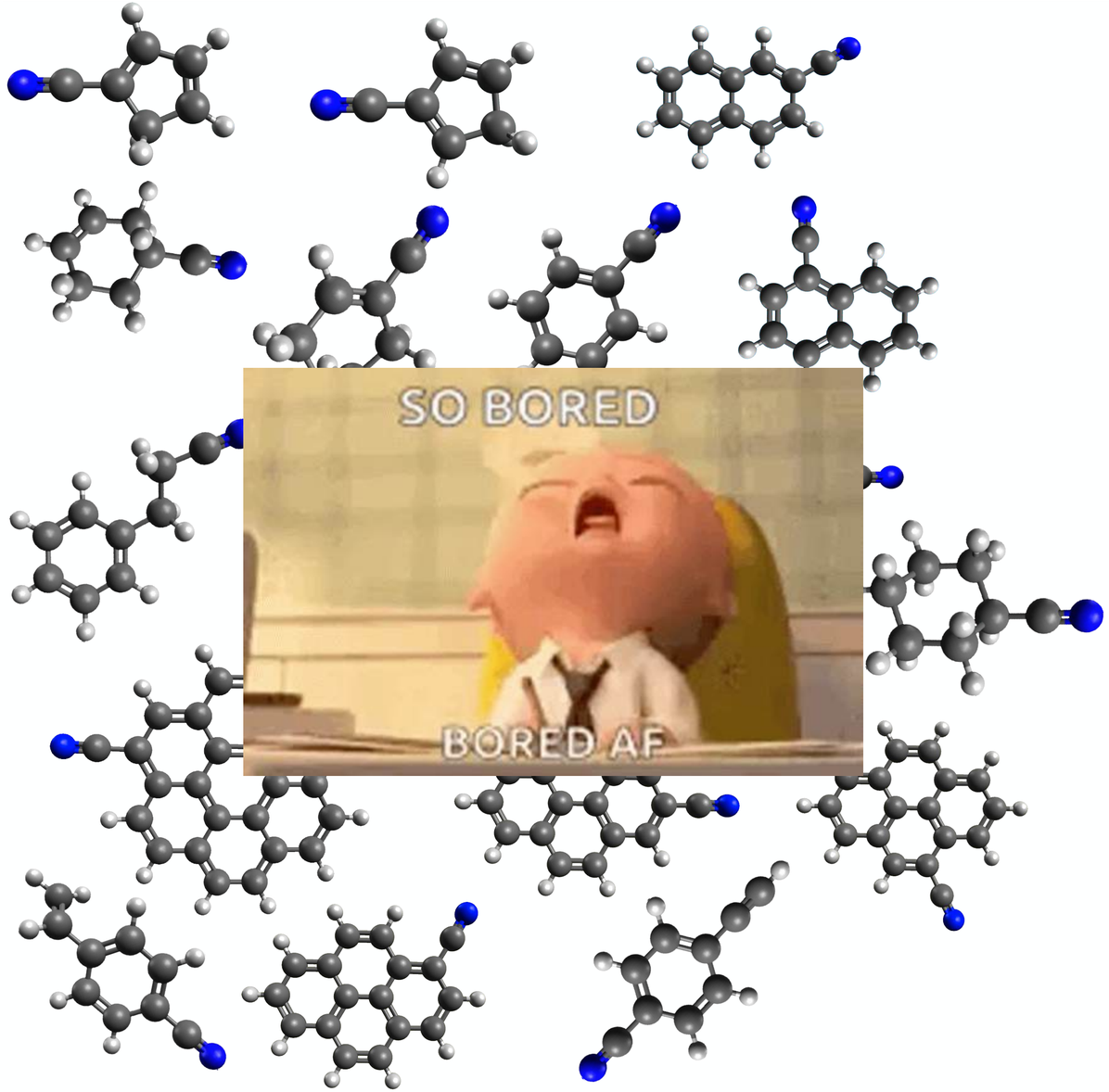
Corpus Composition

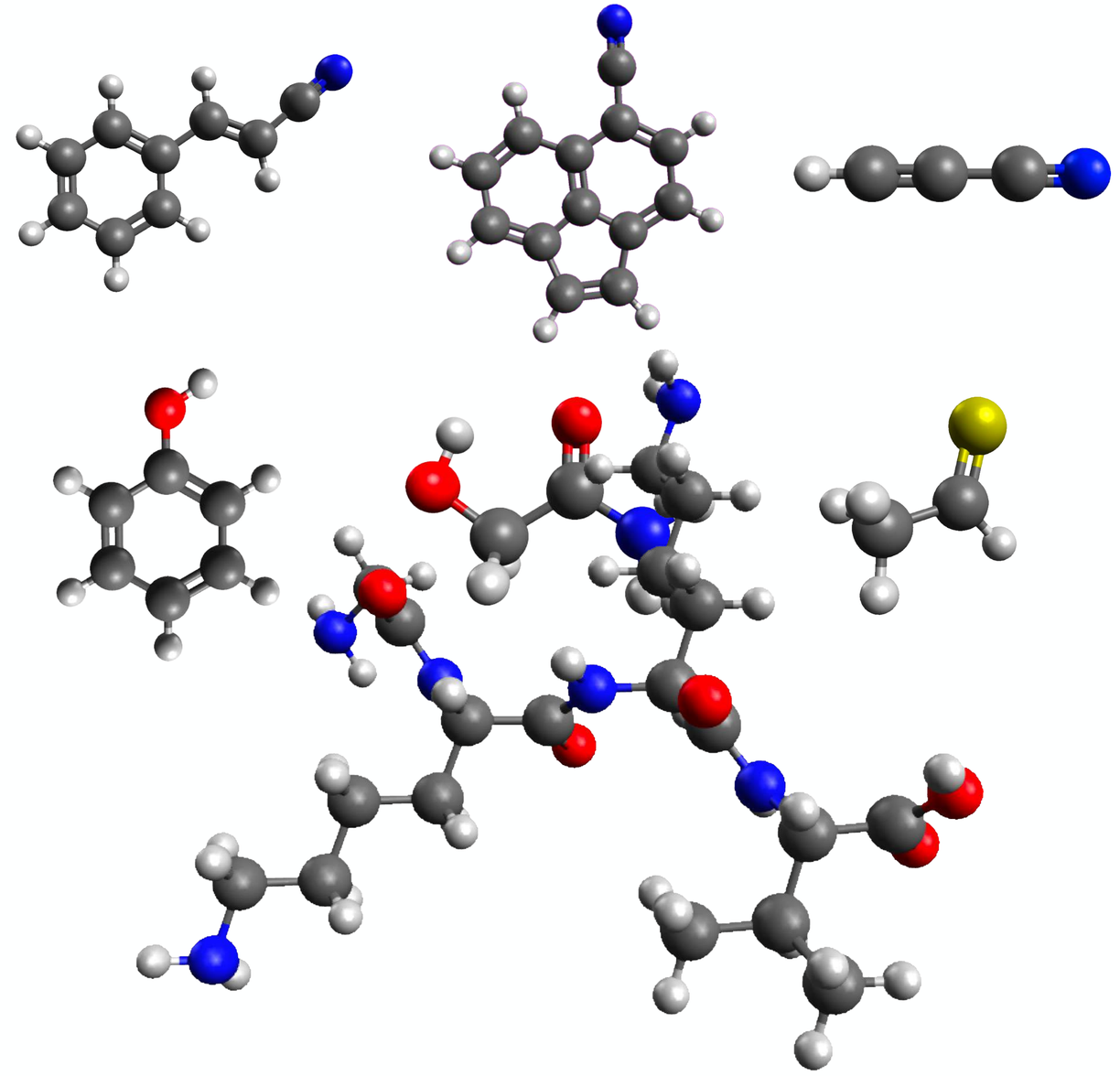
The corpus of molecules you train on will determine what chemical properties and relationships are reflected in your embeddings

Corpus Composition



Corpus Composition






TMC-1
Kelvin Lee
Machine Learning of Chemical Inventories
Lee et al. 2021 ApJL 917, L6

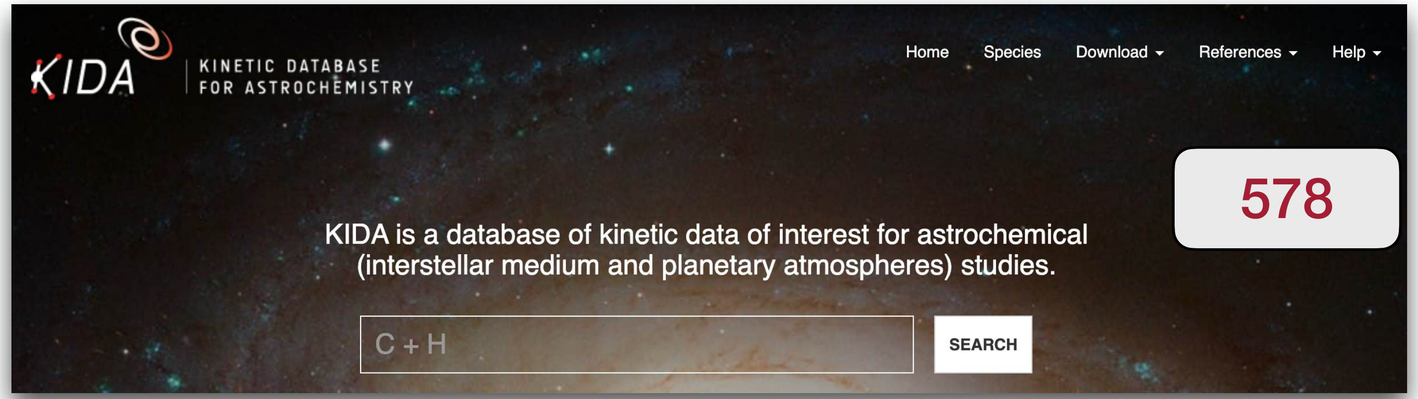
CORPUS COMPOSITION

Lee et al. 2021 ApJL 917, L6

A NEW REFERENCE CHEMICAL COMPOSITION FOR TMC-1

P. GRATIER[1], L. MAJUMDAR[1,2], M. OHISHI[3,4], E. ROUEFF[5,6], J. C. LOISON[7], K. M. HICKSON[7], AND V. WAKELAM[1]

87



578



3139

Subsets of small/medium molecules from …

Quantum chemistry structures and properties of 134 kilo molecules

Raghunathan Ramakrishnan[1], Pavlo O. Dral[2,3], Matthias Rupp[1] & O. Anatole von Lilienfeld[3,4]

QM9

134k

PubChem's BioAssay Database

Yanli Wang ✉, Jewen Xiao, Tugba O. Suzek, Jian Zhang, Jiyao Wang, Zhigang Zhou, Lianyi Han, Karen Karapetyan, Svetlana Dracheva, Benjamin A. Shoemaker ... Show more

Nucleic Acids Research, Volume 40, Issue D1, 1 January 2012, Pages D400–D412,

PCBA

438k

PubChem in 2021: new data content and improved web interfaces

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu ... Show more

Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D1388–D1395,

PubChem

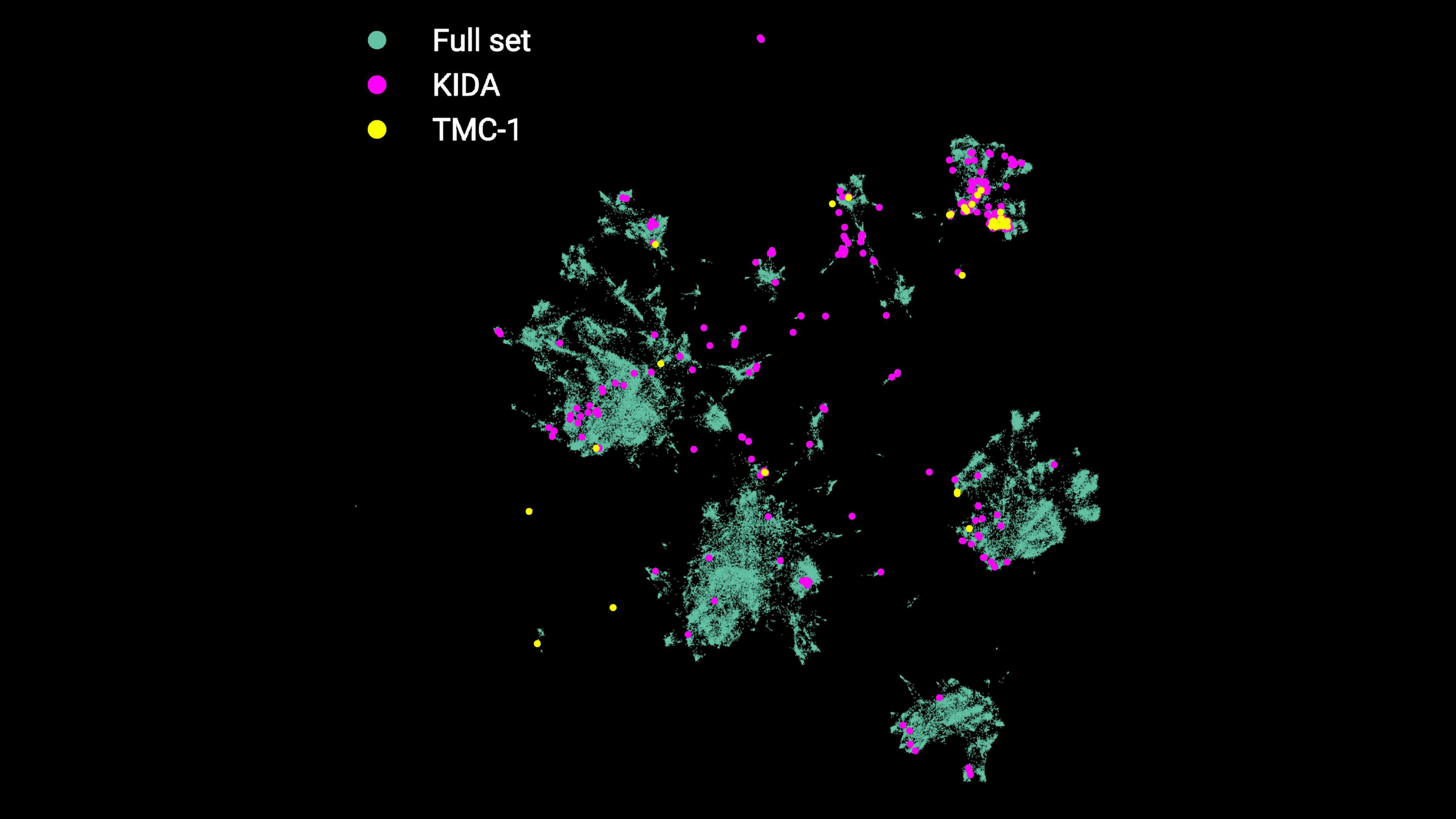
2.4M

ZINC 15 – Ligand Discovery for Everyone

Teague Sterling, and John J. Irwin*

ZINC

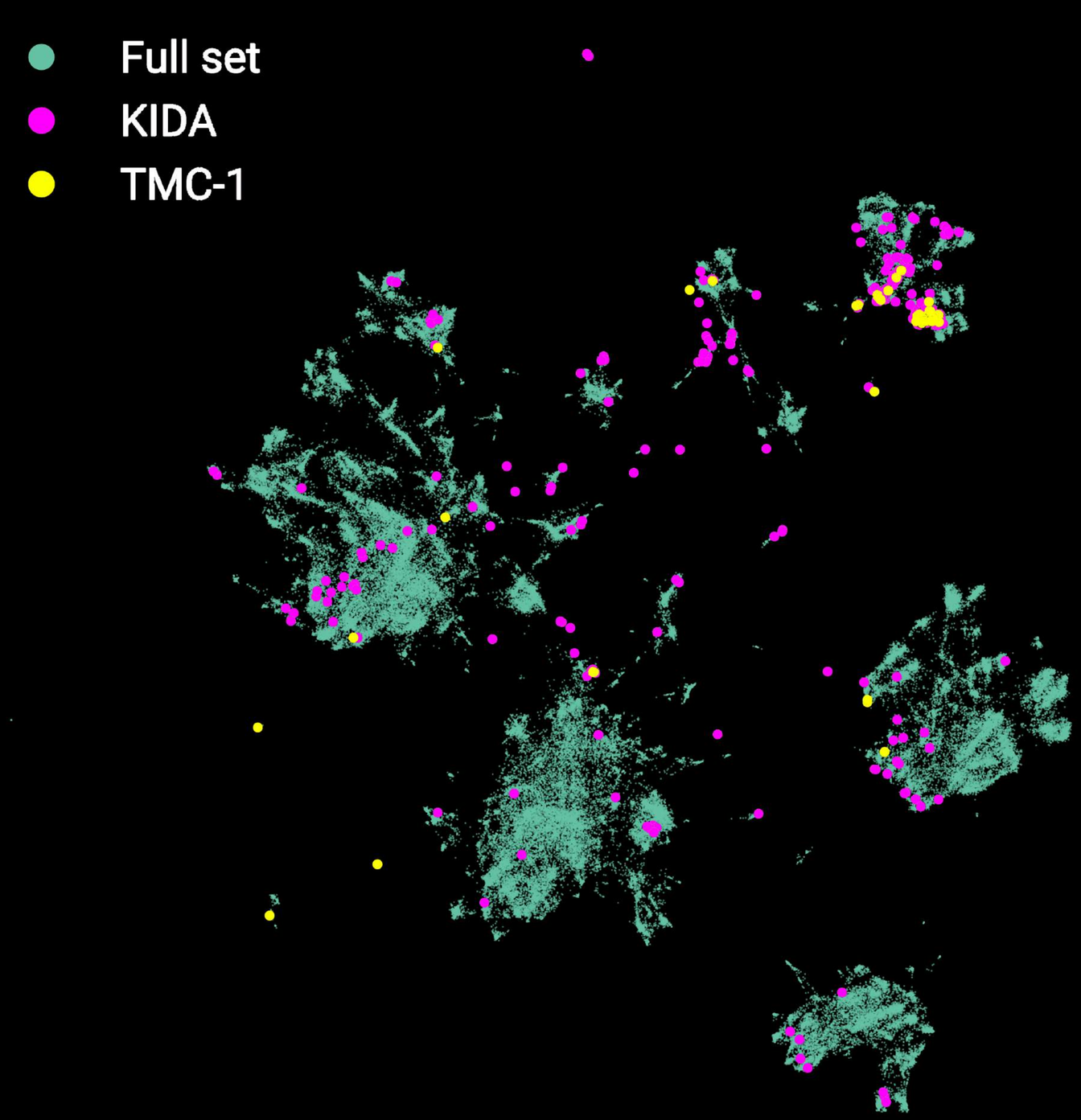
3.8M


Full set
KIDA
TMC-1

Acquired: Machine Chemical Intuition?

Can we learn a relationship between chemical abundance?

If so, what do we actually learn?



Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6

Given

Molecular Vectors
(Coordinates in Chemical Space)

+

Measured Column Densities

We Want

A mathematical relationship that, when given a vector, returns the correct column density

Possible Regressors

Parametric

Pre-defined functional form and parameters for the dataset are learned

Example
Linear Regression

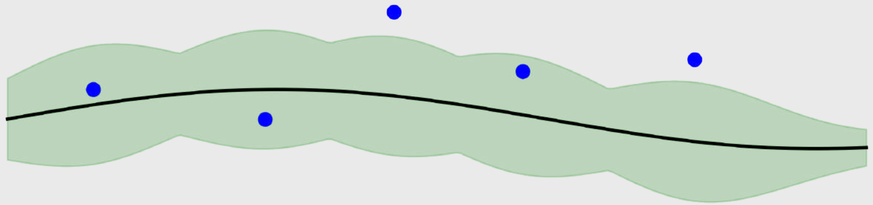
Pre-defined function: $y(x) = \boldsymbol{m}x + \boldsymbol{b}$

m and ***b*** are learned

Non-Parametric

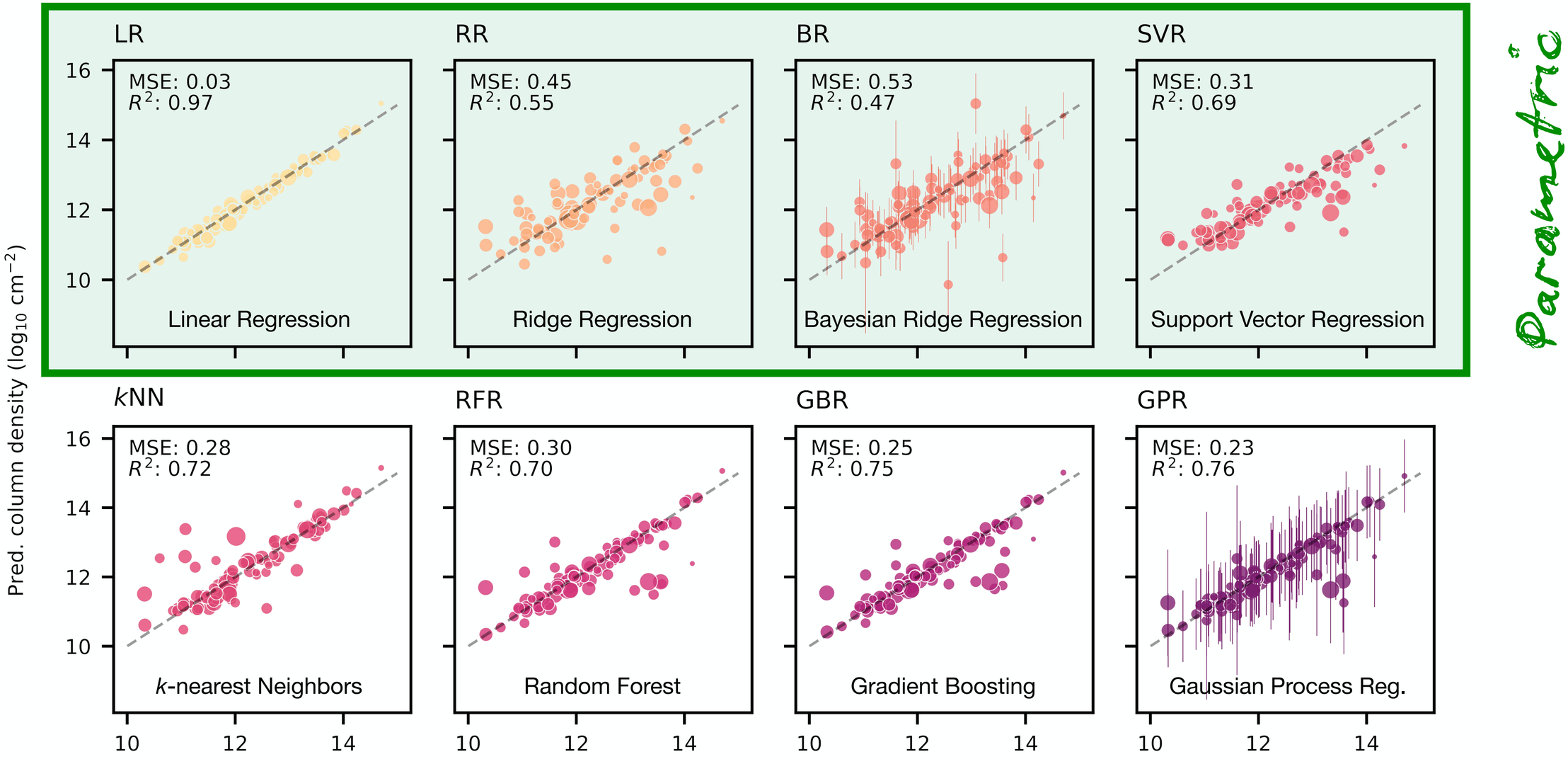
A model with a unique functional form specifically for the dataset is trained

Example
Gaussian Process Regression



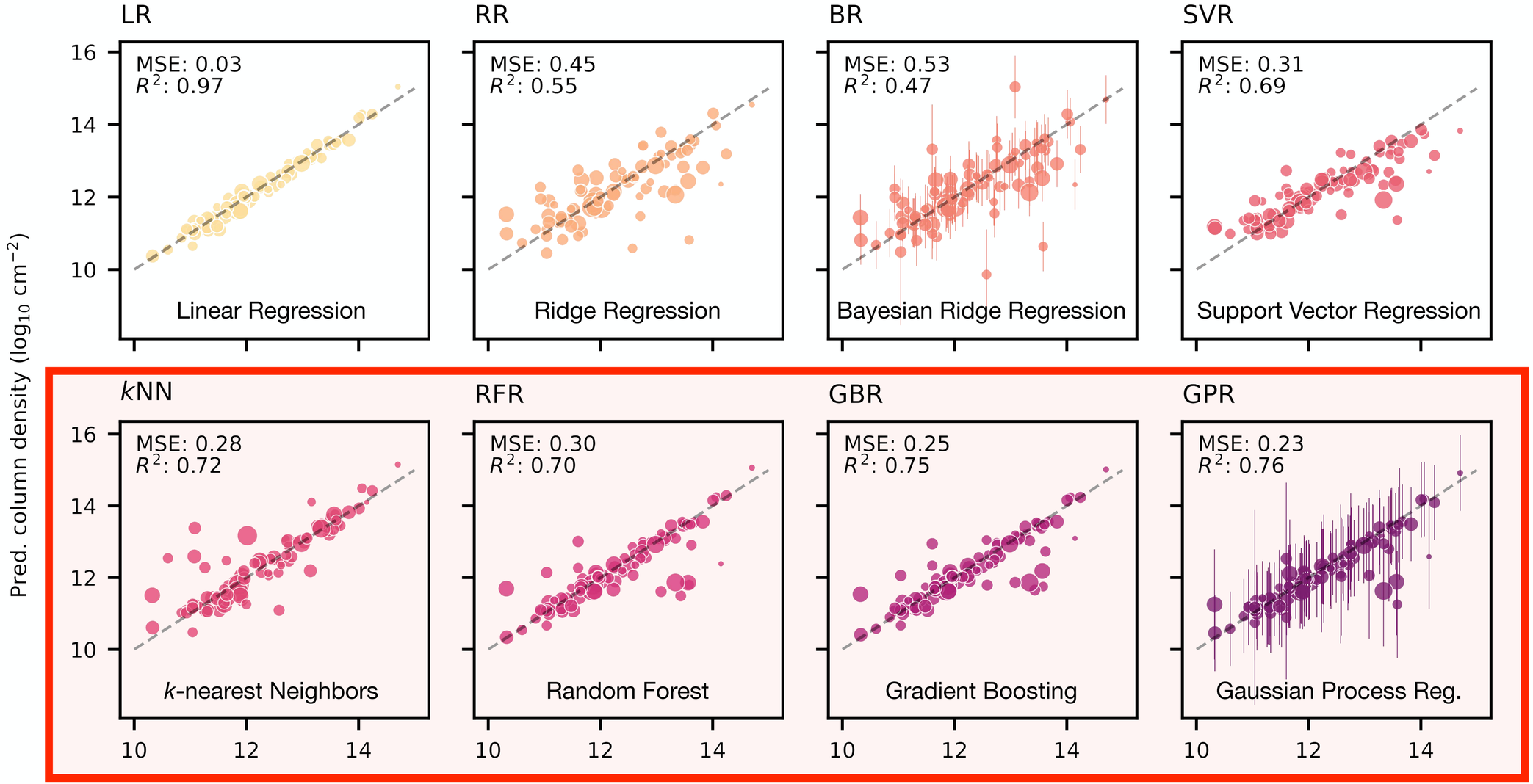
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



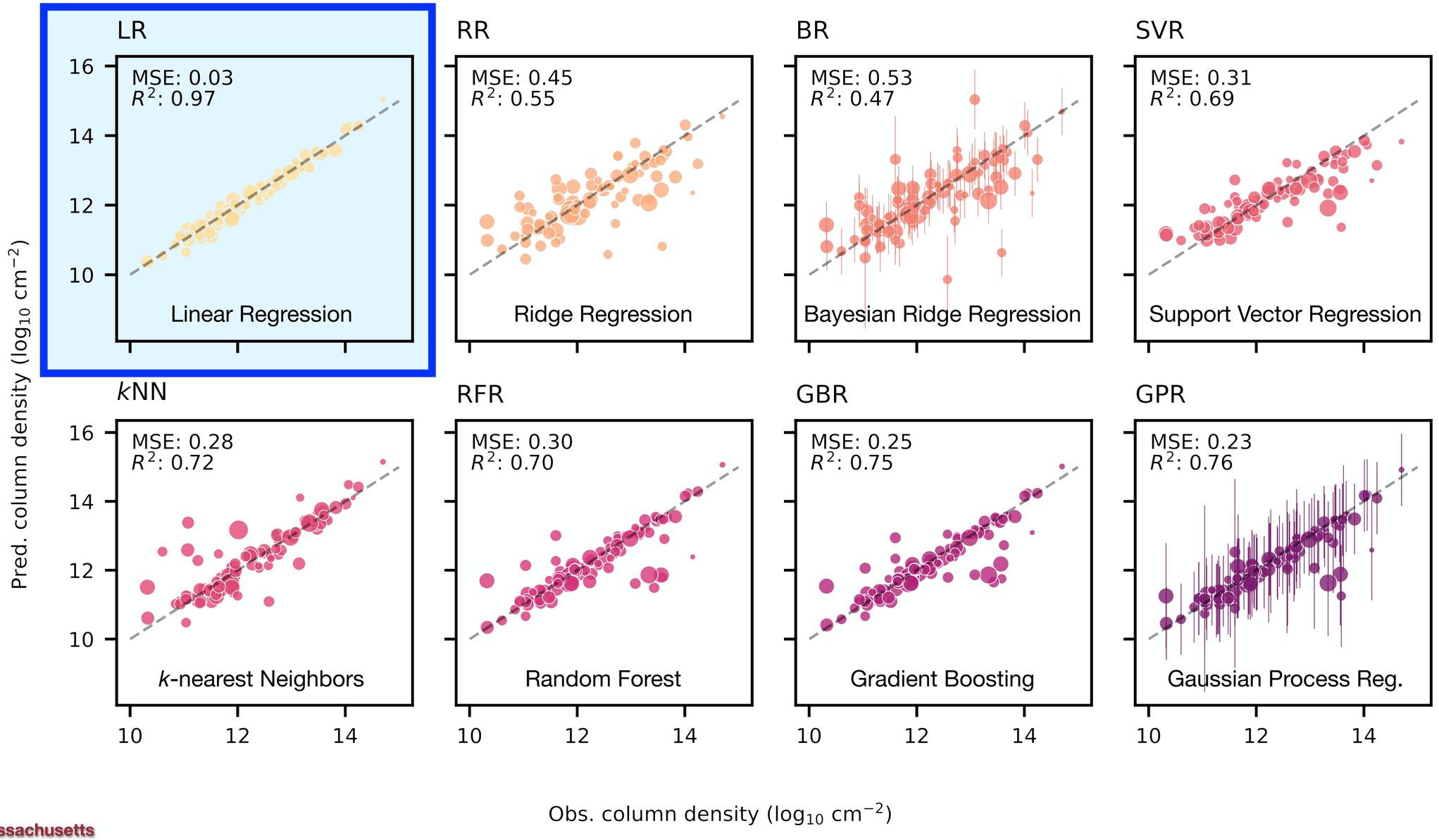
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



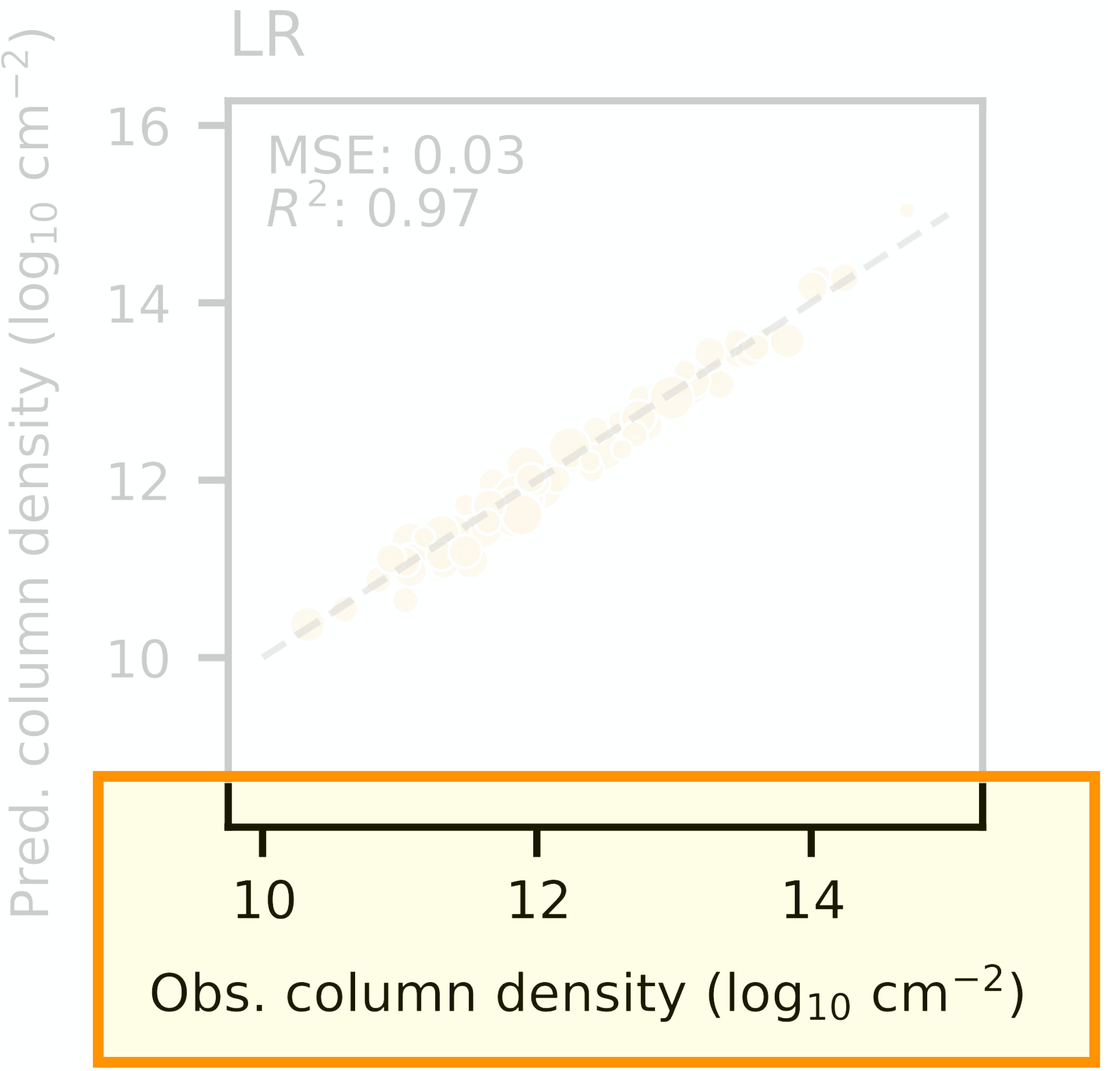
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6

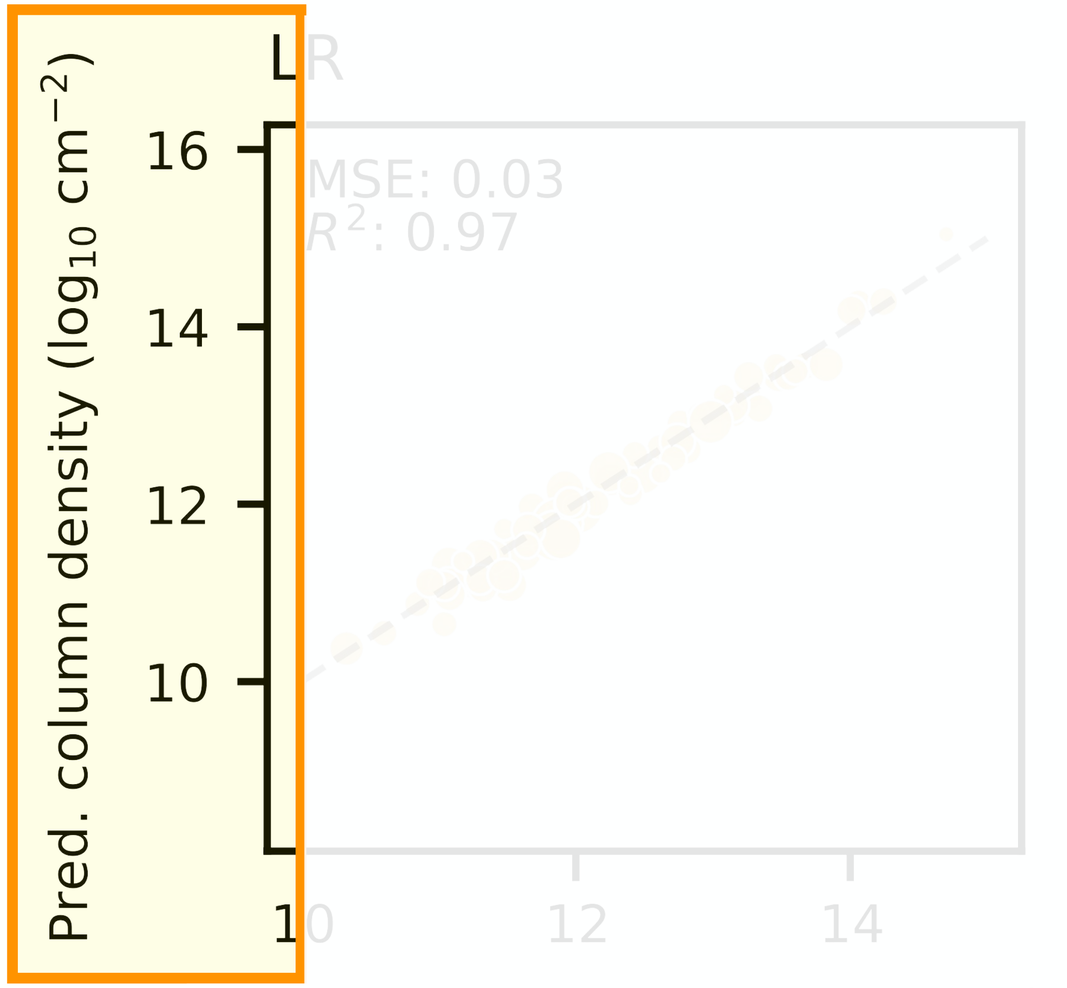


Actual values derived from observations

Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6

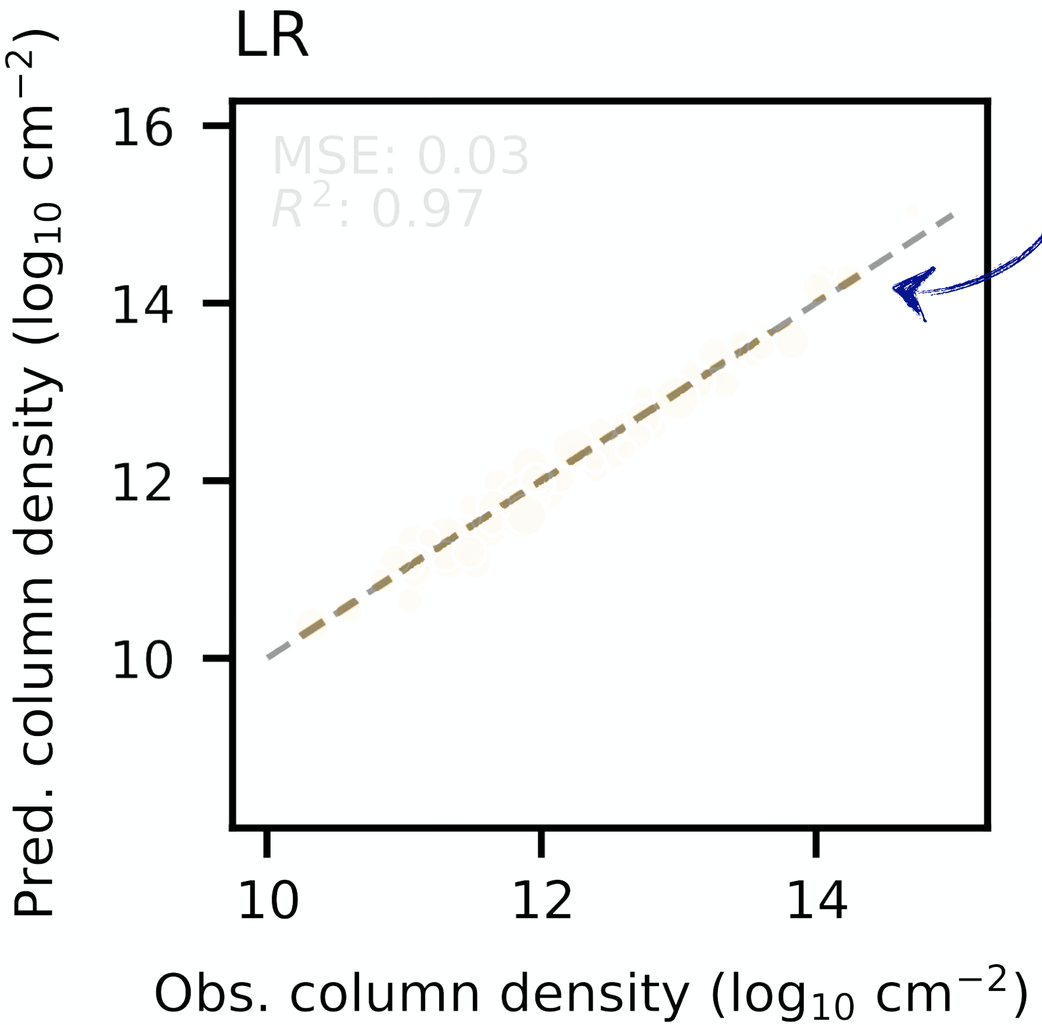
Values returned by model



Actual values derived from observations

Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



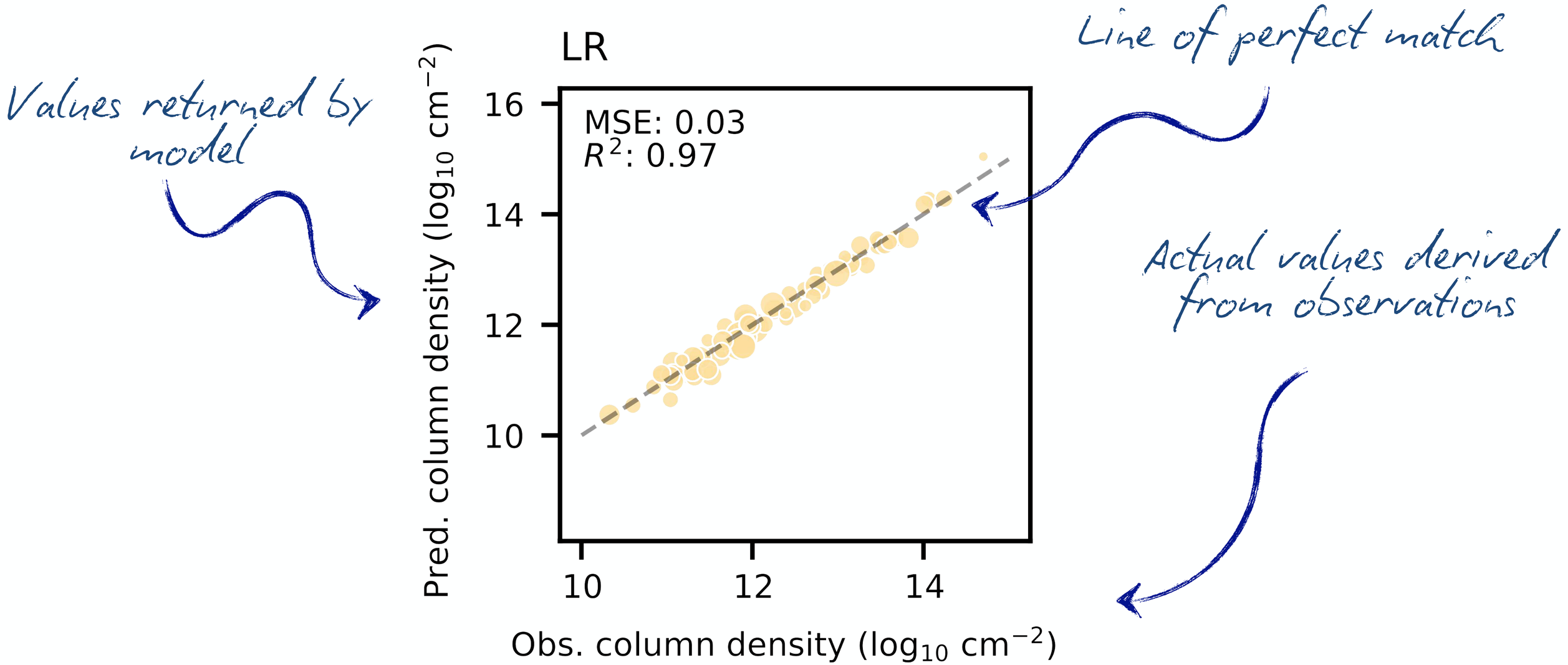
Line of perfect match

Values returned by model

Actual values derived from observations

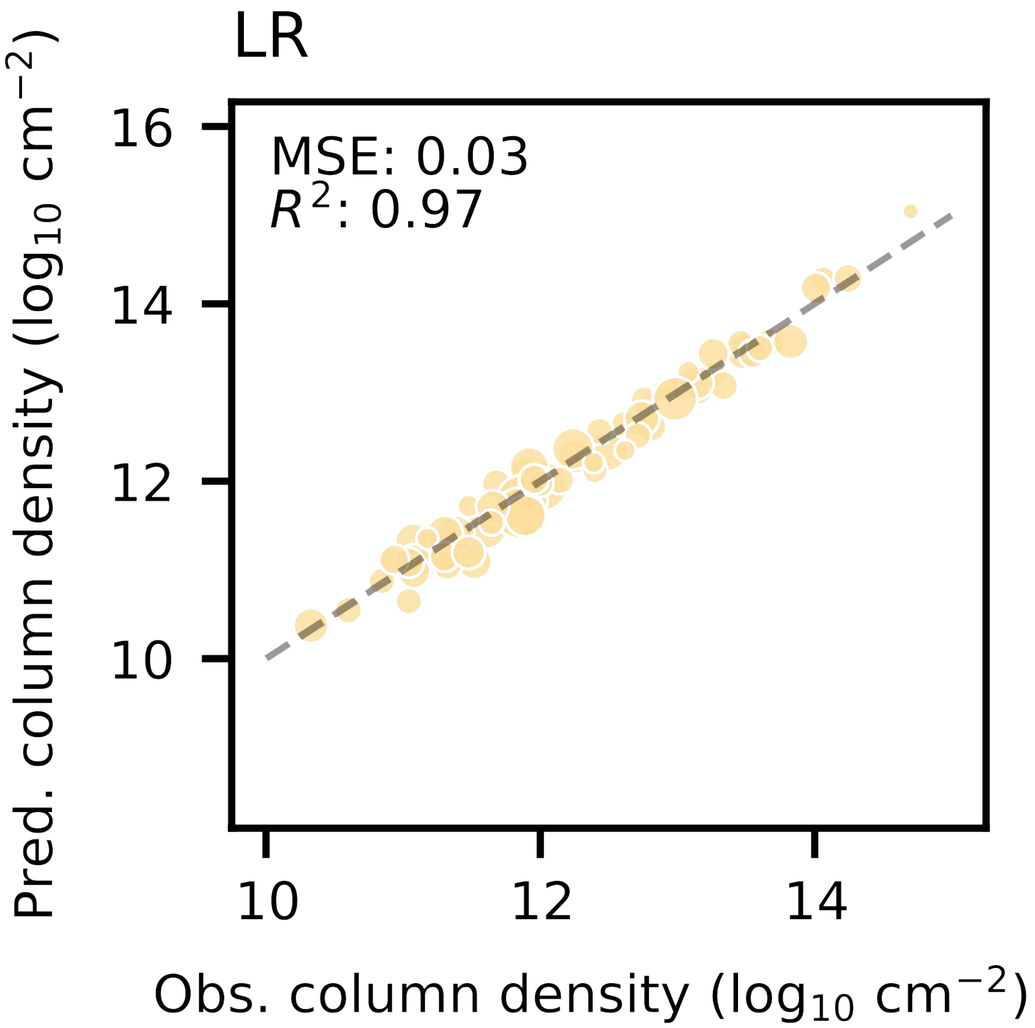
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



Overfit. Why?

mol2vec → 300 dimension vectors

PCA → 70 dimension vectors

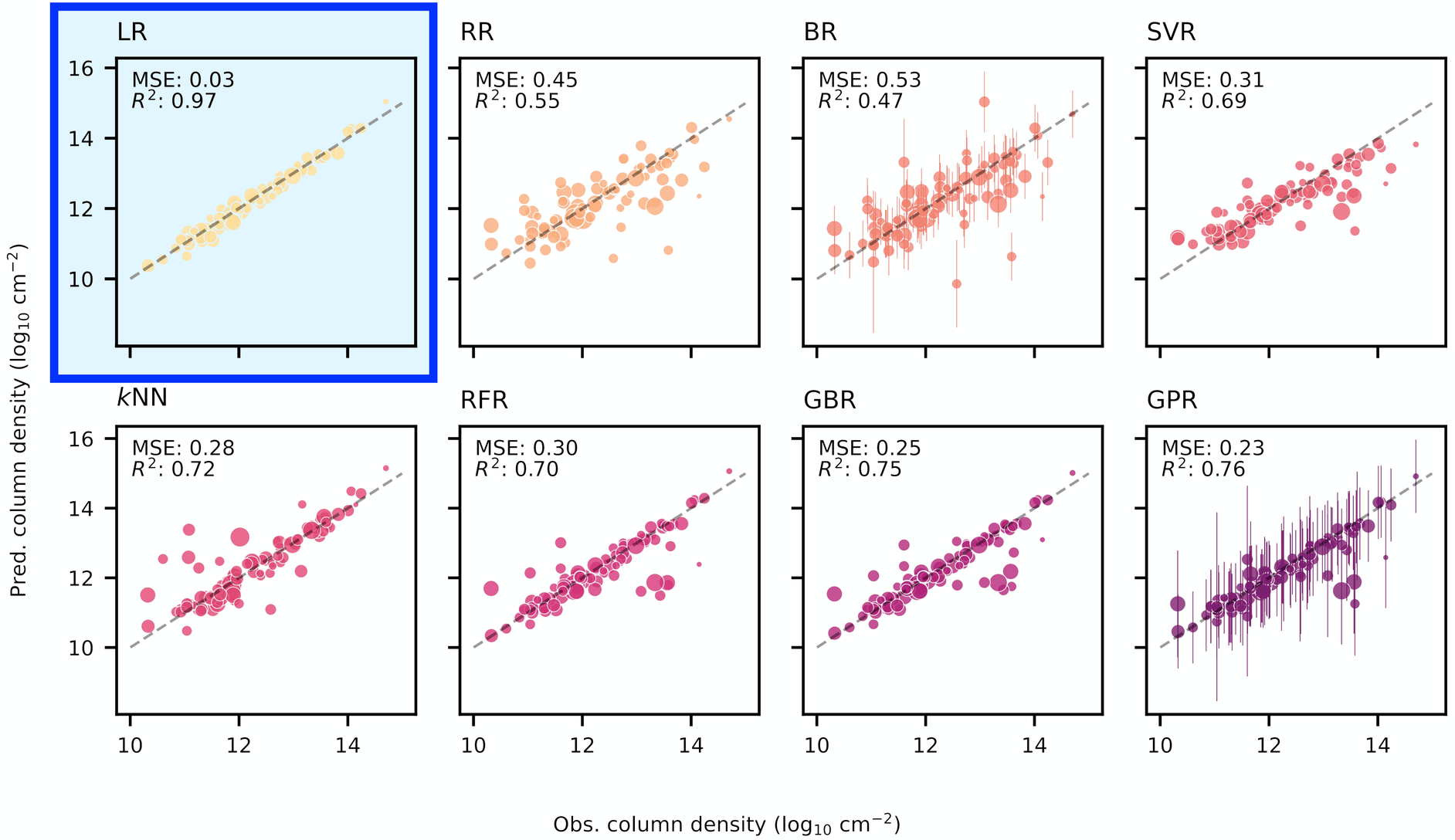
87 data points

Fitting 87* data points with 70 degrees of freedom

**Bootstrapping can help a bit with this, but not eliminate the issue entirely*

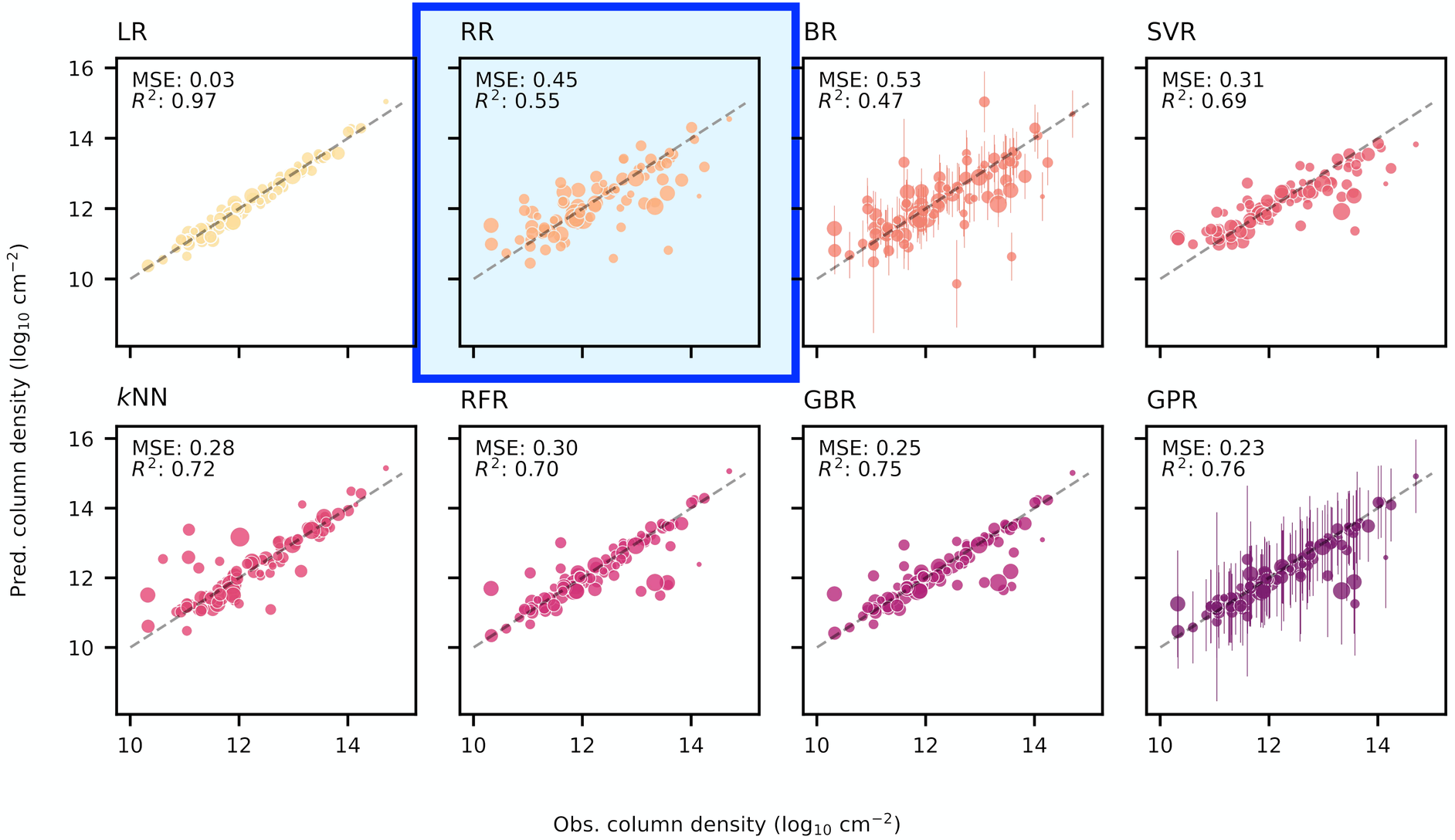
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



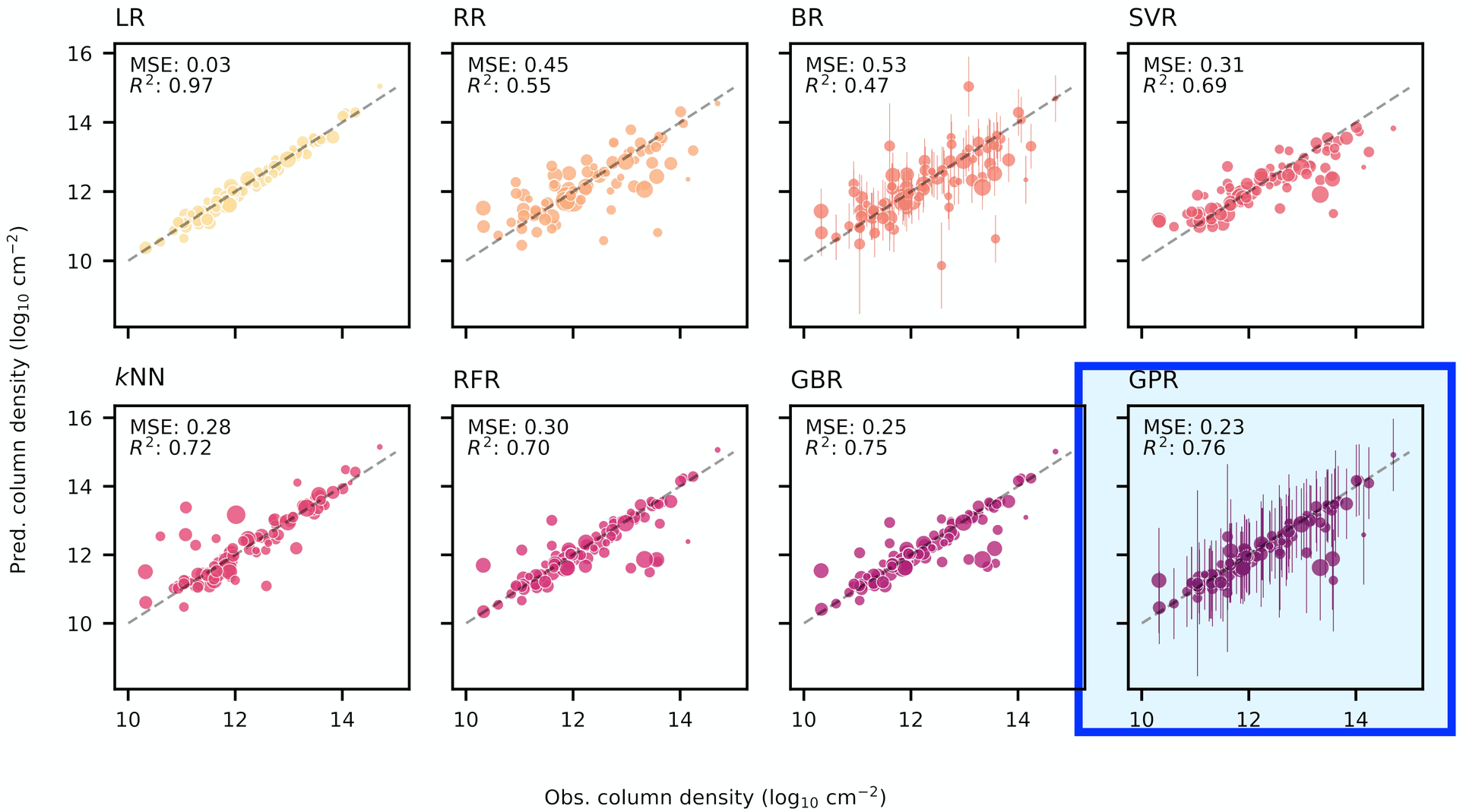
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



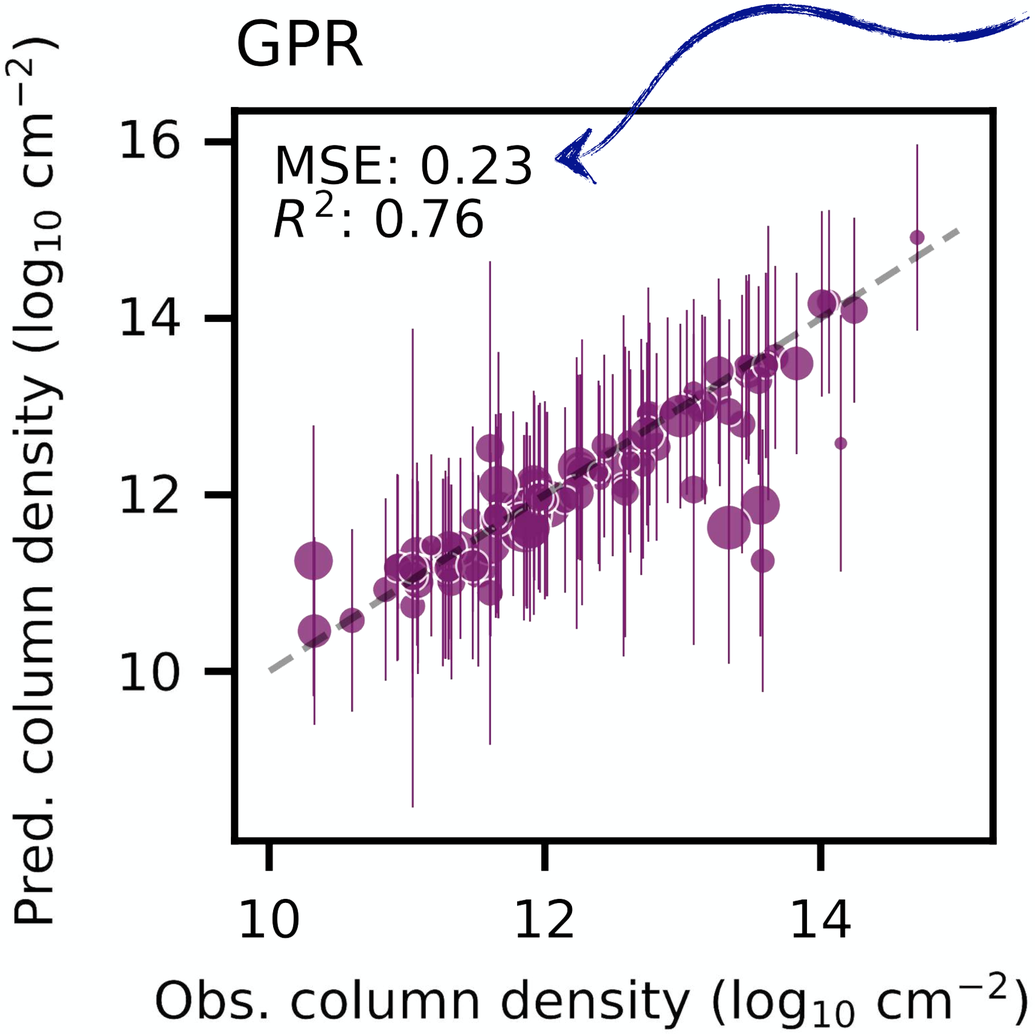
Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6



Abundances of Molecules in TMC-1

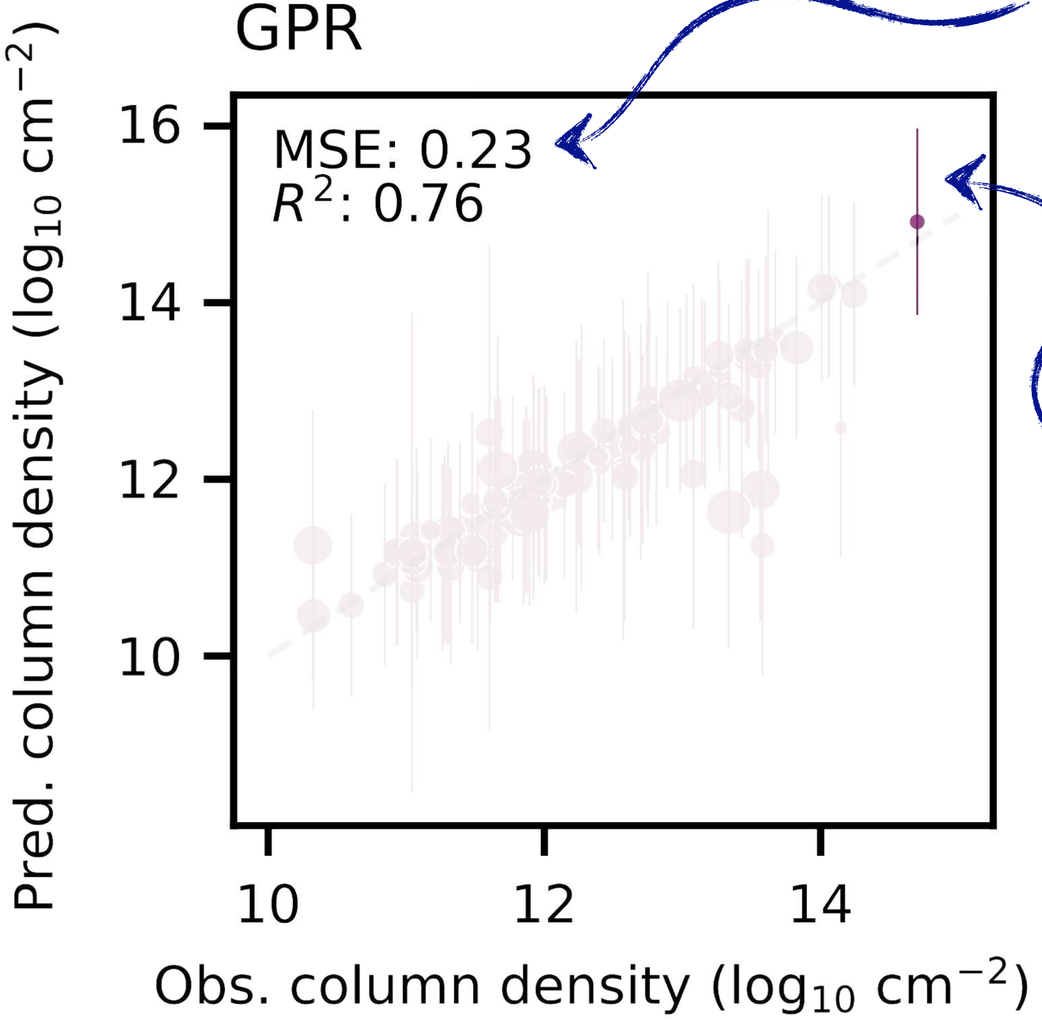
Lee et al. 2021 ApJL 917, L6



Error (factor of ~3) in reproducing fitted data points

Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6

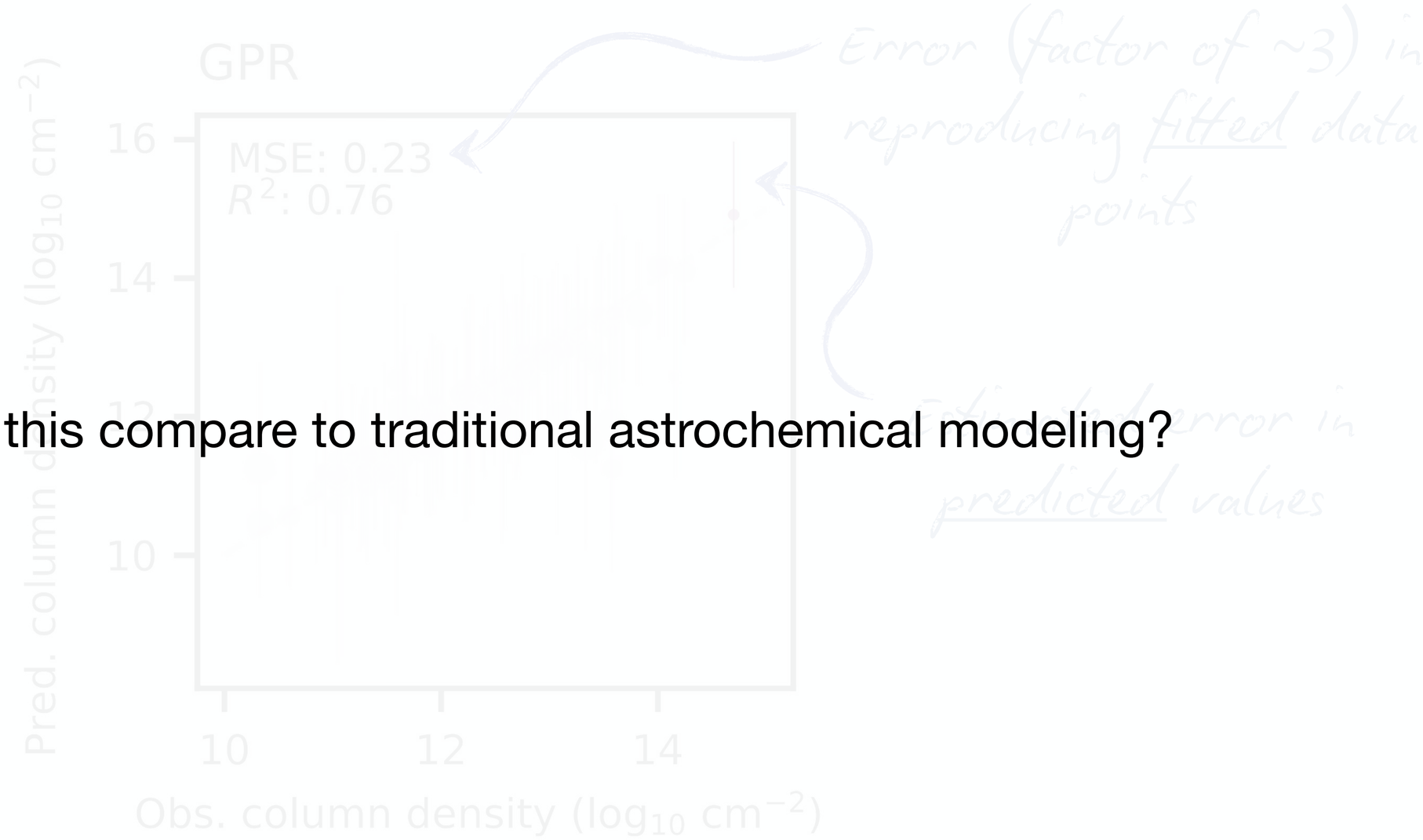


Error (factor of ~3) in reproducing fitted data points

Estimated error in predicted values

Abundances of Molecules in TMC-1

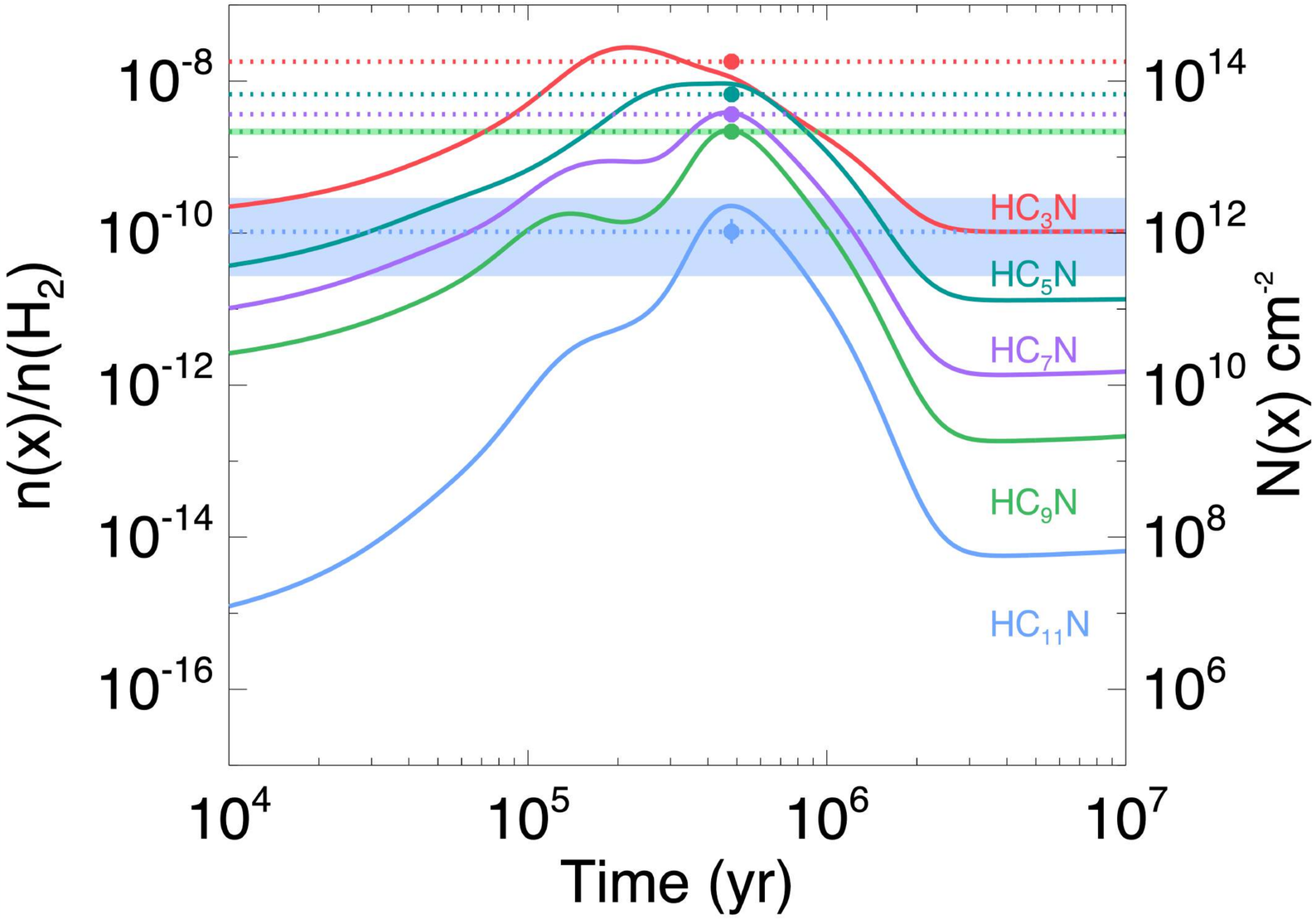
Lee et al. 2021 ApJL 917, L6



How does this compare to traditional astrochemical modeling?

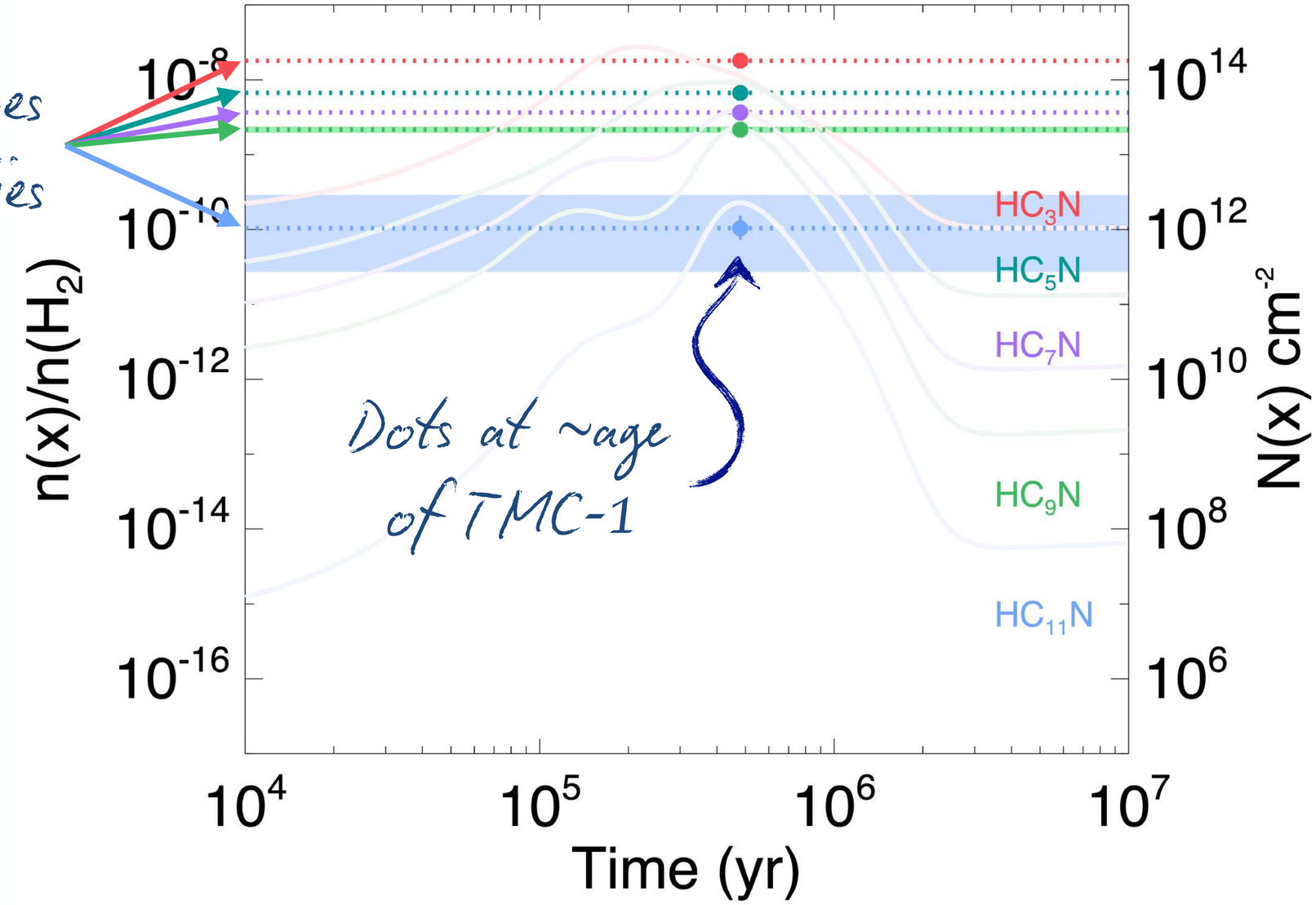
Abundances of Molecules in TMC-1

Loomis et al. 2021 *Nat. Astro.* 5, 188



Abundances of Molecules in TMC-1

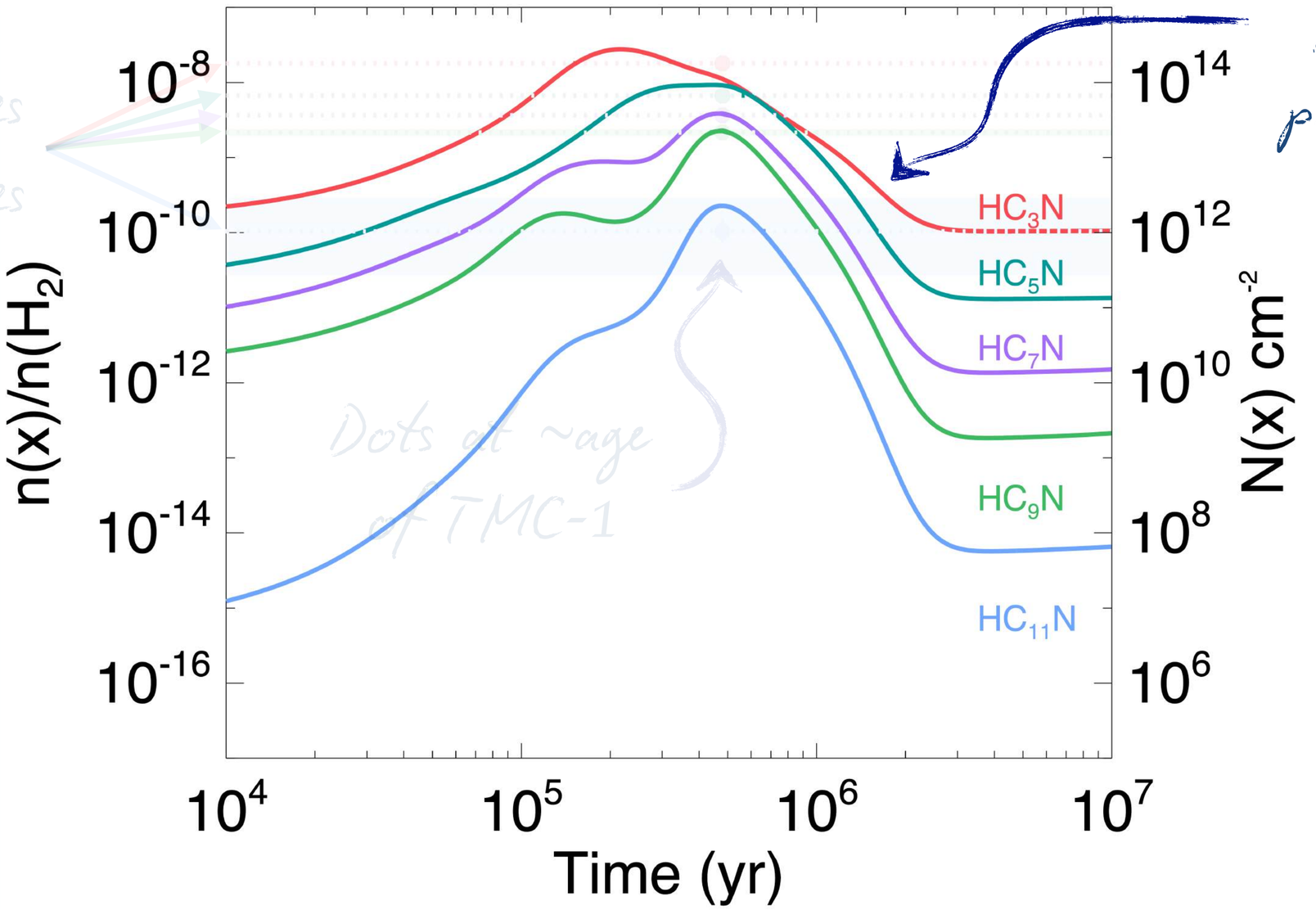
Loomis et al. 2021 *Nat. Astro.* 5, 188



Abundances of Molecules in TMC-1

Loomis et al. 2021 *Nat. Astro.* 5, 188

Observed Values w/ Uncertainties

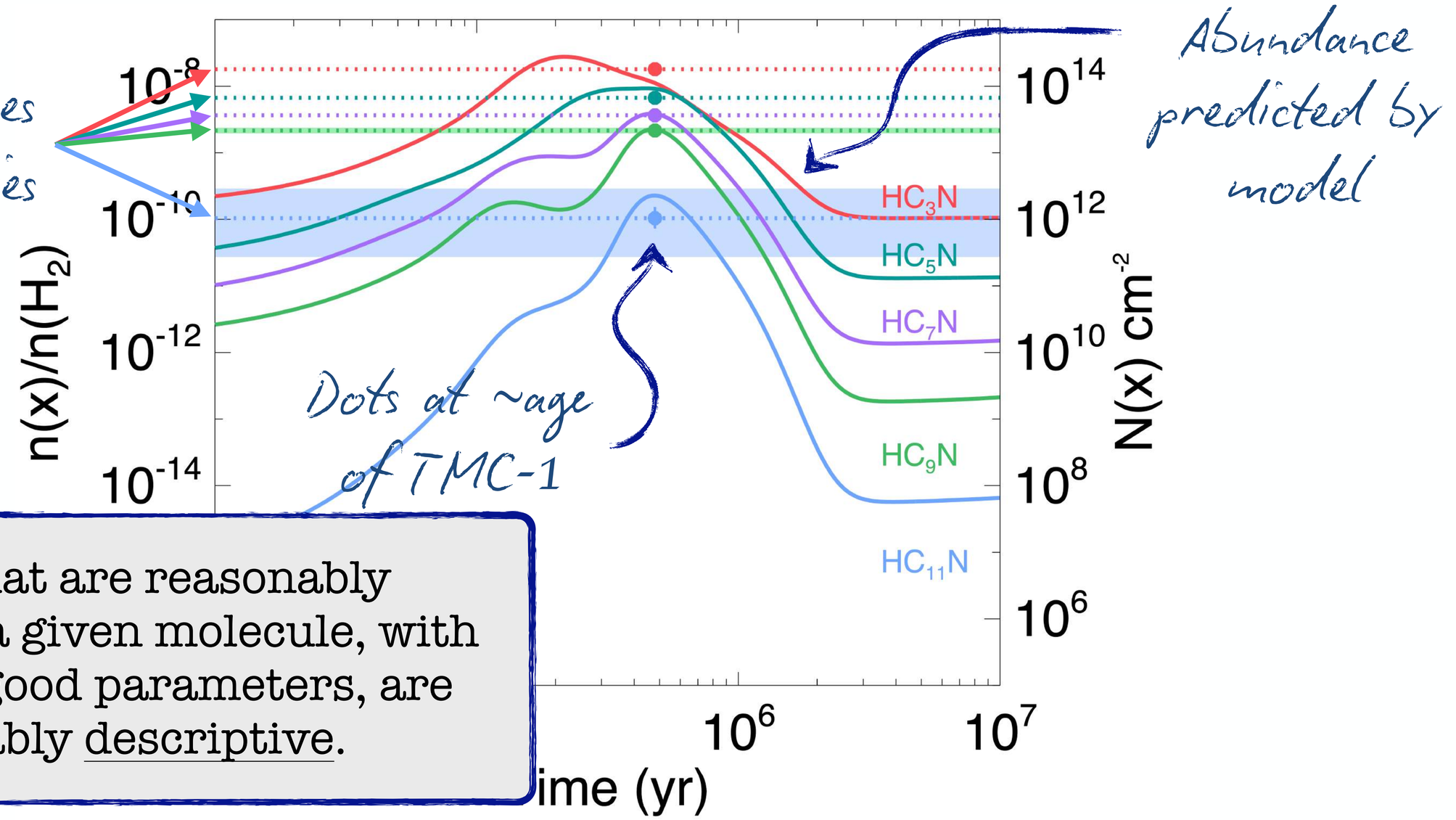


Dots at ~age of TMC-1

Abundance predicted by model

Abundances of Molecules in TMC-1

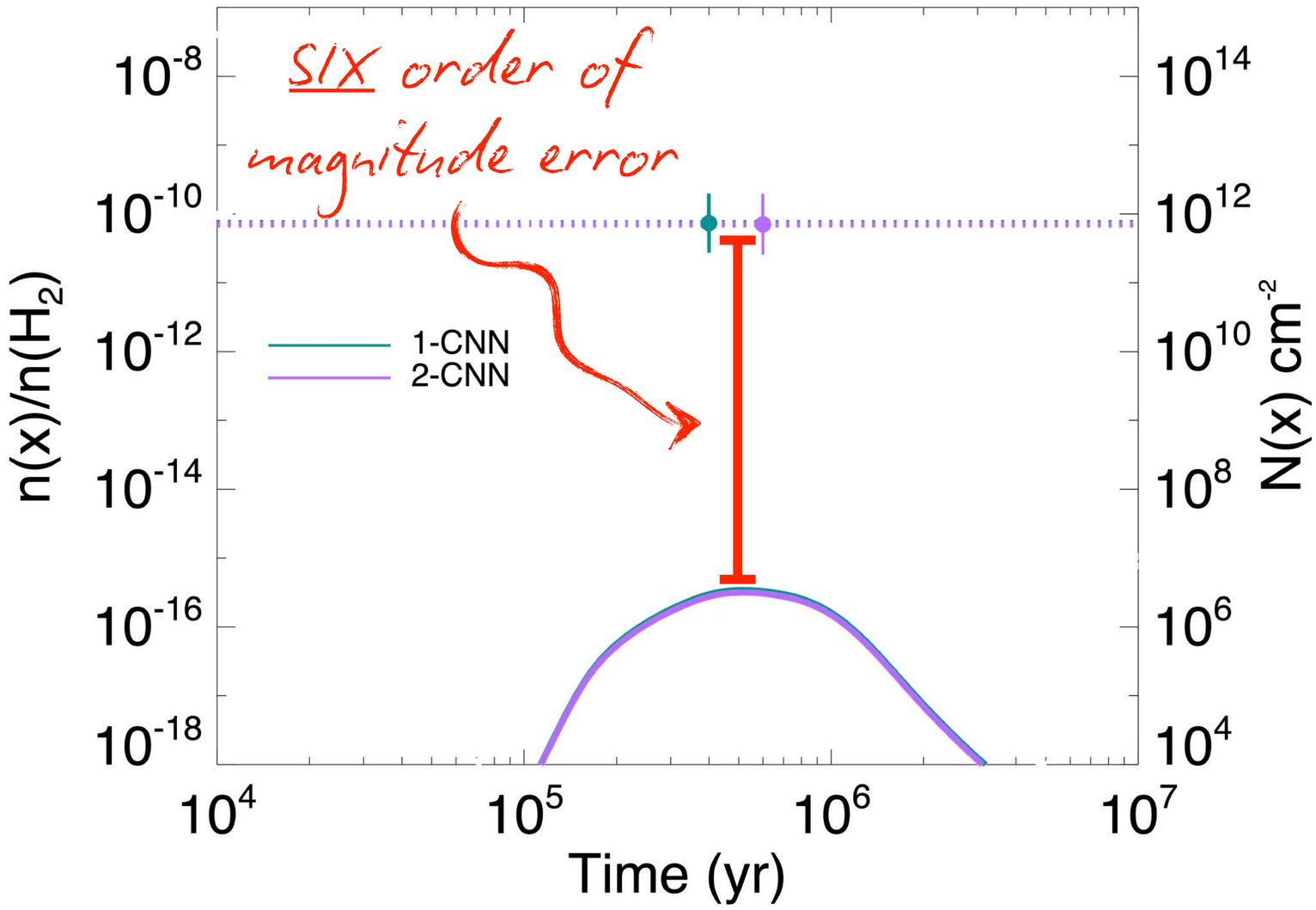
Loomis et al. 2021 *Nat. Astro.* 5, 188

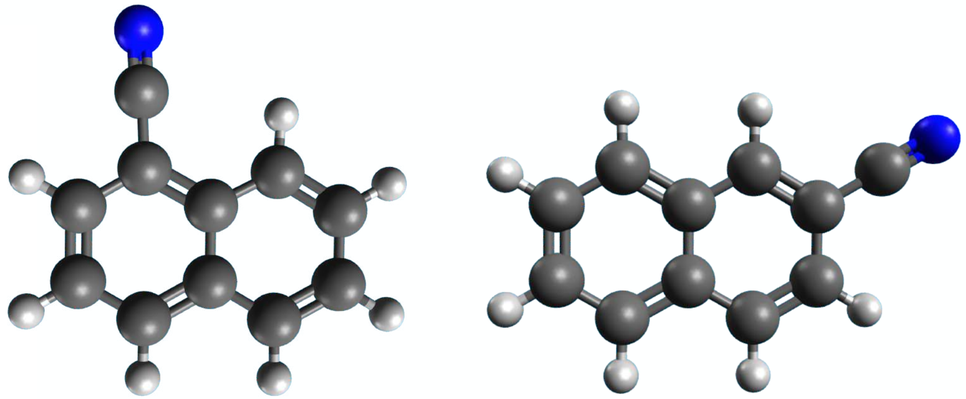


Models that are reasonably complete for a given molecule, with reasonably good parameters, are reasonably descriptive.

Abundances of Molecules in TMC-1

McGuire et al. 2021 *Science* 371, 1265

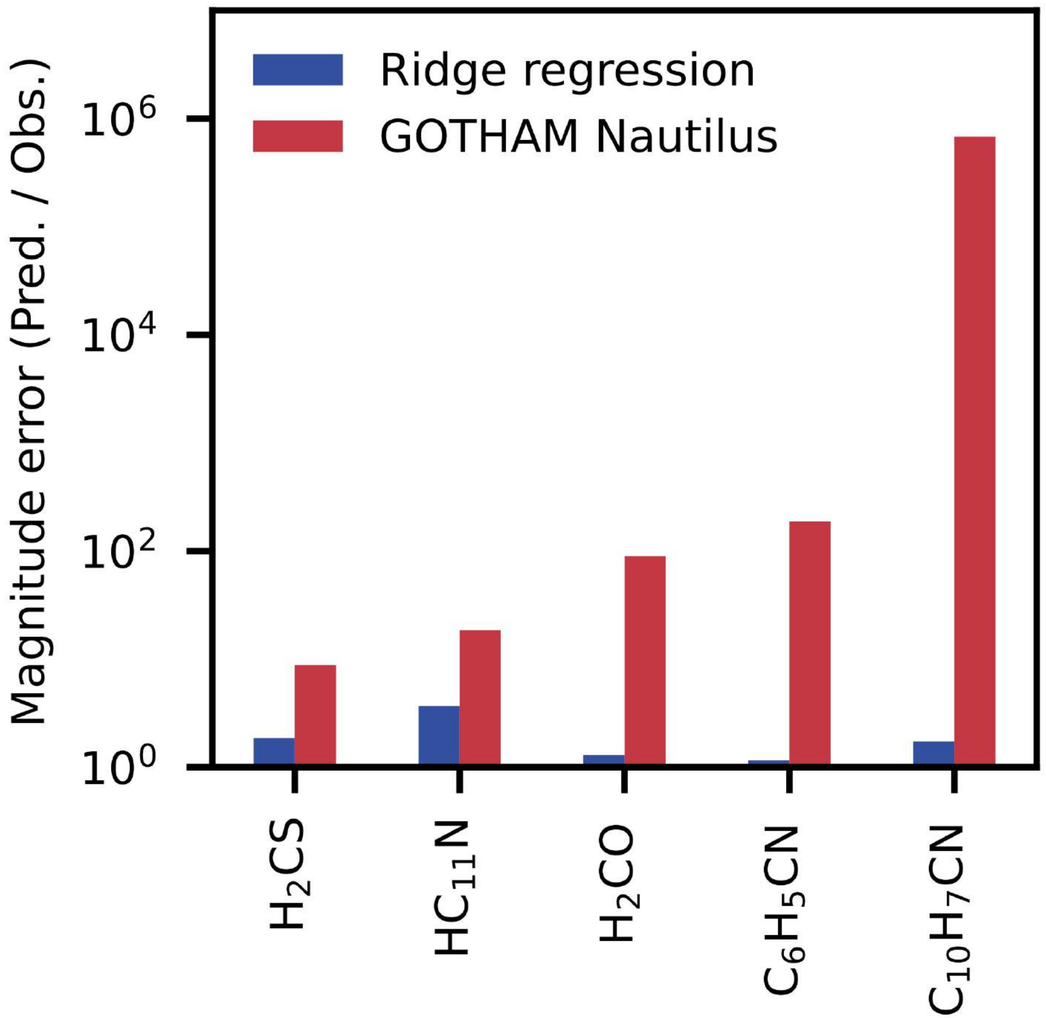






Abundances of Molecules in TMC-1

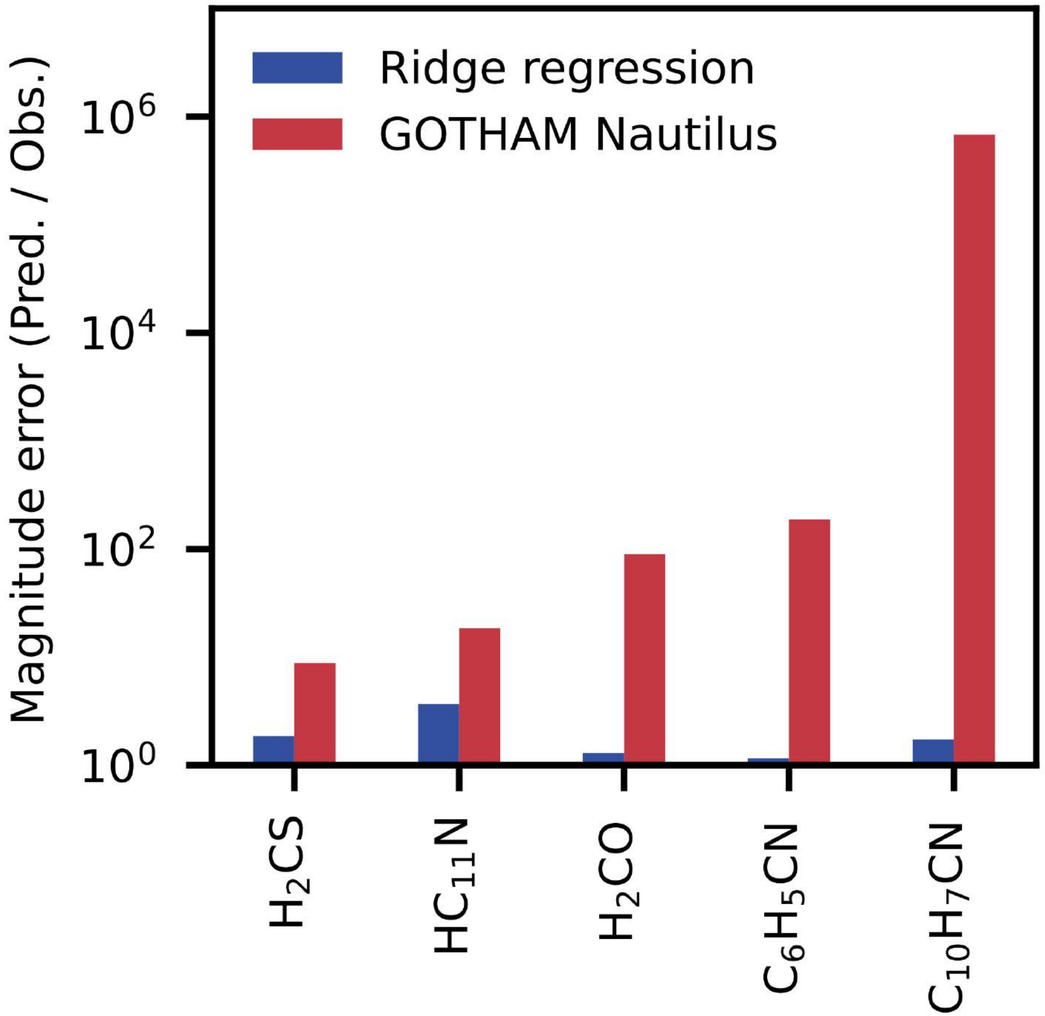
Lee et al. 2021 ApJL 917, L6



Large Bar
=
Large Error

Abundances of Molecules in TMC-1

Lee et al. 2021 ApJL 917, L6

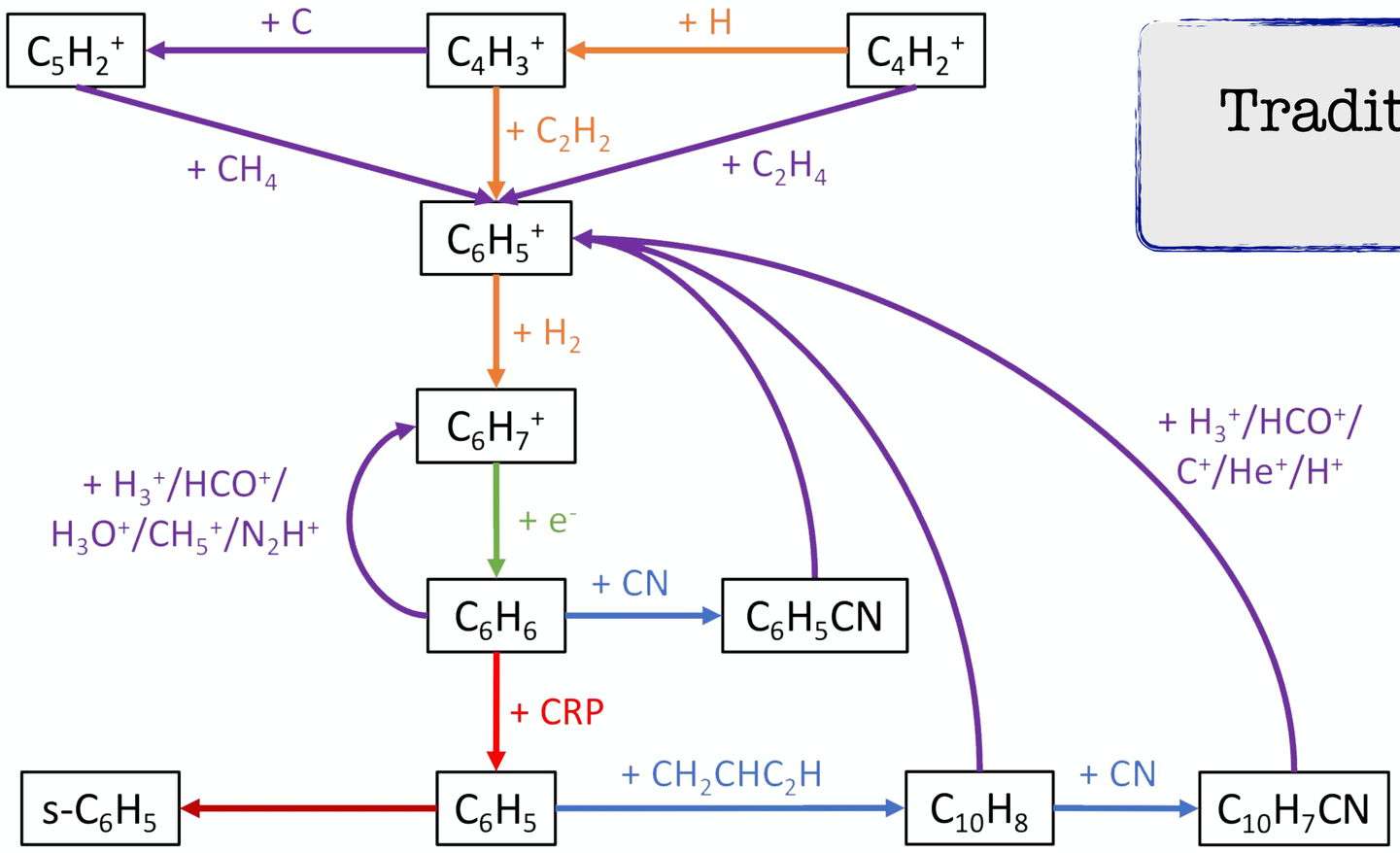


Models of abundances using machine-learned chemical intuition are remarkably descriptive, although overfit.

How about predictive capabilities, though?

Abundances of Molecules in TMC-1

Byrne et al. 2024 *PCCP* 26, 26734



Traditional models are very rarely predictive.

Abundances of Molecules in TMC-1

Byrne et al. 2024 *PCCP* 26, 26734

Astrochemical Modeling of Propargyl Radical Chemistry in TMC-1

Alex N. Byrne[1], Ci Xue[1], Ilsa R. Cooke[2], Michael C. McCarthy[3], and Brett A. McGuire[1,4]

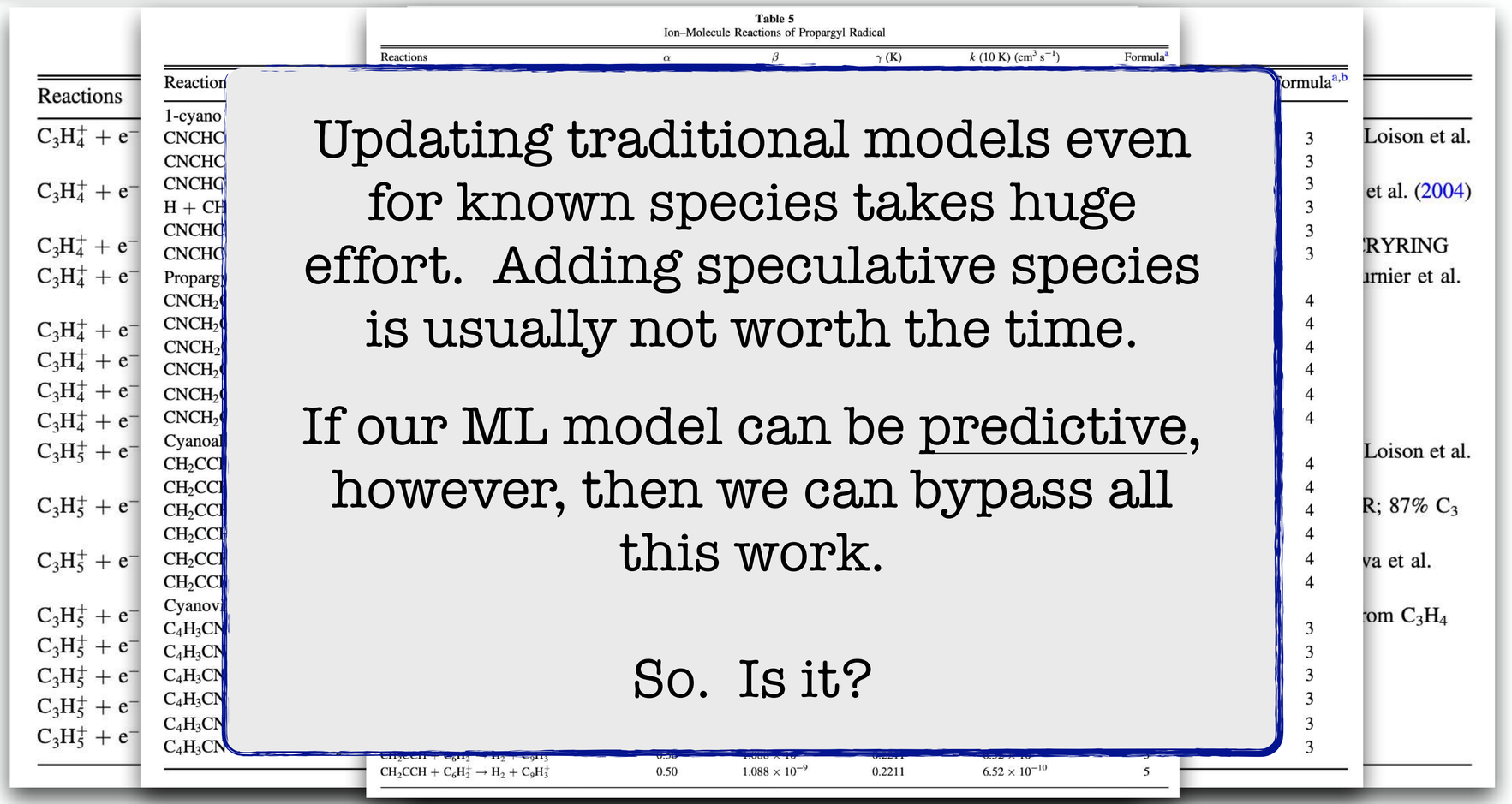


Table 5
Ion–Molecule Reactions of Propargyl Radical

Reactions	α	β	γ (K)	k (10 K) (cm^3 s^{-1})	Formula[a]
$CH_2CCH + C_6H_2^+ \rightarrow H_2 + C_9H_3^+$	0.50	1.088×10^{-9}	0.2211	6.52×10^{-10}	5

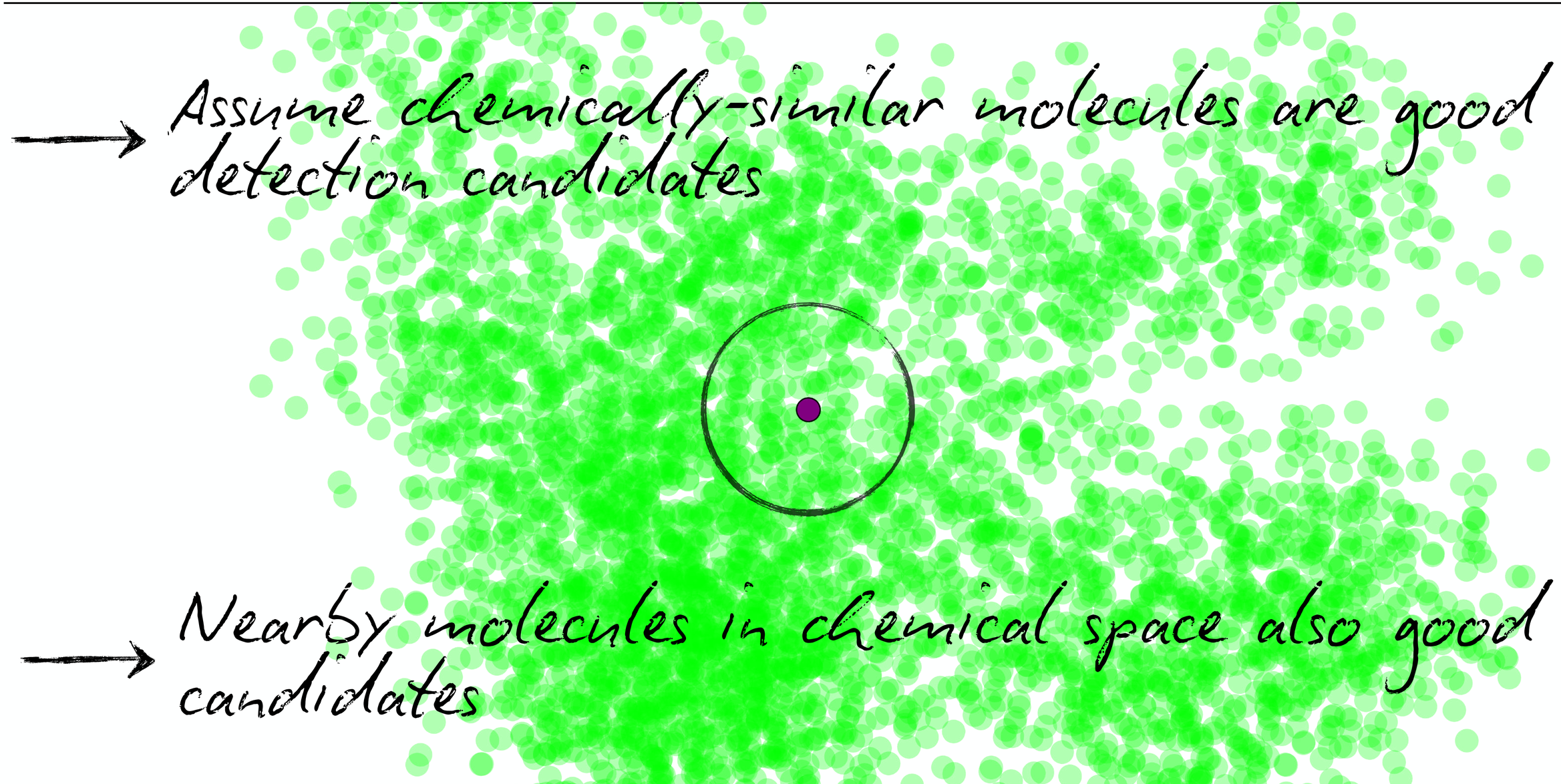
Updating traditional models even for known species takes huge effort. Adding speculative species is usually not worth the time.

If our ML model can be predictive, however, then we can bypass all this work.

So. Is it?

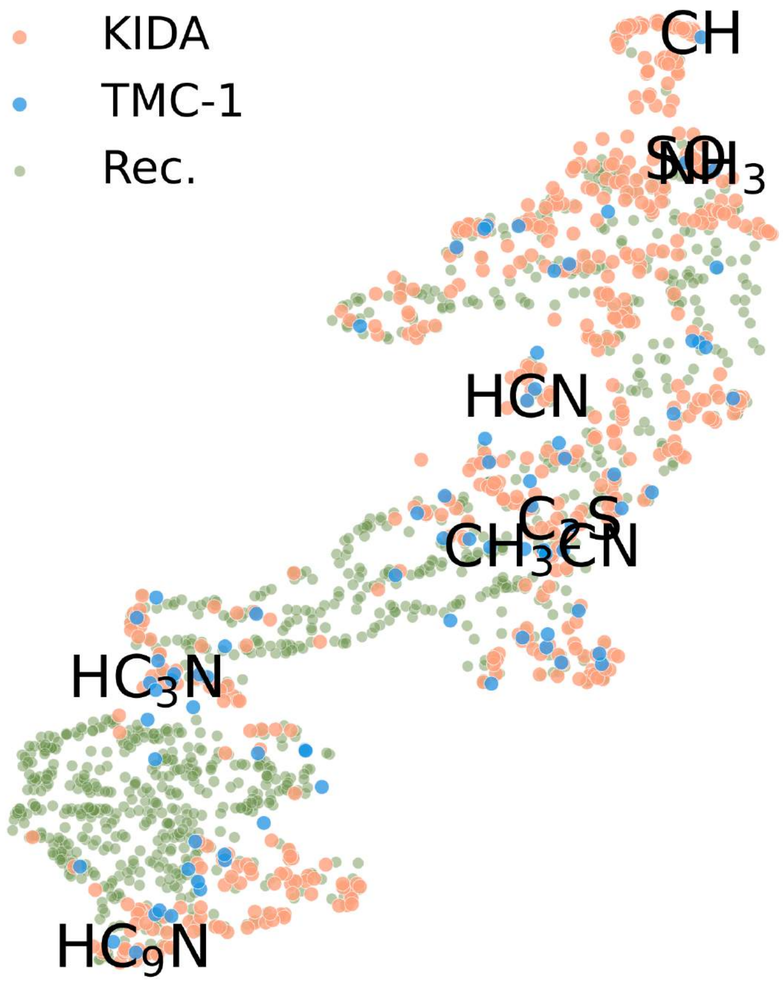
Predicting Abundances of Molecules in TMC-1

Lee et al. 2021 *ApJL* 917, L6

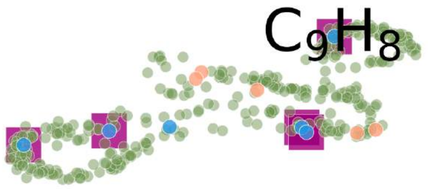


Predicting Abundances of Molecules in TMC-1

Lee et al. 2021 *ApJL* 917, L6

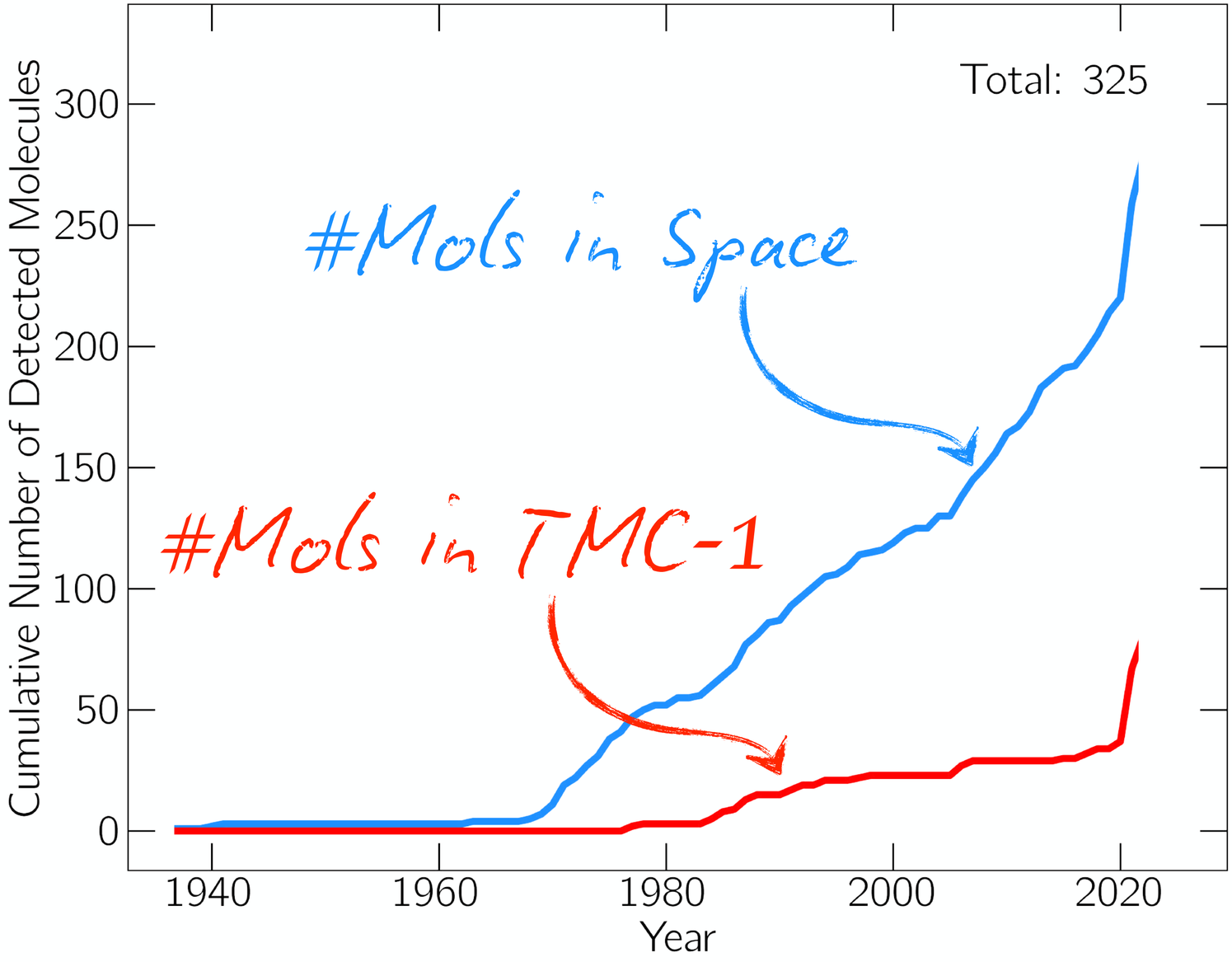


Generate a list of ~1500 recommended molecules



Abundances of Molecules in TMC-1

Lee et al. 2021 *ApJL* 917, L6
Toru Shay et al. 2025 *ApJ* 985, 123

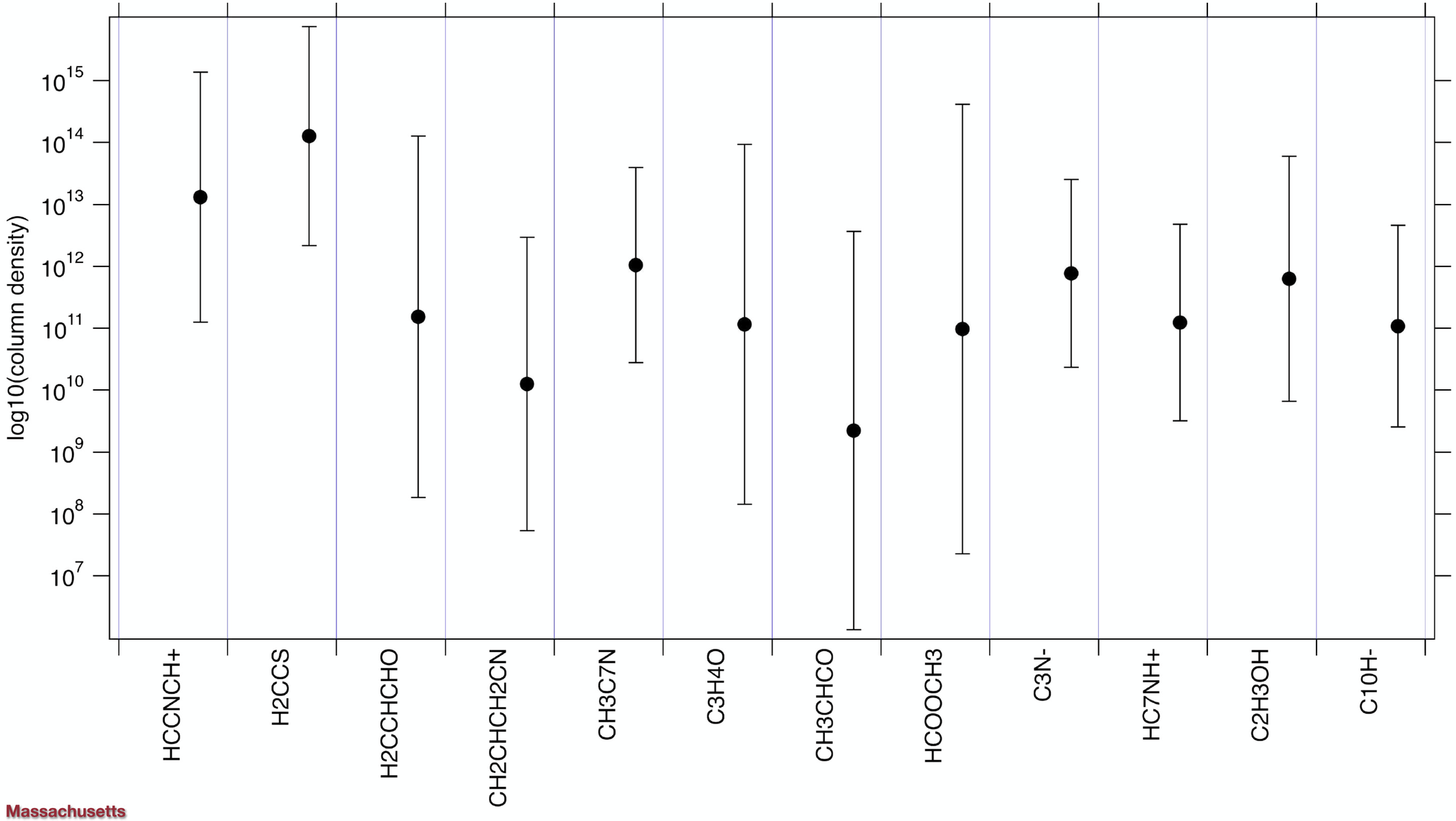


50 Total New Molecules in TMC-1 since Lee+2021

(New in the source + new in space)

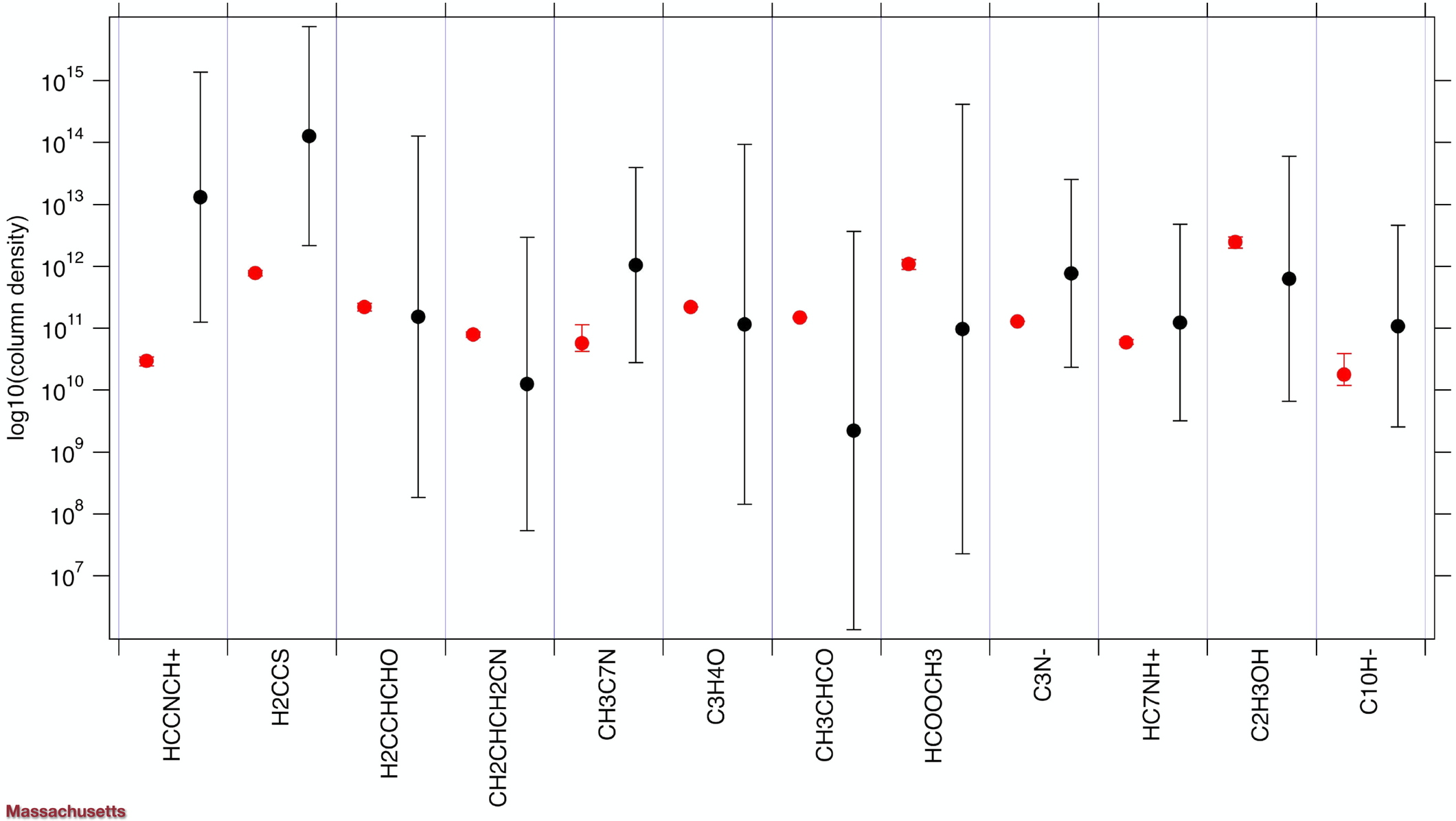
Abundances of Molecules in TMC-1

Lee et al. 2021 *ApJL* 917, L6
Toru Shay et al. 2025 *ApJ* 985, 123



Abundances of Molecules in TMC-1

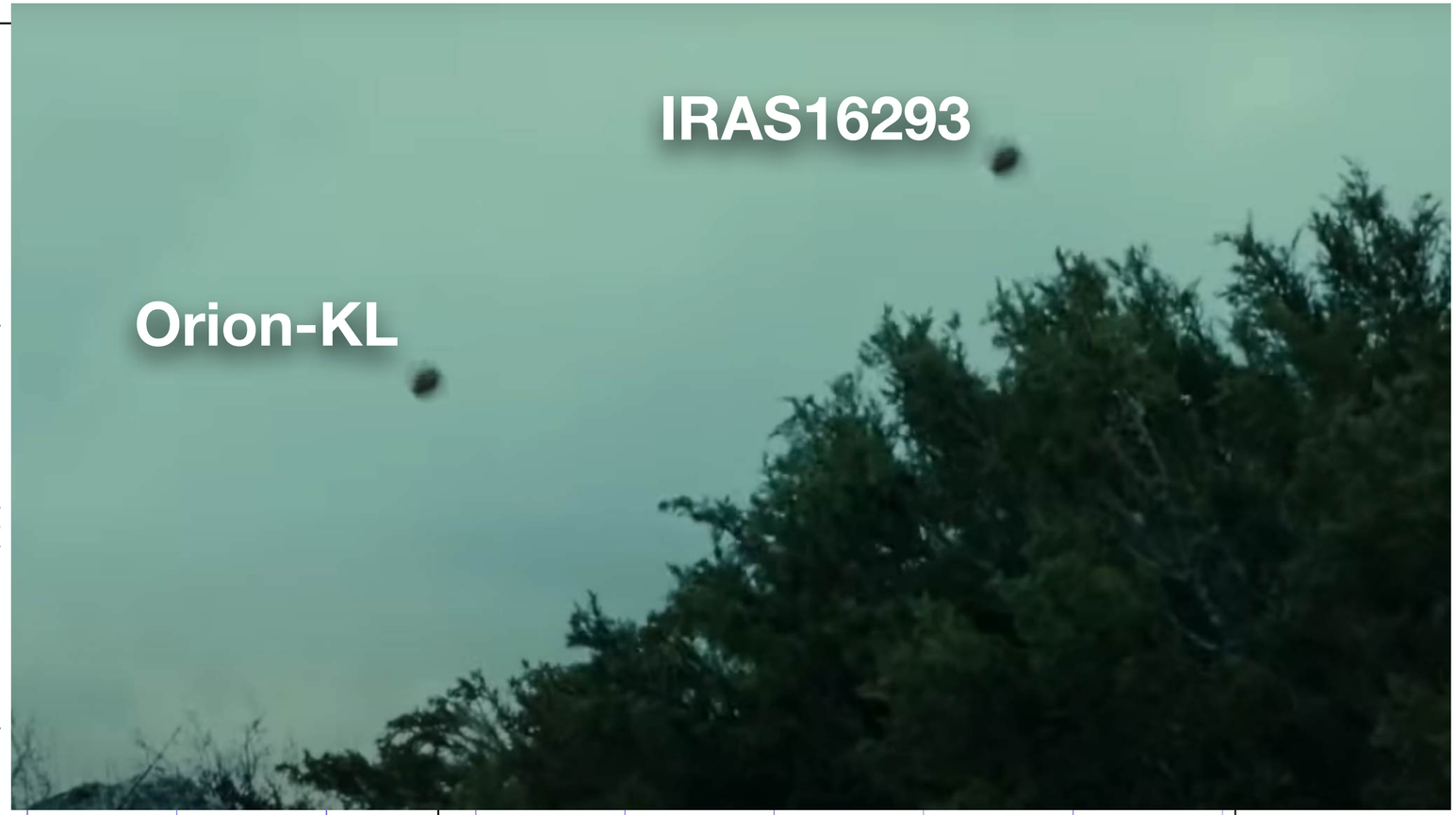
Lee et al. 2021 *ApJL* 917, L6
Toru Shay et al. 2025 *ApJ* 985, 123



Abundances of Molecules in TMC-1

Lee et al. 2021 *ApJL* 917, L6
Toru Shay et al. 2025 *ApJ* 985, 123





CH3C7N
C3H4O
CH3CHCO
HCOOCH3
C3N-
HC7NH+
C2H3OH
C10H-

Abundances of Molecules in TMC-1

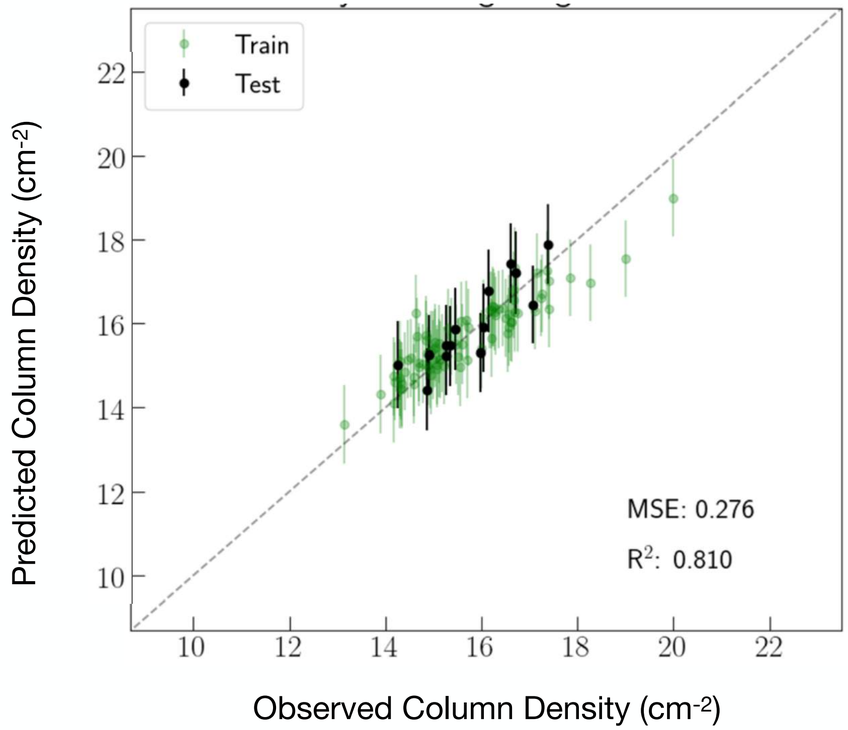
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]



Abundances of Molecules in TMC-1

Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

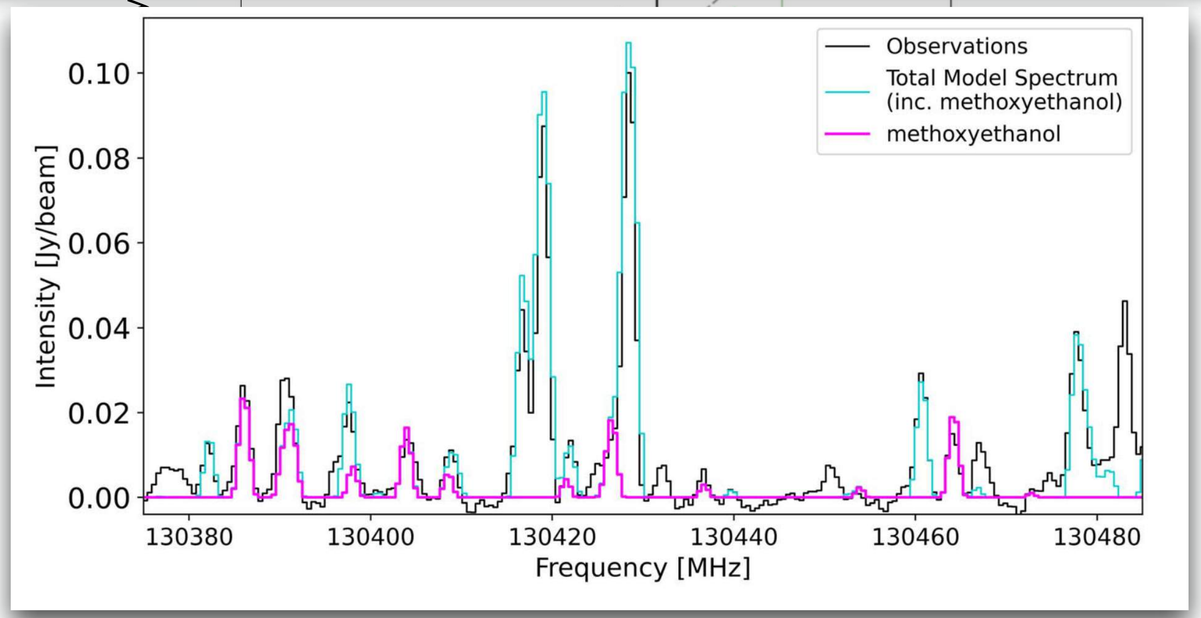
Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

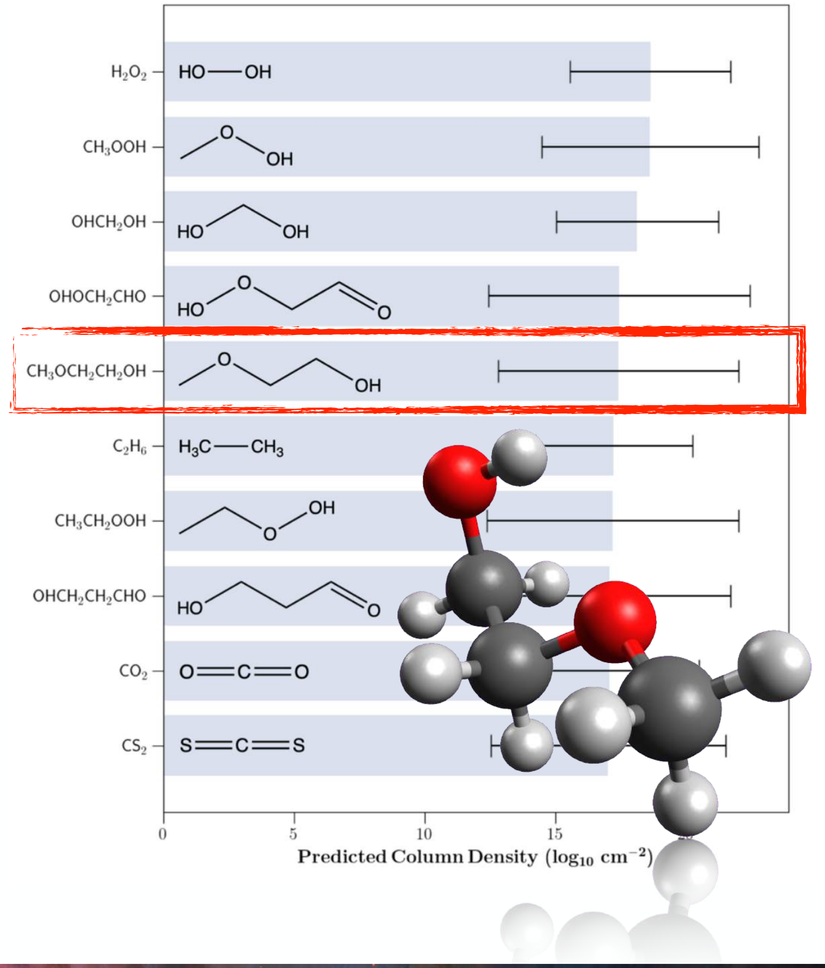
Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]

Rotational Spectrum and First Interstellar Detection of 2-methoxyethanol Using ALMA Observations of NGC 6334I

Zachary T. P. Fried[1], Samer J. El-Abd[2,3], Brian M. Hays[4], Gabi Wenzel[1], Alex N. Byrne[1], Laurent Margulès[4], Roman A. Motiyenko[4], Steven T. Shipman[5], Maria P. Horne[6], Jes K. Jørgensen[7], Crystal L. Brogan[3], Todd R. Hunter[3], Anthony J. Remijan[3], Andrew Lipnicky[3], Ryan A. Loomis[3], and Brett A. McGuire[1,3]





Abundances of Molecules in TMC-1

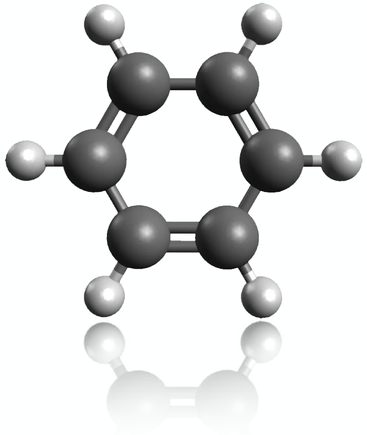
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]



SMILES → c1ccccc1 → (mol2vec) → 300-d embedding → (PCA) → 70-d embedding → (Manual Isos) → 89-d embedding →

SELFIES → [C][=C][C][=C][C][=C][Ring1][=Branch1] → (VICGAE) → 32-d embedding →

Token-level Context natively included, plus supports isotopes

Abundances of Molecules in TMC-1

Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

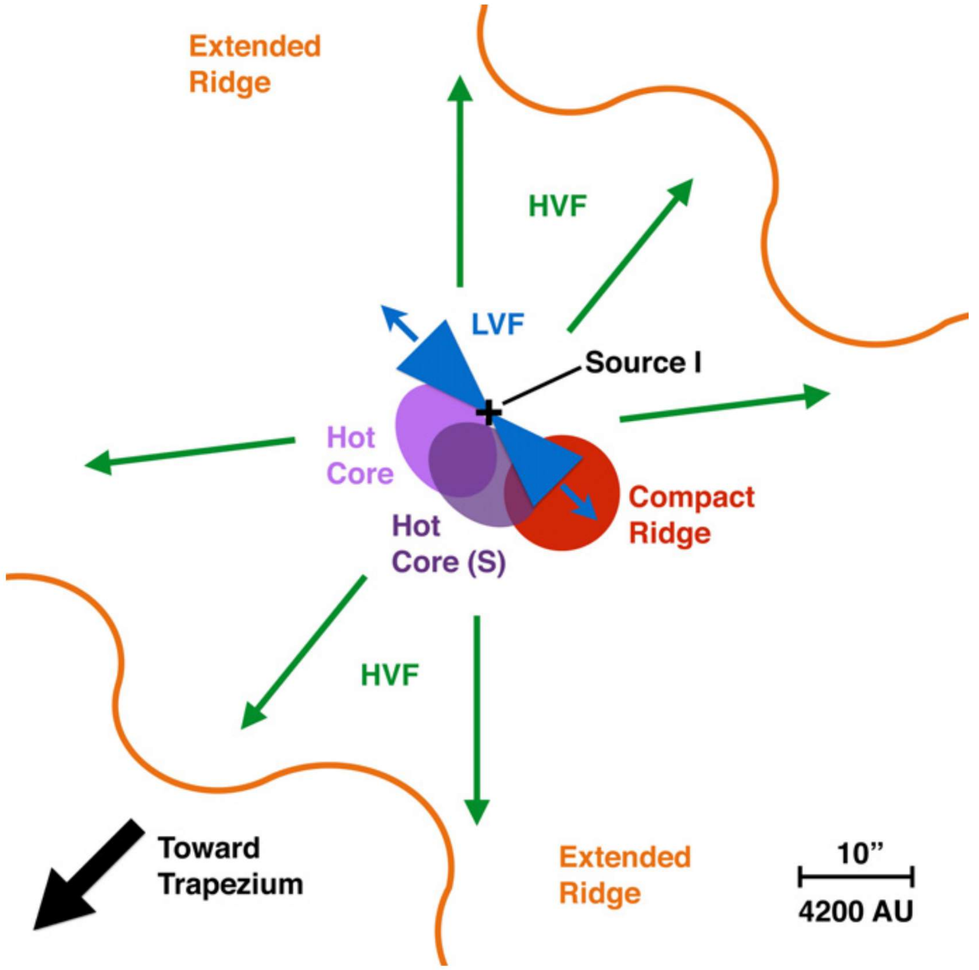
Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]

Hot Core

Compact Ridge

Extended Ridge

Plateau



Abundances of Molecules in TMC-1

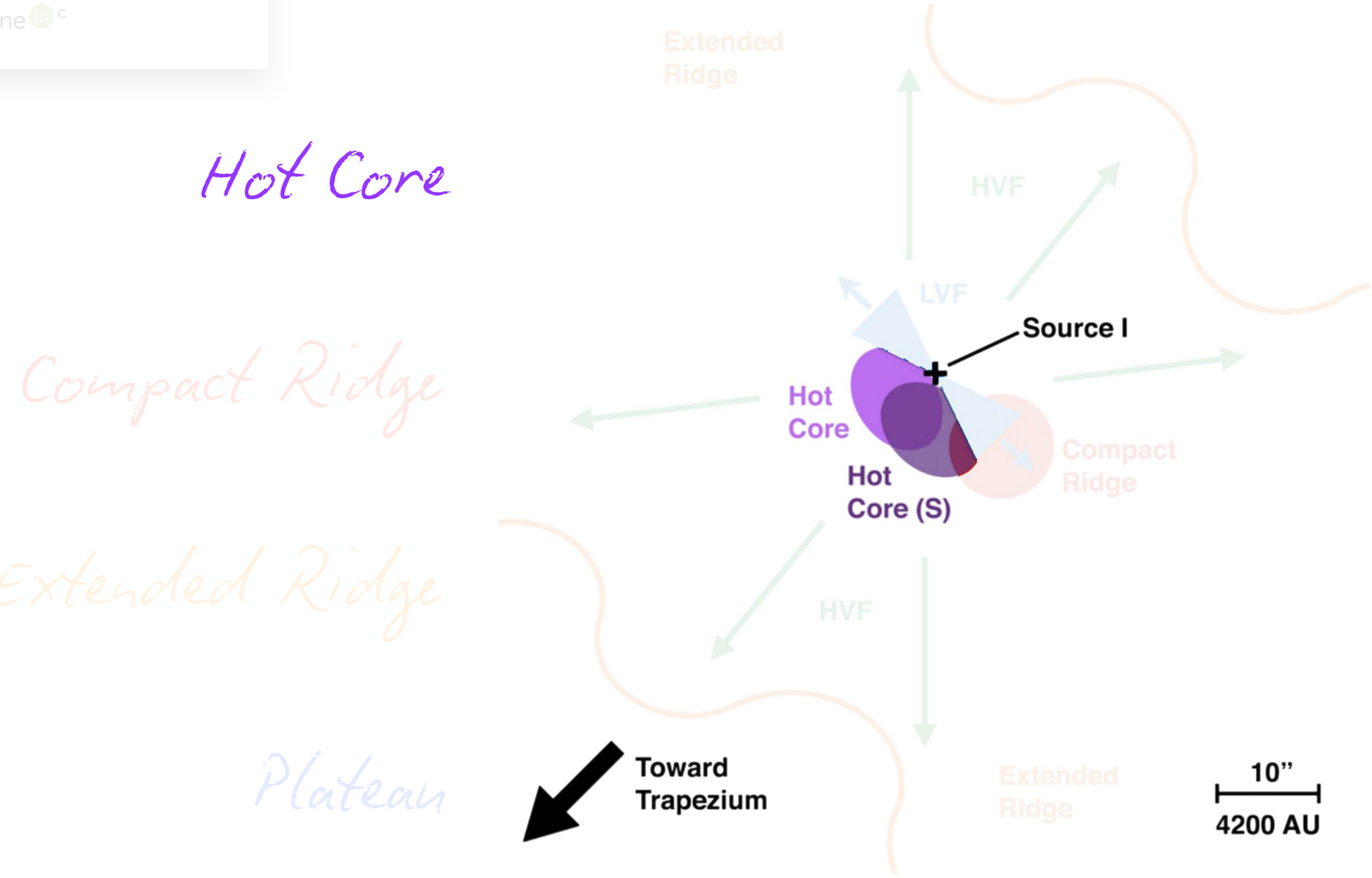
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]



Abundances of Molecules in TMC-1

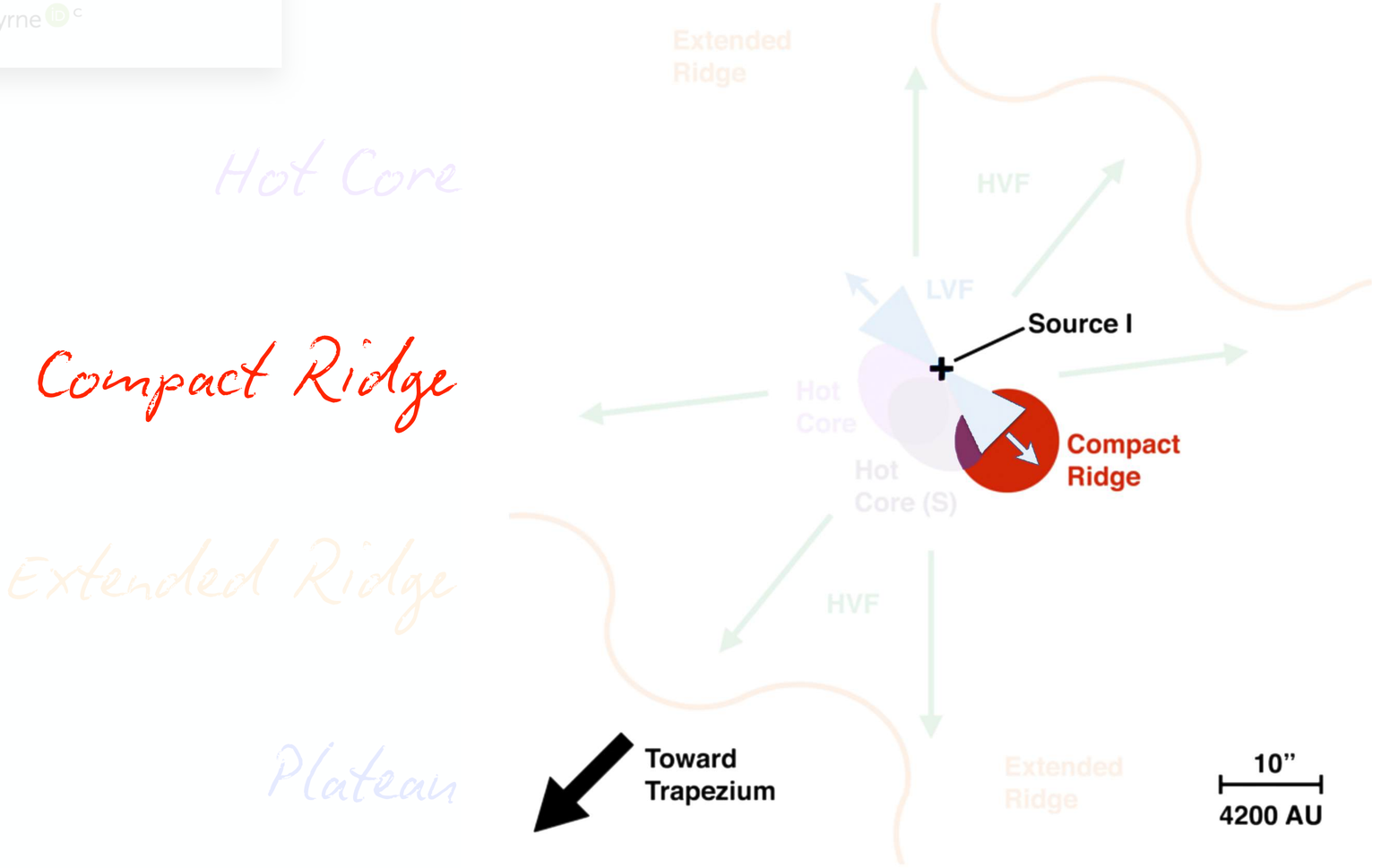
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]



Abundances of Molecules in TMC-1

Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

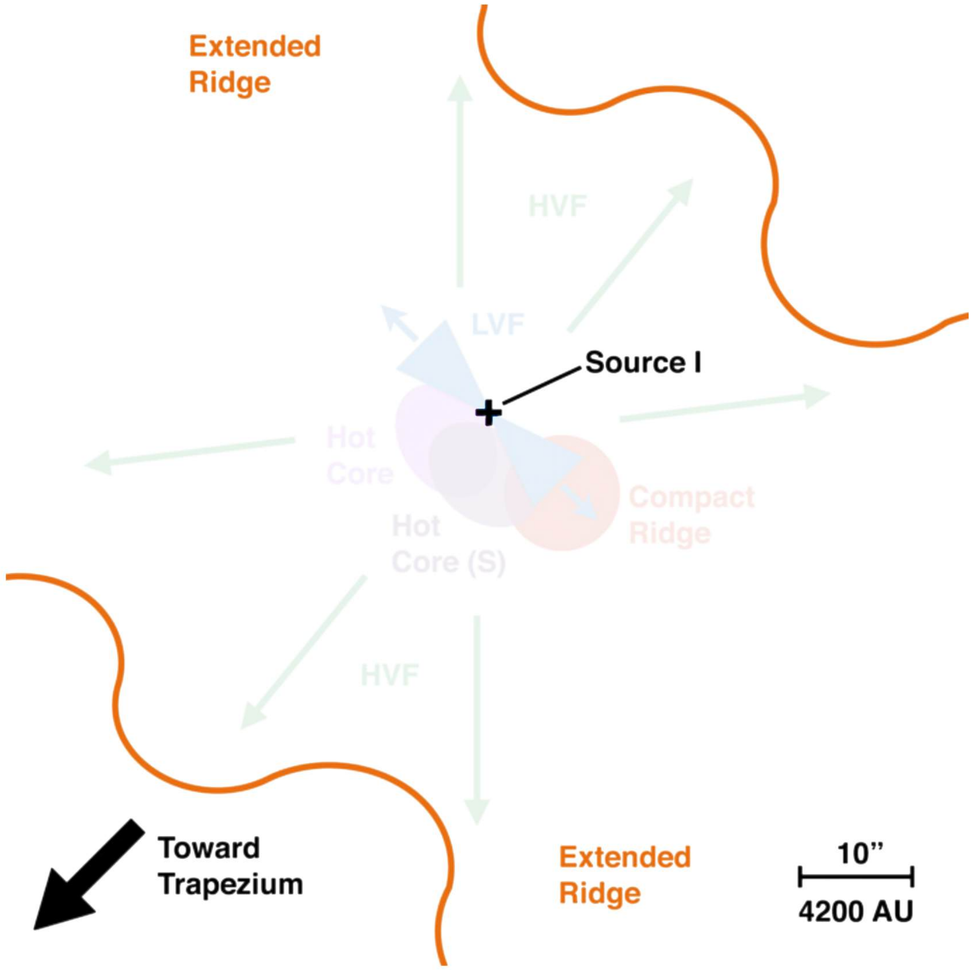
Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]

Hot Core

Compact Ridge

Extended Ridge

Plateau



Abundances of Molecules in TMC-1

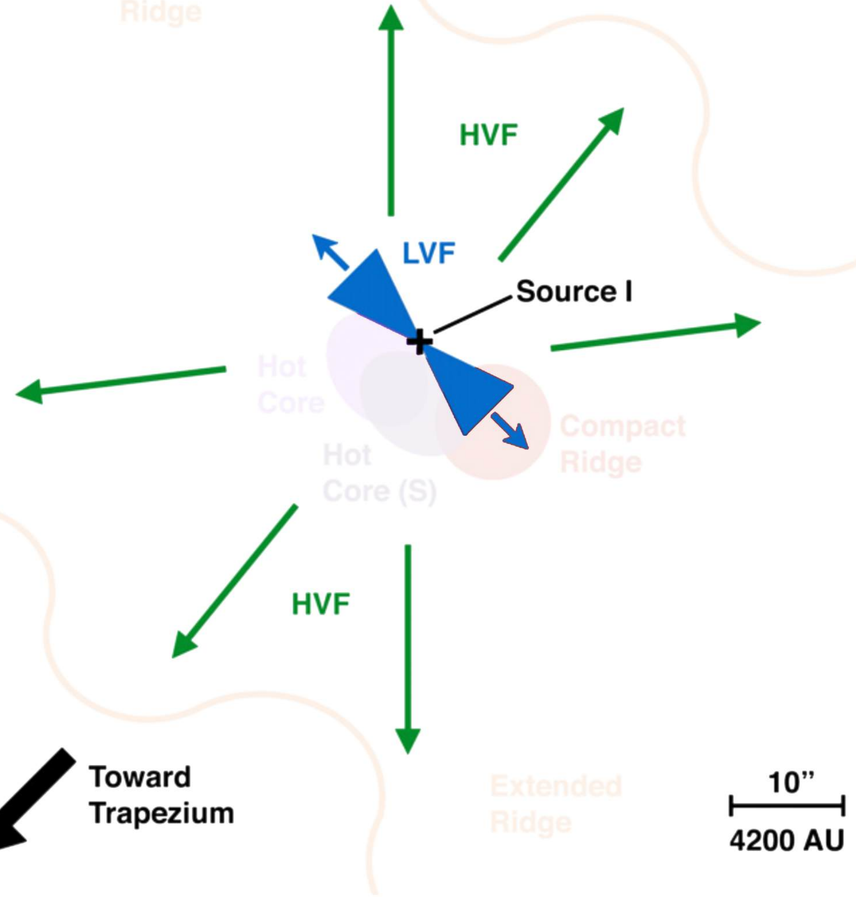
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]



Hot Core

Compact Ridge

Extended Ridge

Plateau

Abundances of Molecules in TMC-1

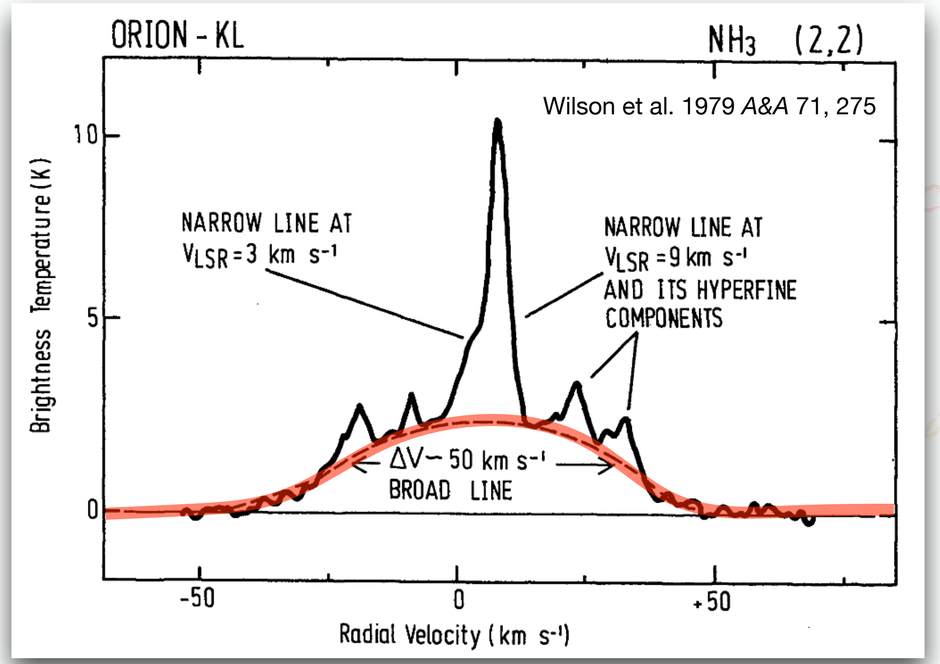
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

FURTHER OBSERVATIONAL STUDIES OF THE HIGH VELOCITY
MOLECULAR SOURCE ("PLATEAU") IN ORION

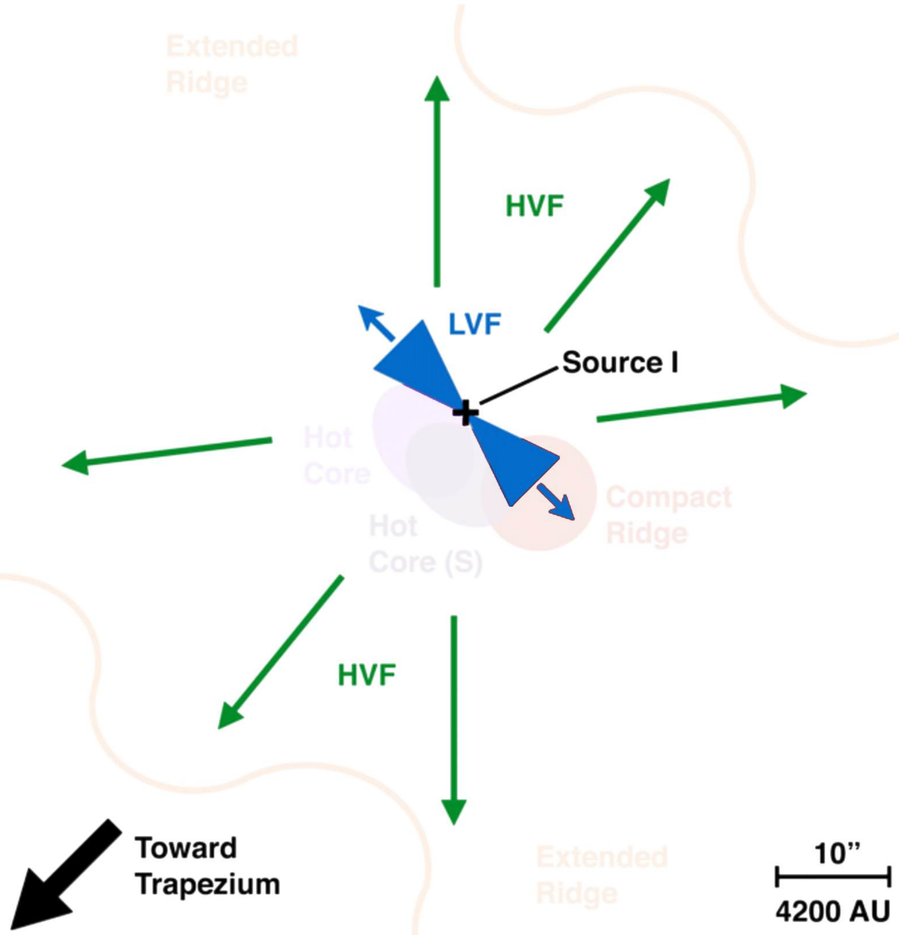
T.B.H. Kuiper
Jet Propulsion Laboratory, California Institute of Technology
E.N. Rodriguez Kuiper
Western Laboratories, Ball Aerospace Systems Divisions
B. Zuckerman
Astronomy Program, University of Maryland

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]



Plateau



Abundances of Molecules in TMC-1

Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

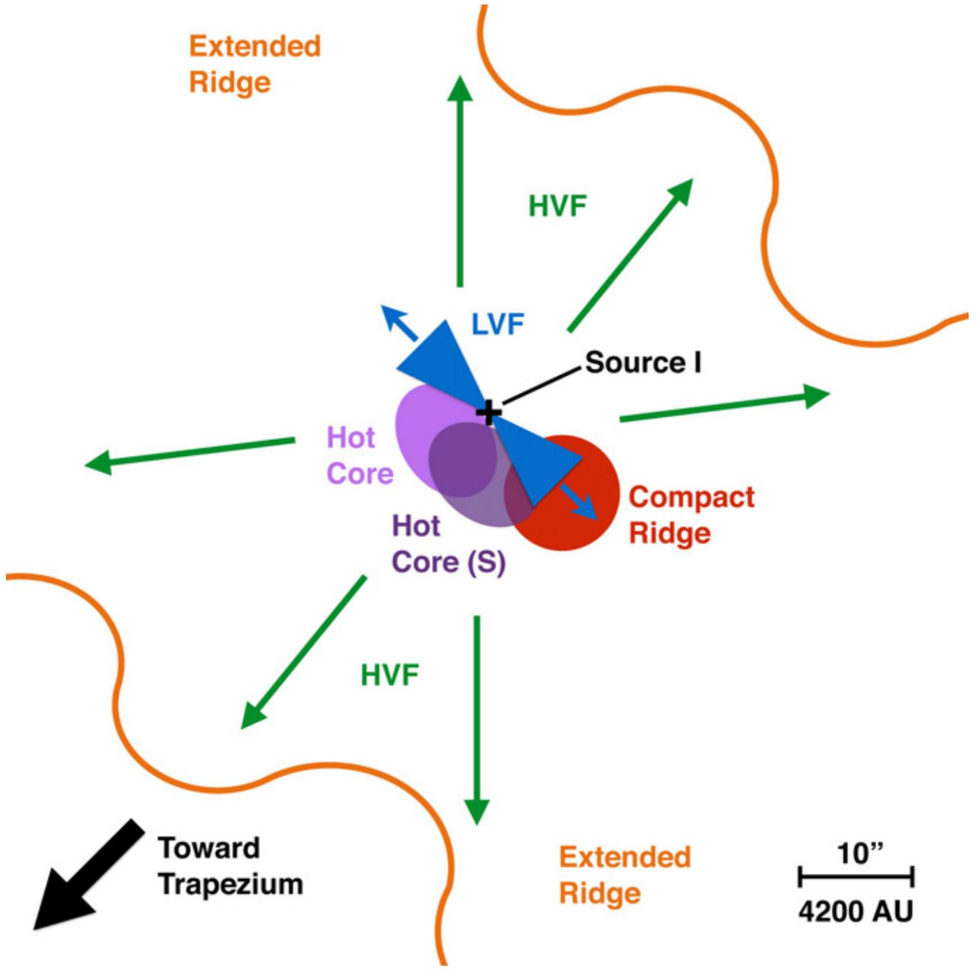
Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]

Hot Core

Compact Ridge

Extended Ridge

Plateau



Abundances of Molecules in TMC-1

Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]

$$[x_1\ x_2\ x_3\ \ldots\ \ldots\ x_{32}]T_{rot}\ v_{lsr}\ \Delta v\ vib\ iso\ env]$$

Abundances of Molecules in TMC-1

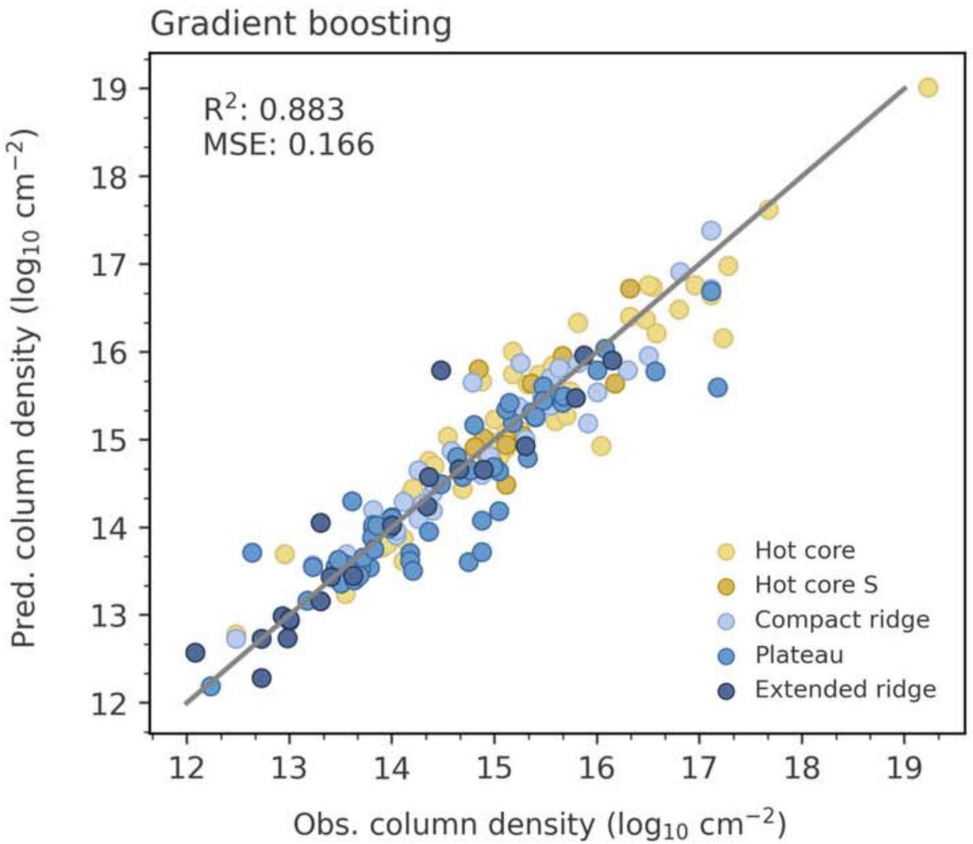
Fried et al. 2023 *Dig. Disc.* 2, 952
Scolati et al. 2023 *ApJ* 959, 108
Crockett et al. 2014 *ApJ* 787, 112

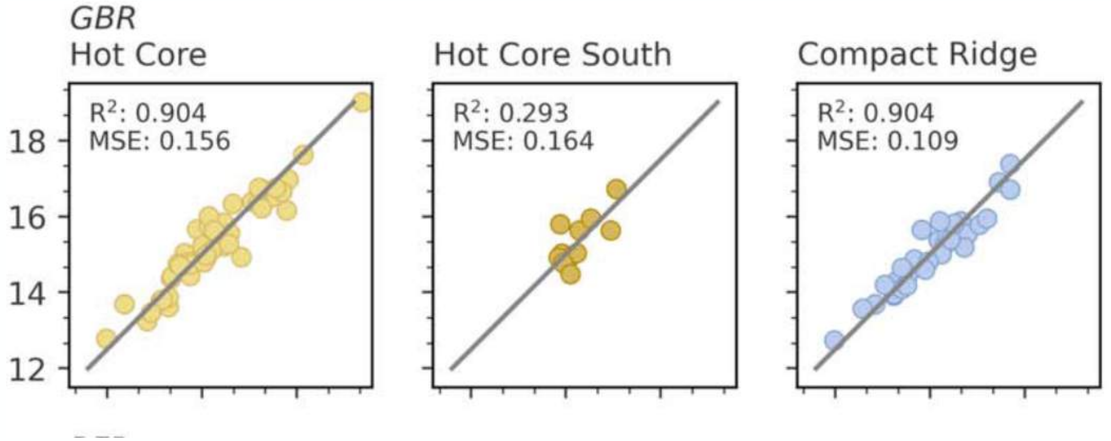
Implementation of rare isotopologues into machine learning of the chemical inventory of the solar-type protostellar source IRAS 16293-2422

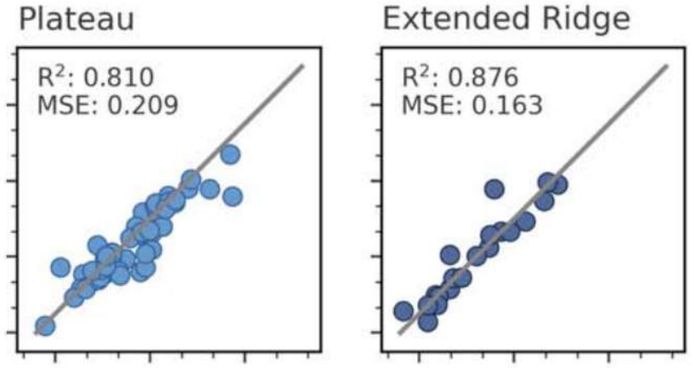
Zachary T. P. Fried, *[a] Kin Long Kelvin Lee, [b] Alex N. Byrne [c] and Brett A. McGuire *[de]

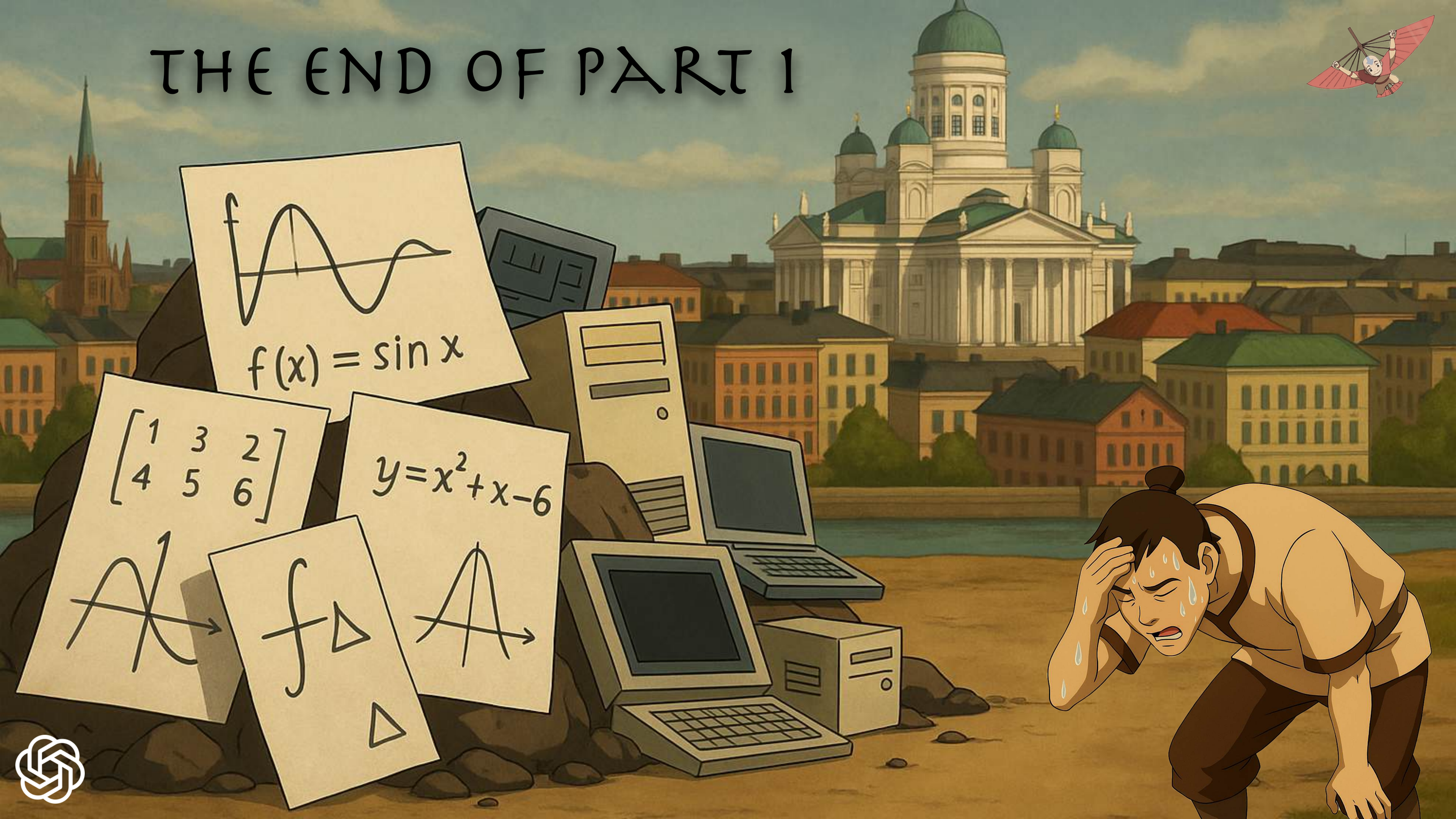
Explaining the Chemical Inventory of Orion KL through Machine Learning

Haley N. Scolati[1], Anthony J. Remijan[2], Eric Herbst[1,3], Brett A. McGuire[2,4], and Kin Long Kelvin Lee[5]








THE END OF PART 1
f
f(x) = sin x
$\begin{bmatrix} 1 & 3 & 2 \\ 4 & 5 & 6 \end{bmatrix}$
$y=x^2+x-6$

Book Two: Chemical Similarity







Part II

Chemical Similarity



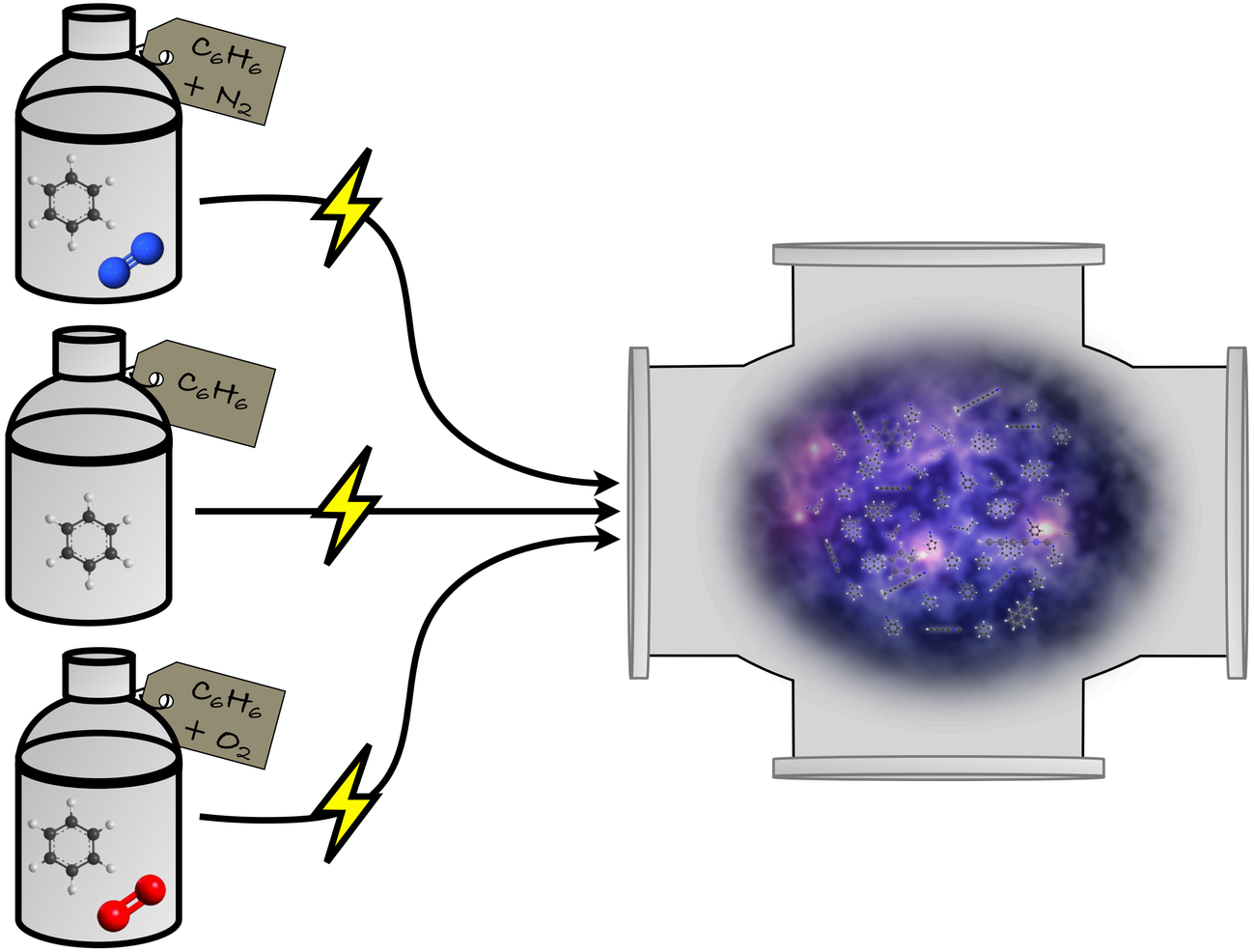


Automated Mixture Analysis via Structural Evaluation

Fried & McGuire 2024 *JPCA* 128, 8254



Mixture Analysis is Hard Work

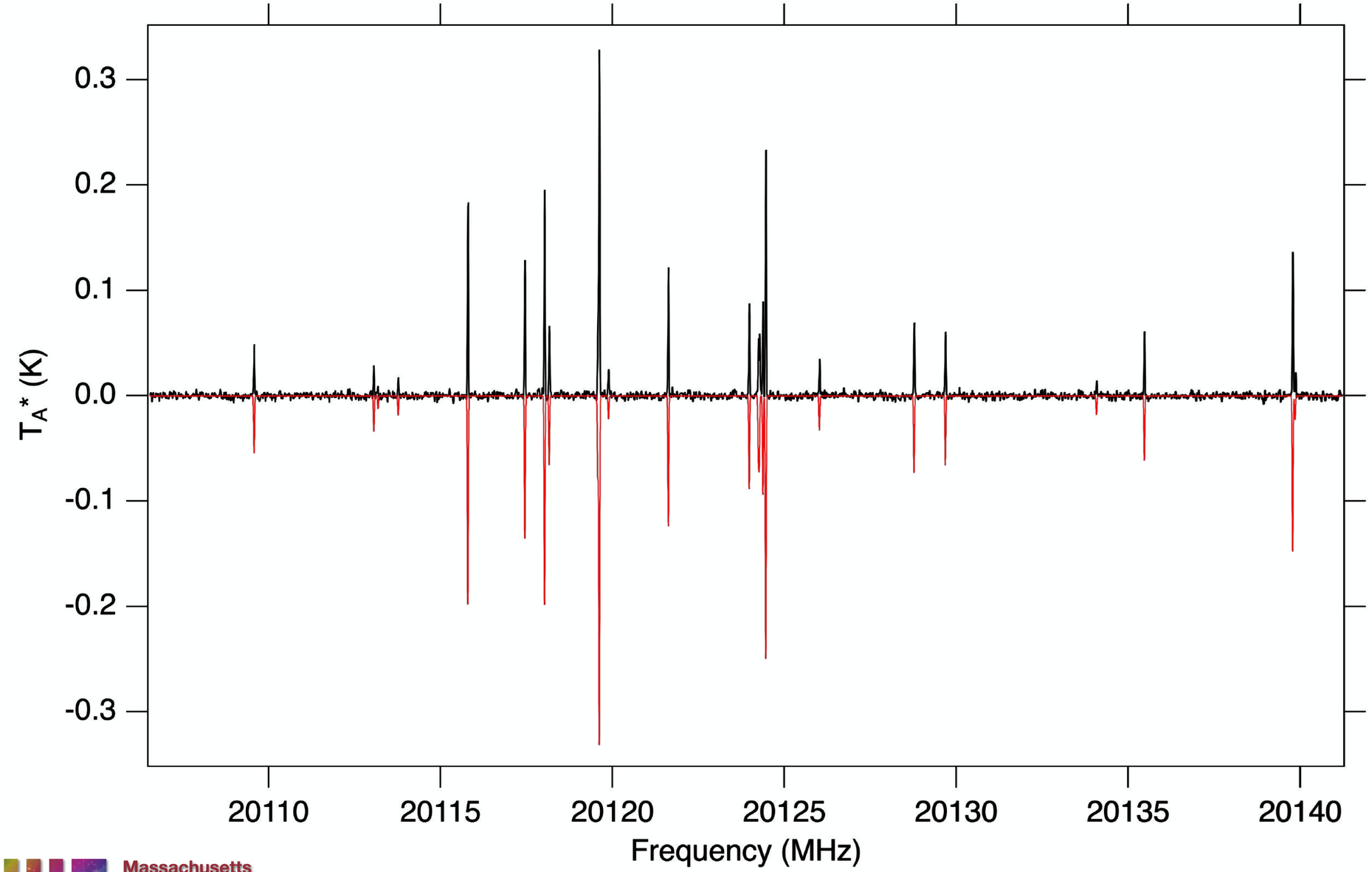




Mixture Analysis is Hard Work



Mixture Analysis is Hard Work

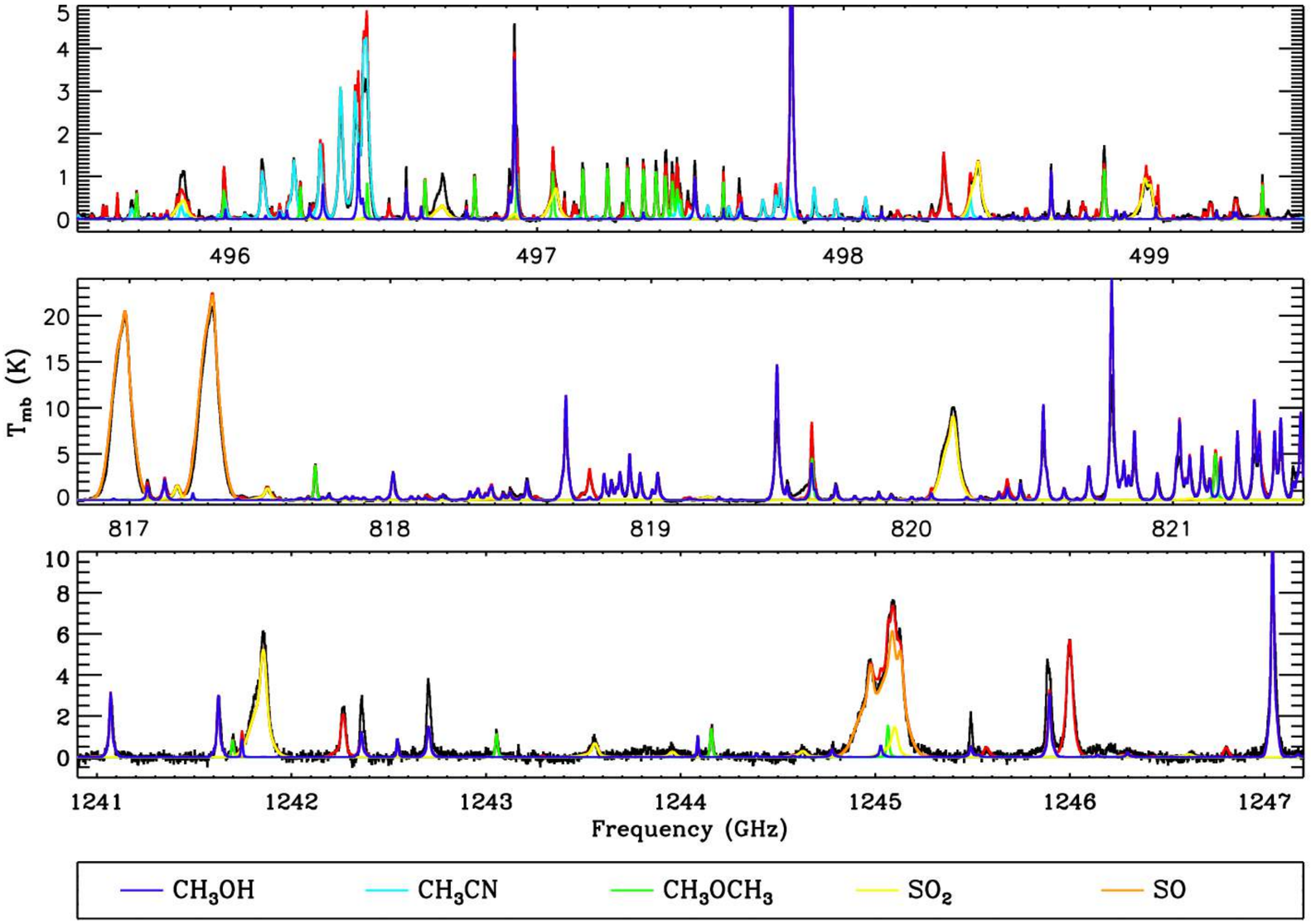


TMC-1
GBT Obs.

CH_2CN
Lab Freqs.

Mixture Analysis is Hard Work

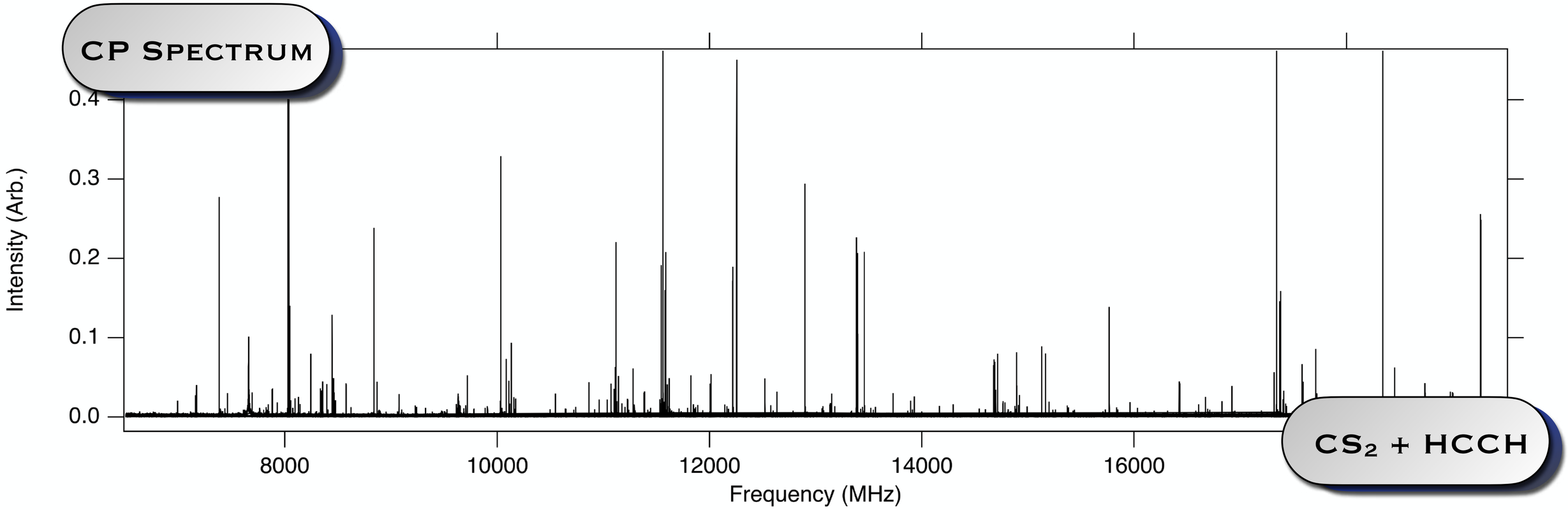
Crockett et al. 2014 *ApJ* 787, 112



Took a team of about a dozen people months to identify the molecules, velocity components, and physical parameters

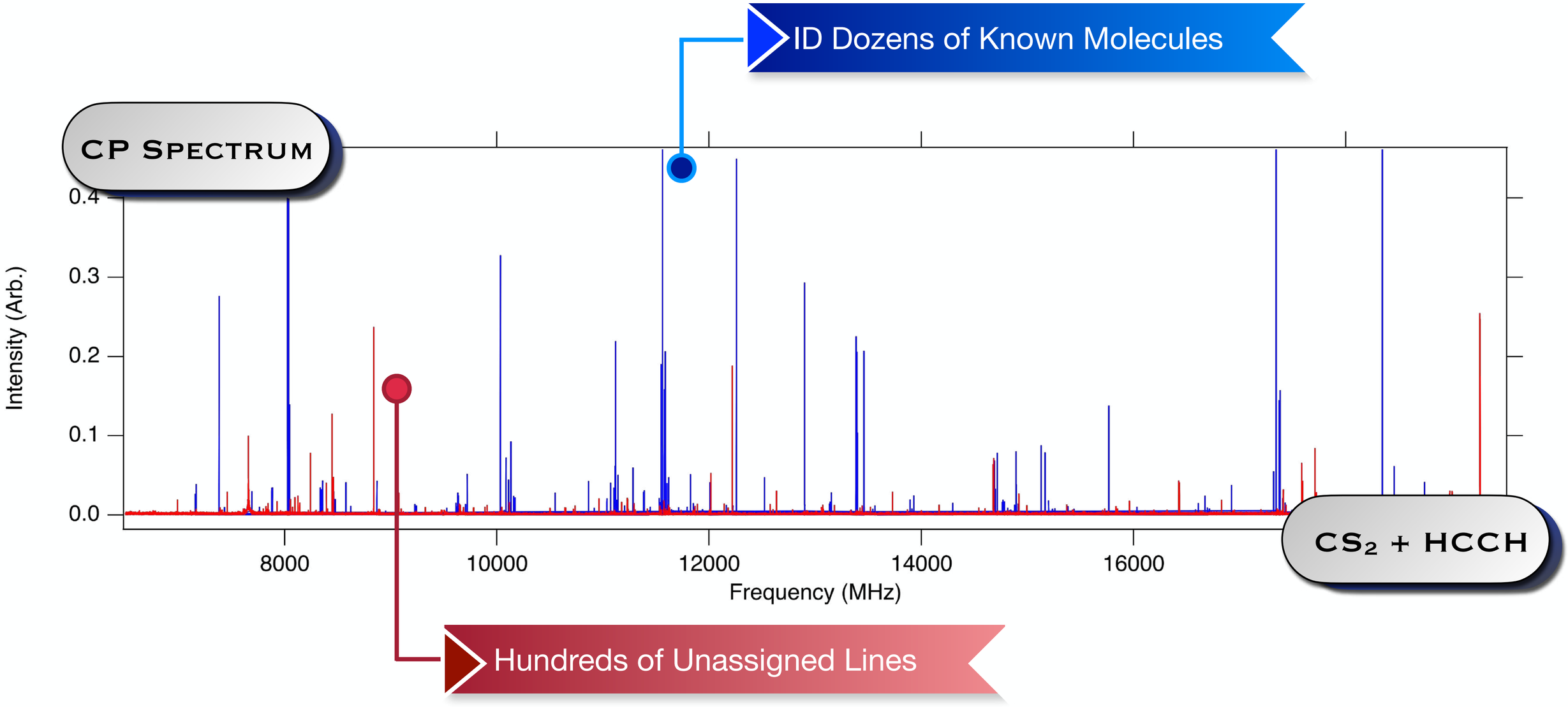
Mixture Analysis is Hard Work

McGuire et al. 2018 *PCCP* 20, 13870
Crabtree et al. 2016 *JCP* 144, 124201
Martin-Drumel et al. 2016 *JCP* 144, 124202



Mixture Analysis is Hard Work

McGuire et al. 2018 *PCCP* 20, 13870
Crabtree et al. 2016 *JCP* 144, 124201
Martin-Drumel et al. 2016 *JCP* 144, 124202

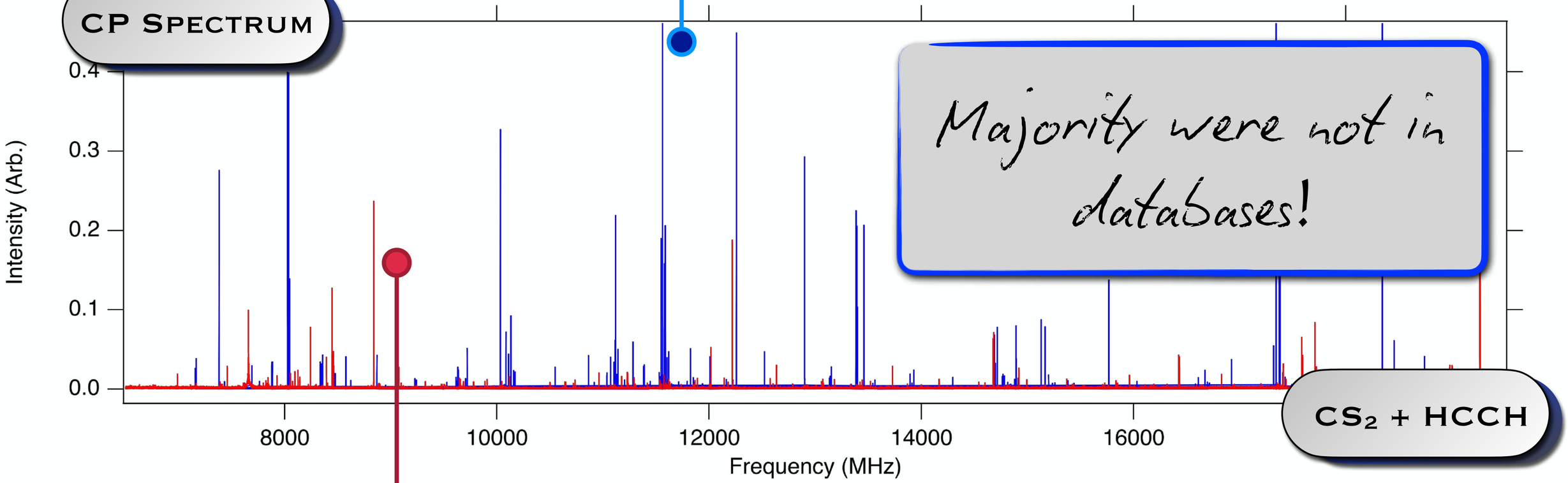


Mixture Analysis is Hard Work

McGuire et al. 2018 *PCCP* 20, 13870
Crabtree et al. 2016 *JCP* 144, 124201
Martin-Drumel et al. 2016 *JCP* 144, 124202

ID Dozens of Known Molecules

CP Spectrum



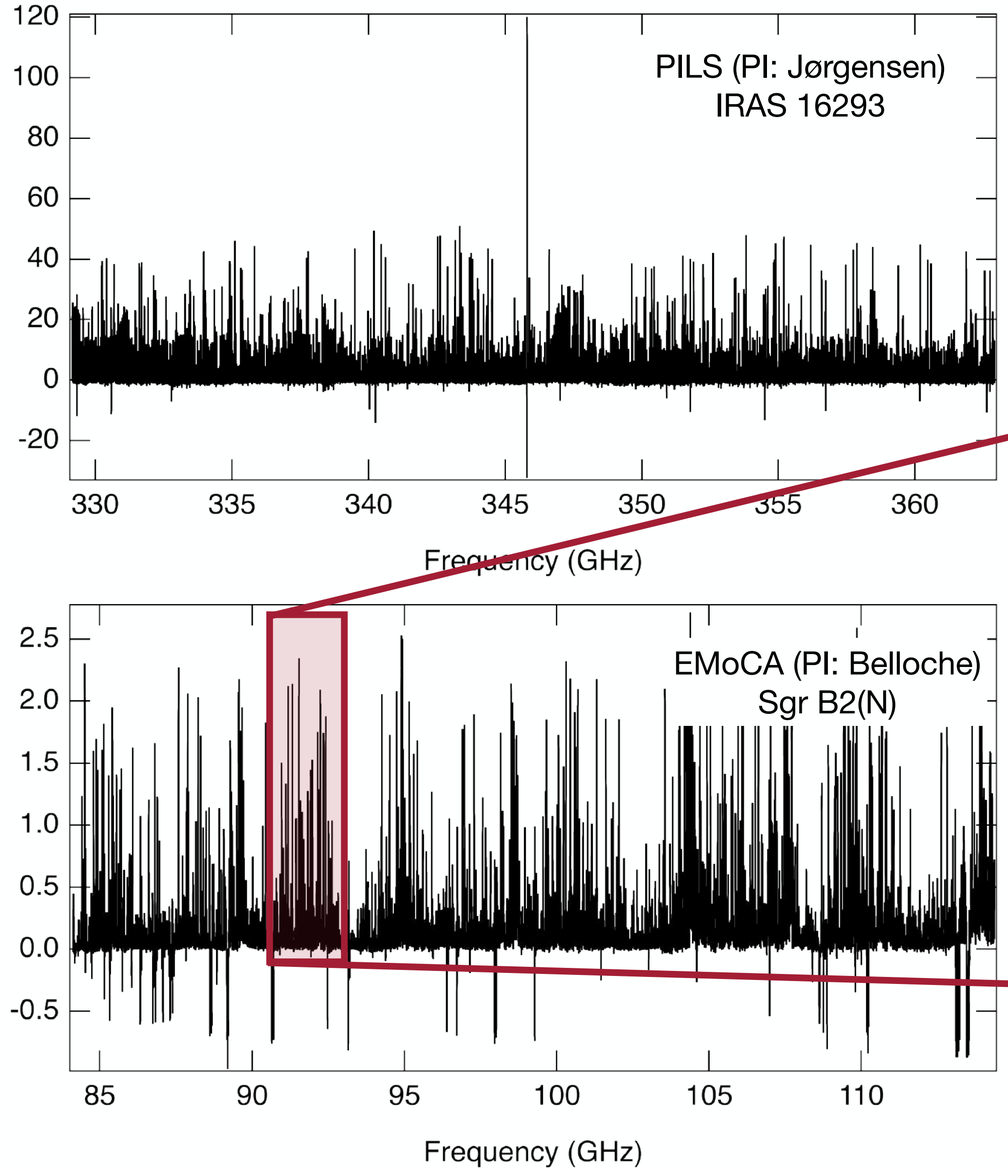
Majority were not in databases!

CS_2 + HCCH

Hundreds of Unassigned Lines

SPECTRAL LINE SURVEYS: FROM RARE → COMMON

Jørgensen et al. 2016 *A&A* 595, A117
Belloche et al. 2014 *Science* 345, 1584

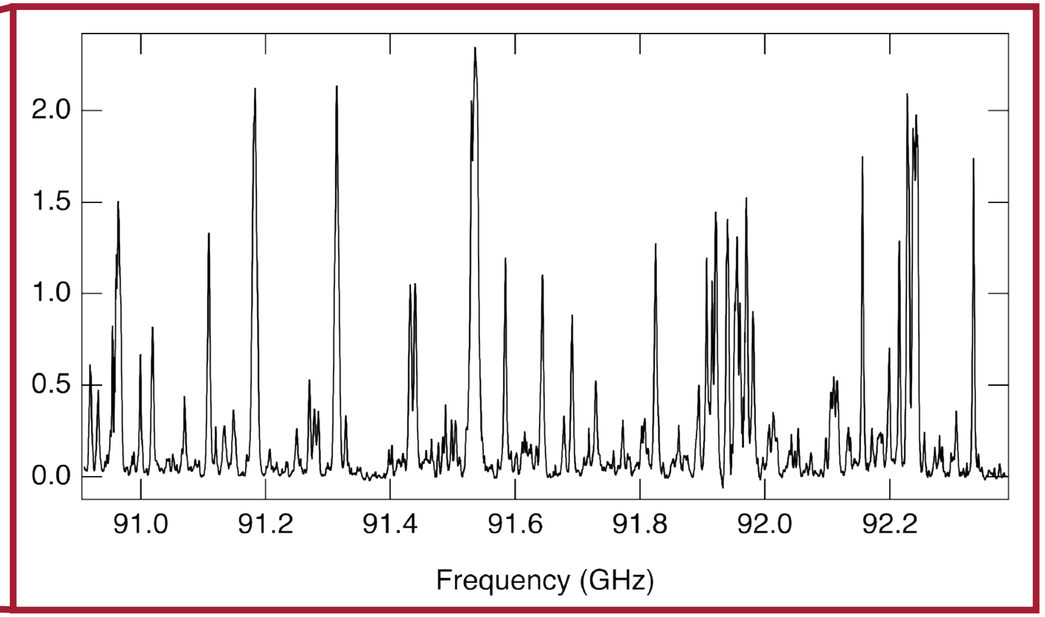


Broadband Spectral Line Surveys used to be …

… a few dozen GHz

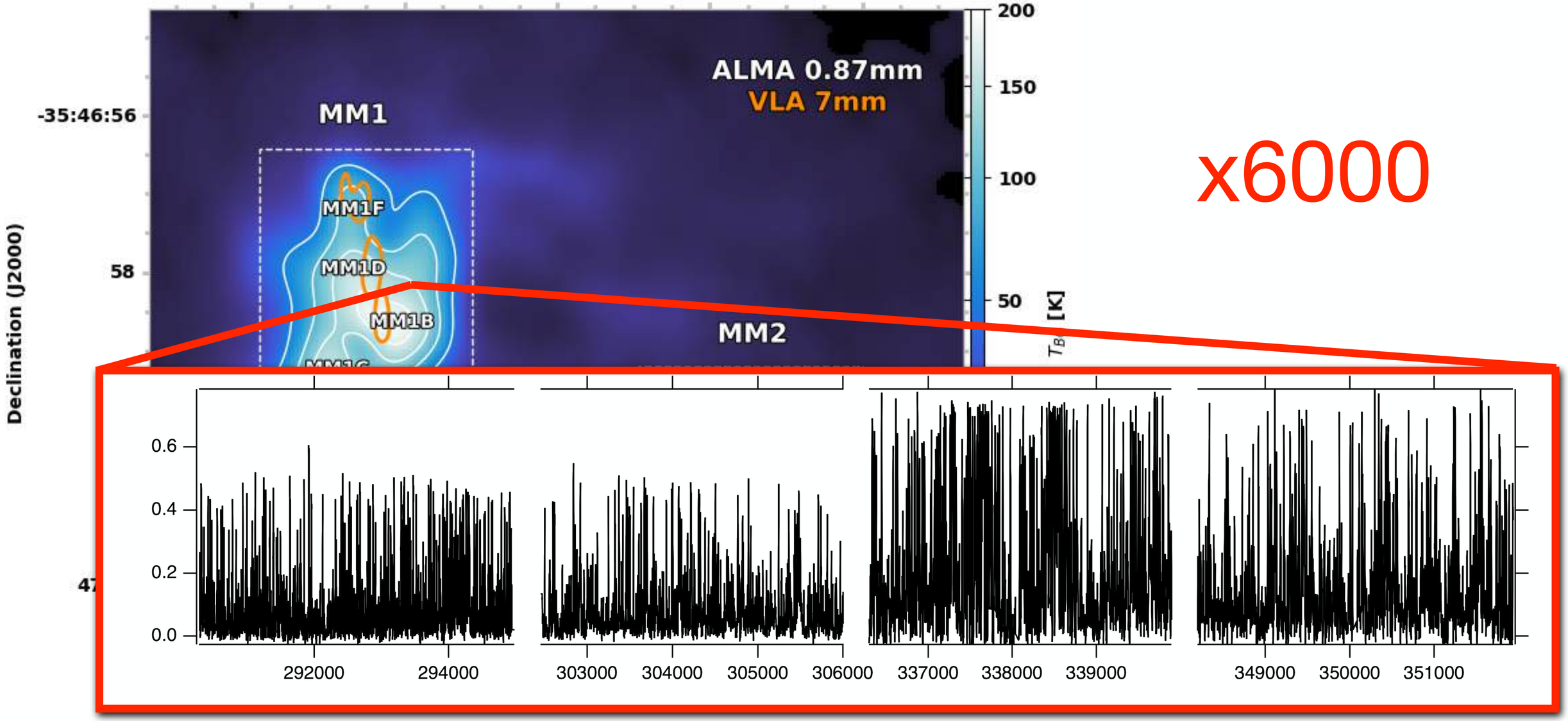
… analyzed toward a few pixels

… an ALMA Large Project



SPECTRAL LINE SURVEYS: FROM RARE → COMMON

El-Abd et al. 2024 *ApJ* 965, 14



SPECTRAL LINE SURVEYS: FROM MODEST → AHHHHHHHH!

The ALMA 2030 Development Roadmap, a Board-approved vision for ALMA sustainability and enhancement, identified the WSU initiatives as top priority. The goals of the WSU include an increase in bandwidth by a factor of 2 (ultimately x4), as well as increased digital sensitivity and improved sensitivity of key receivers.

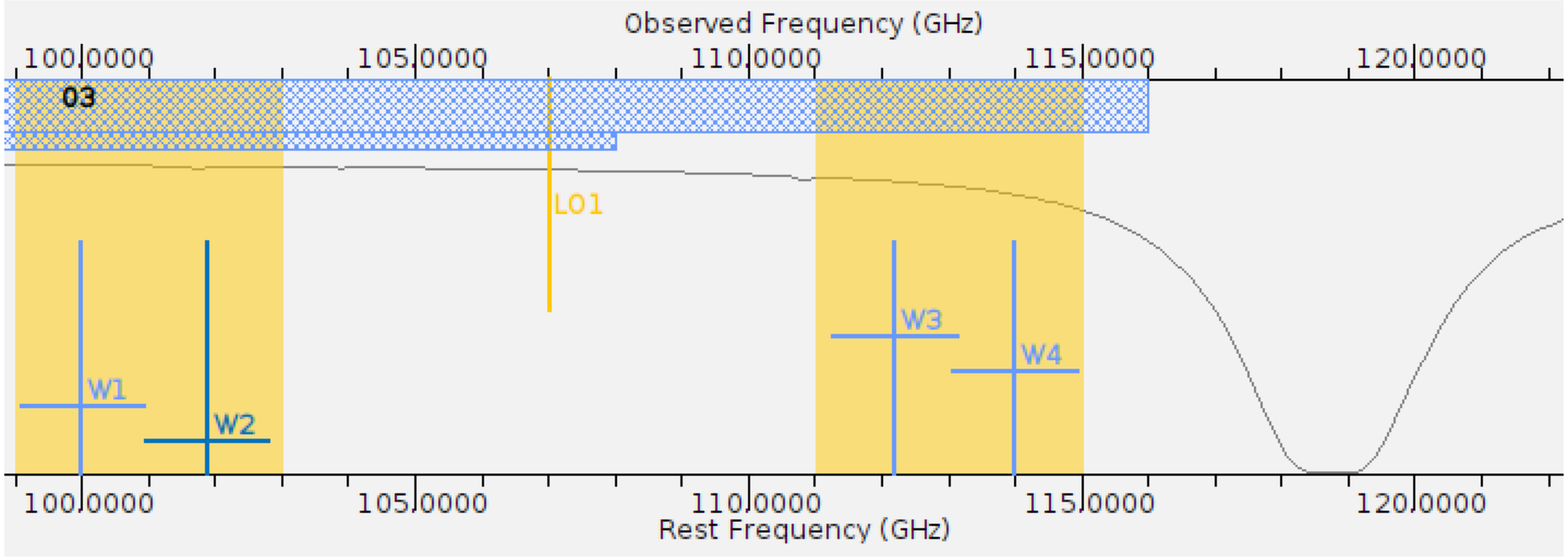


SPECTRAL LINE SURVEYS: FROM MODEST → AHHHHHHHH!

ALMA in 2025

Can set up 4x 1875 MHz basebands

So ~7.5 GHz total bandwidth … at 3 km/s spectral resolution!

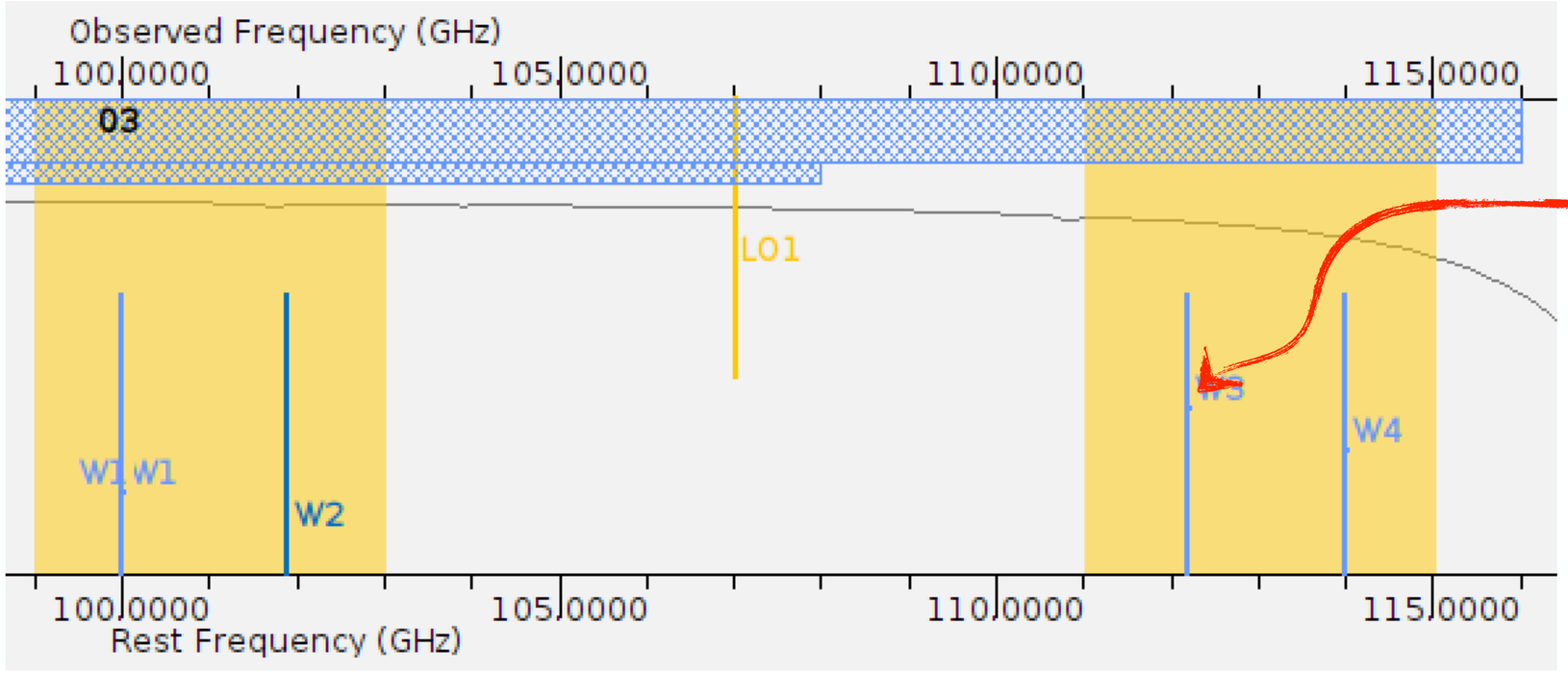


SPECTRAL LINE SURVEYS: FROM MODEST → AHHHHHHHH!

ALMA in 2025

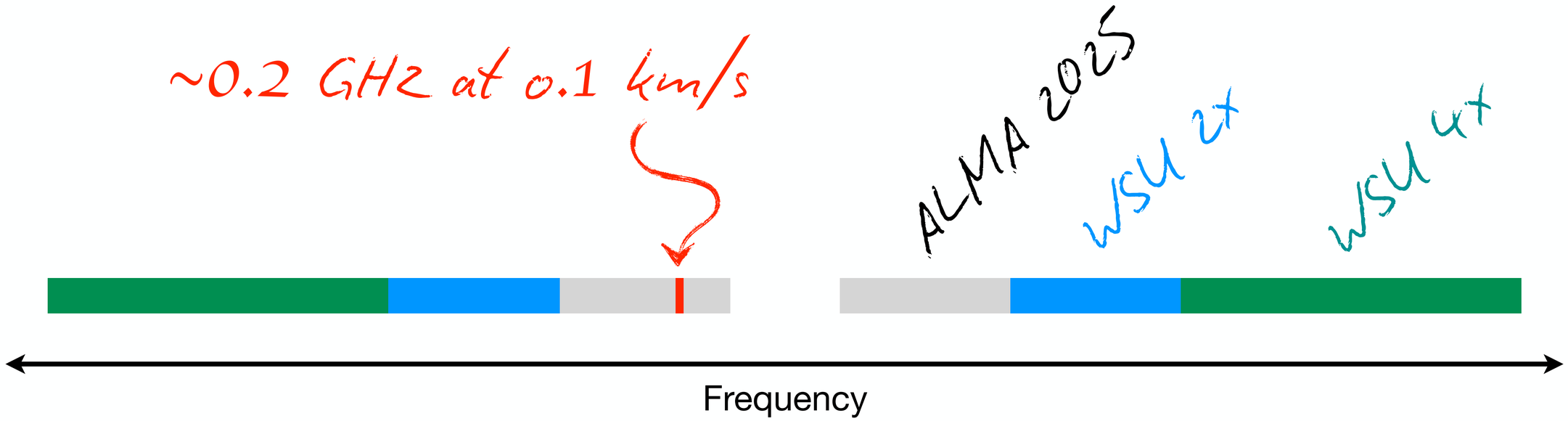
ALMA has only a few thousand spectral channels.

To get 0.1 km/s resolution, you must sacrifice bandwidth … 97% of it!!!



58 MHz windows to get 0.1 km/s

SPECTRAL LINE SURVEYS: FROM MODEST → AHHHHHHHH!



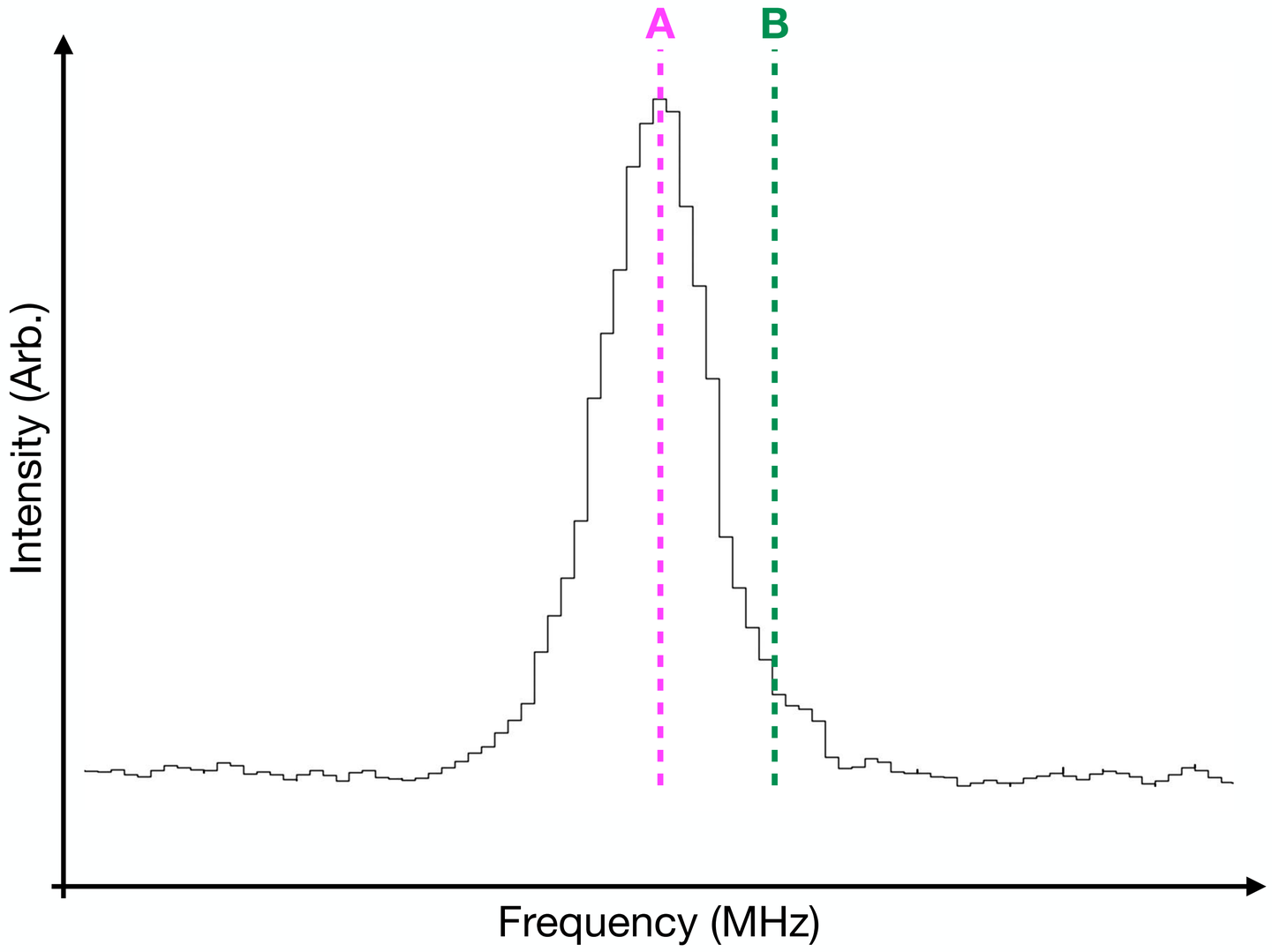
For spectral line observations, the WSU will provide not 2-4x, but 16-32x (or more!) the spectral grasp

Spectral Line Surveys with WSU / ngVLA / SKA



Picking Apart a Spectrum

Fried & McGuire 2024 *JPCA* 128, 8254

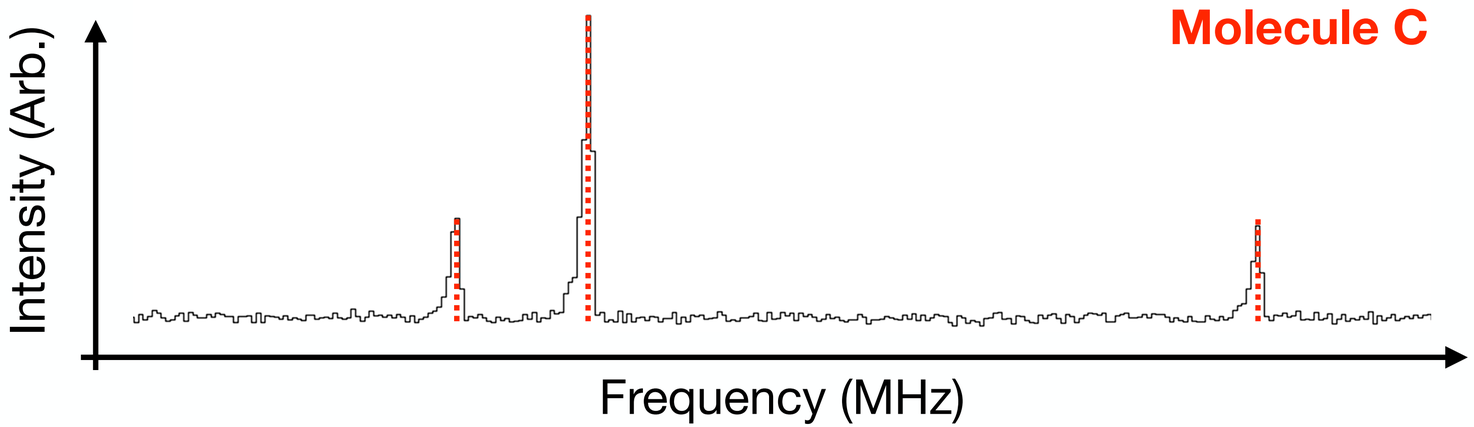


Go line-by-line and check for closest match in database

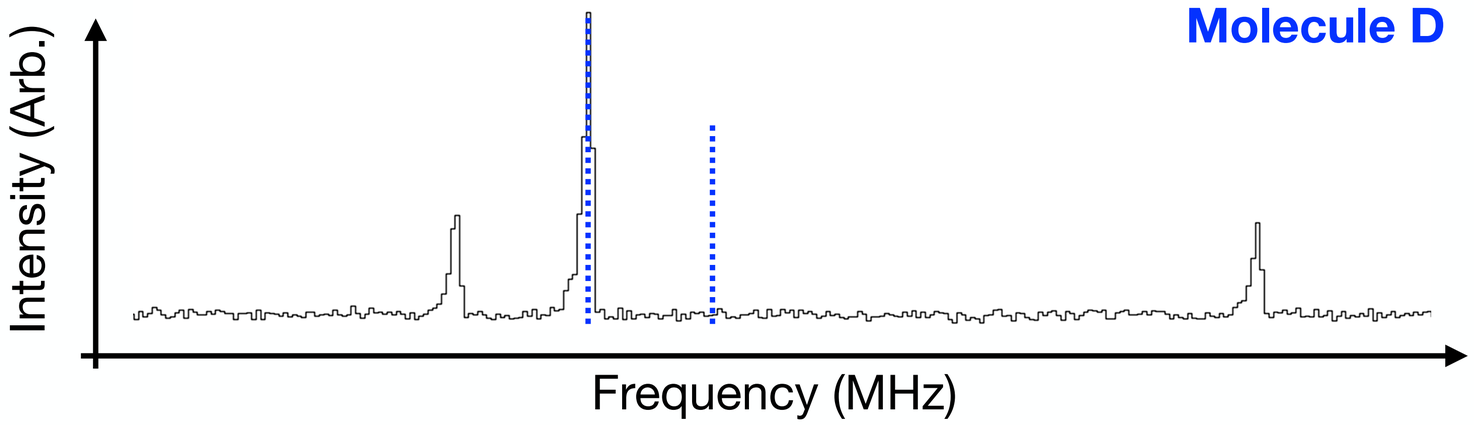
Trivial conventional computational problem

Picking Apart a Spectrum

Fried & McGuire 2024 *JPCA* 128, 8254



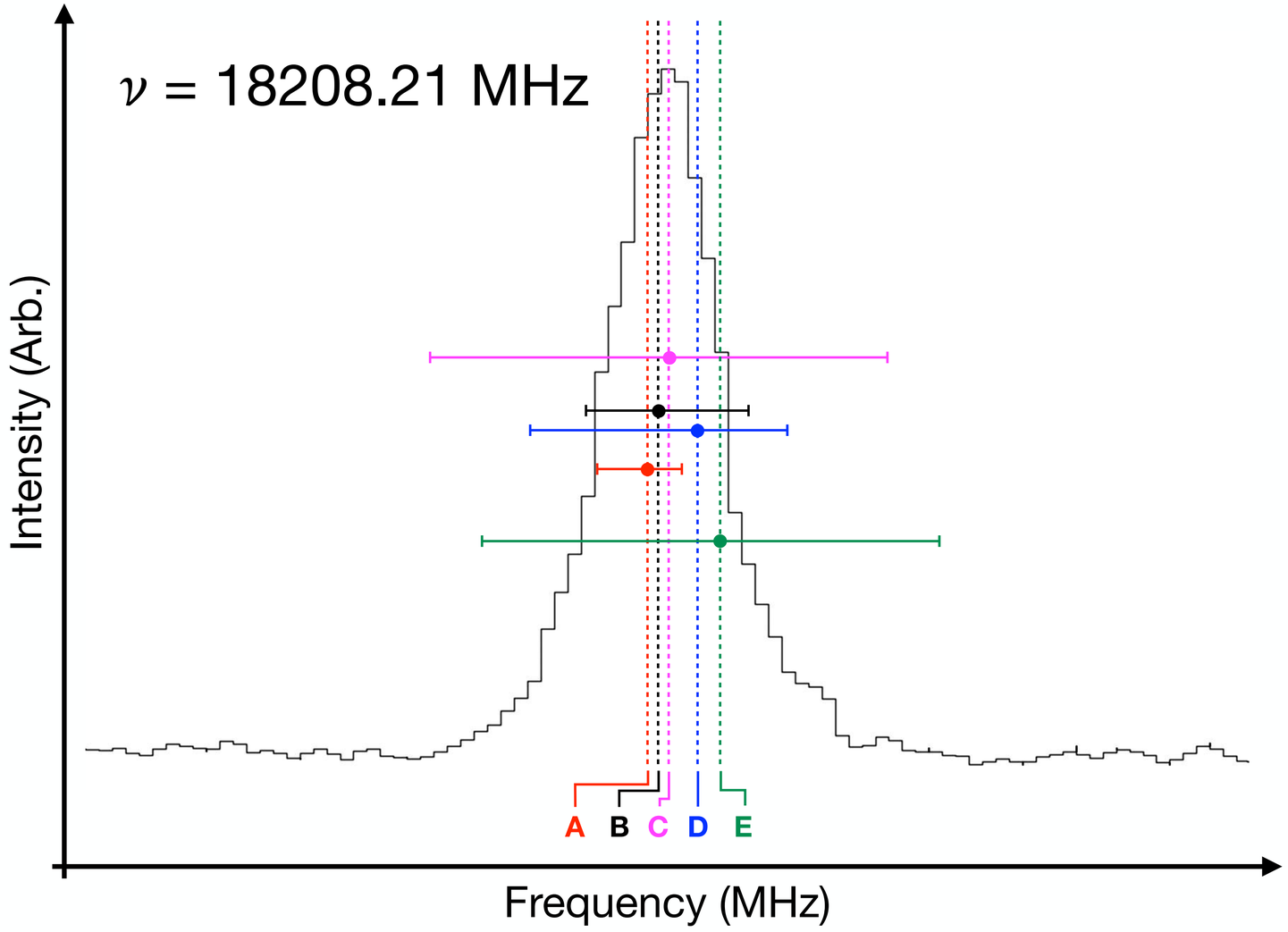
Check for other confirming lines in the spectrum

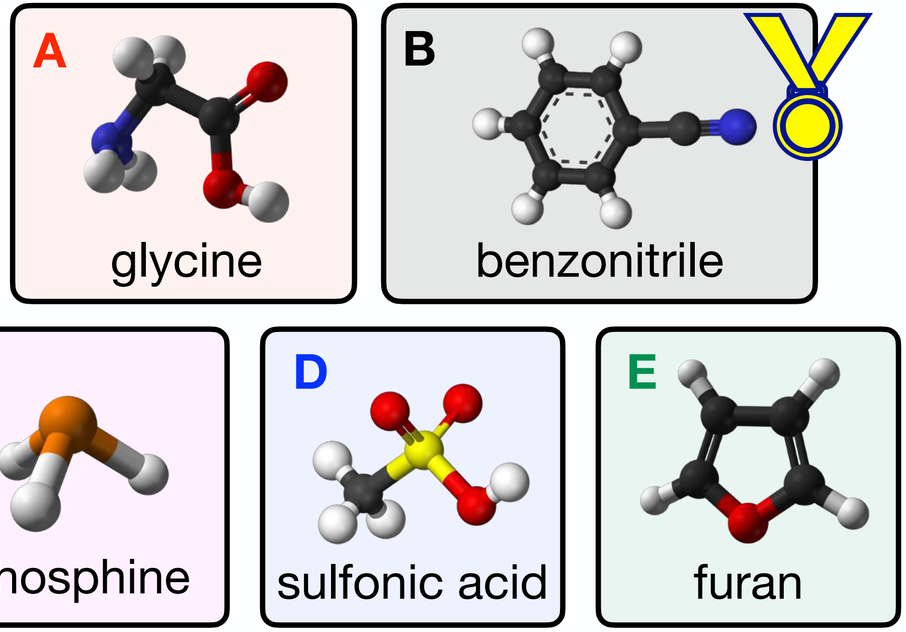


~trivial conventional computational problem

A Practical Example

Fried & McGuire 2024 *JPCA* 128, 8254

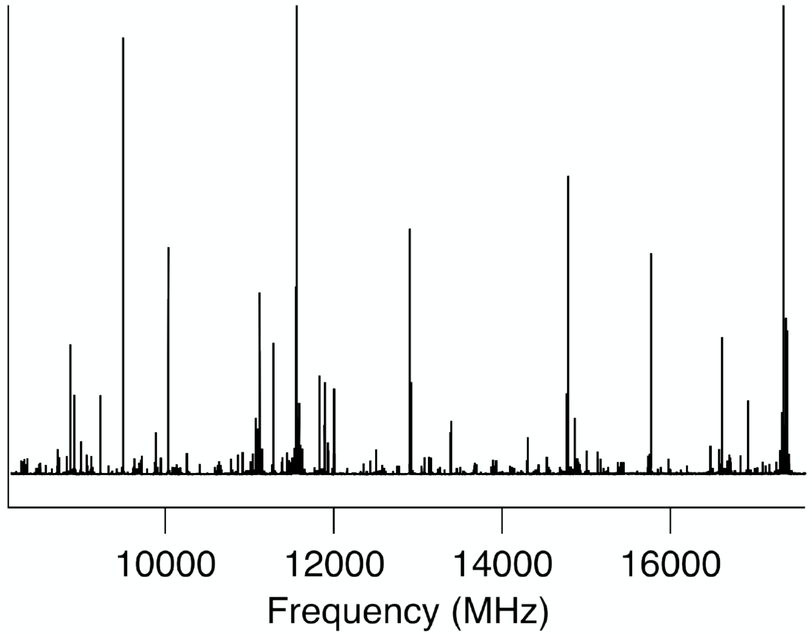




Candidate	Catalog - Exp.
A	0.201 MHz
B	0.121 MHz
C	0.020 MHz
D	0.356 MHz
E	0.549 MHz

AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ*, in prep

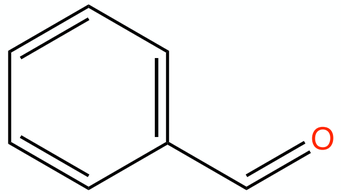


Line Survey

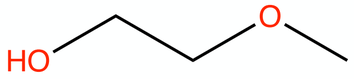
Peak Finder →

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

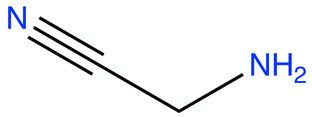
Database Query →



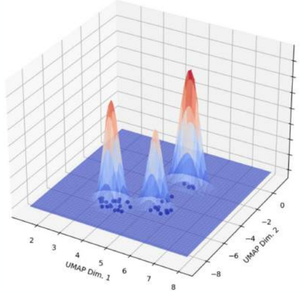
8314.92 MHz



8135.07 MHz



8135.14 MHz



Chemical Intuition Calculation

Assumption

Assumption

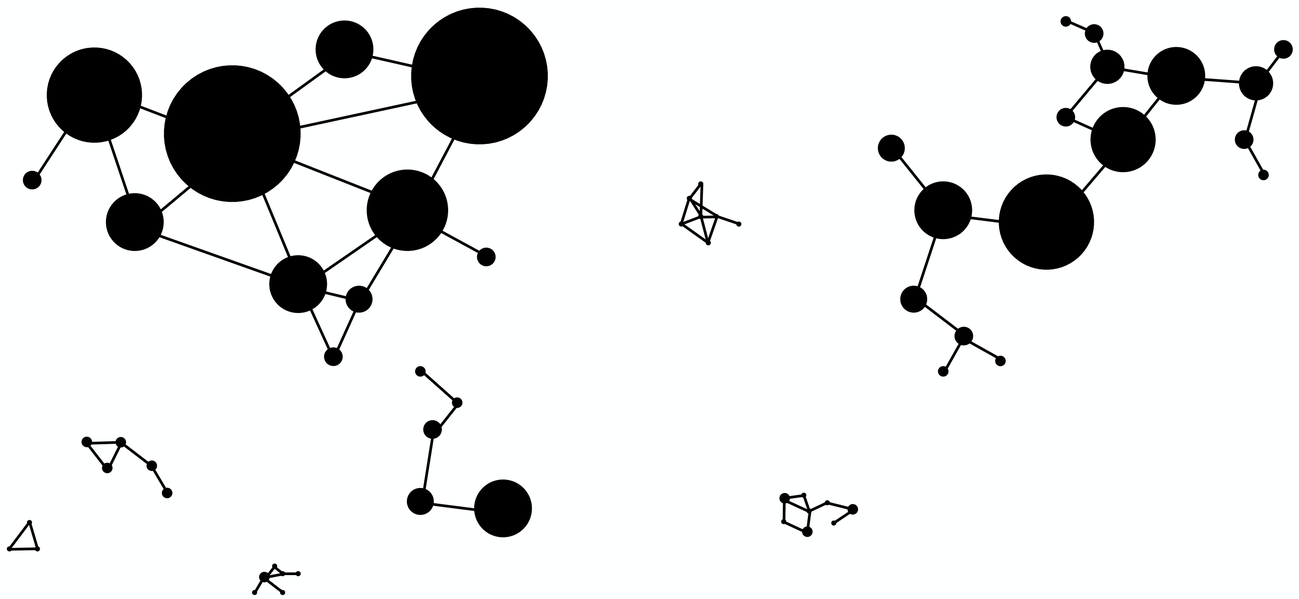
Molecules in a mixture are chemically related

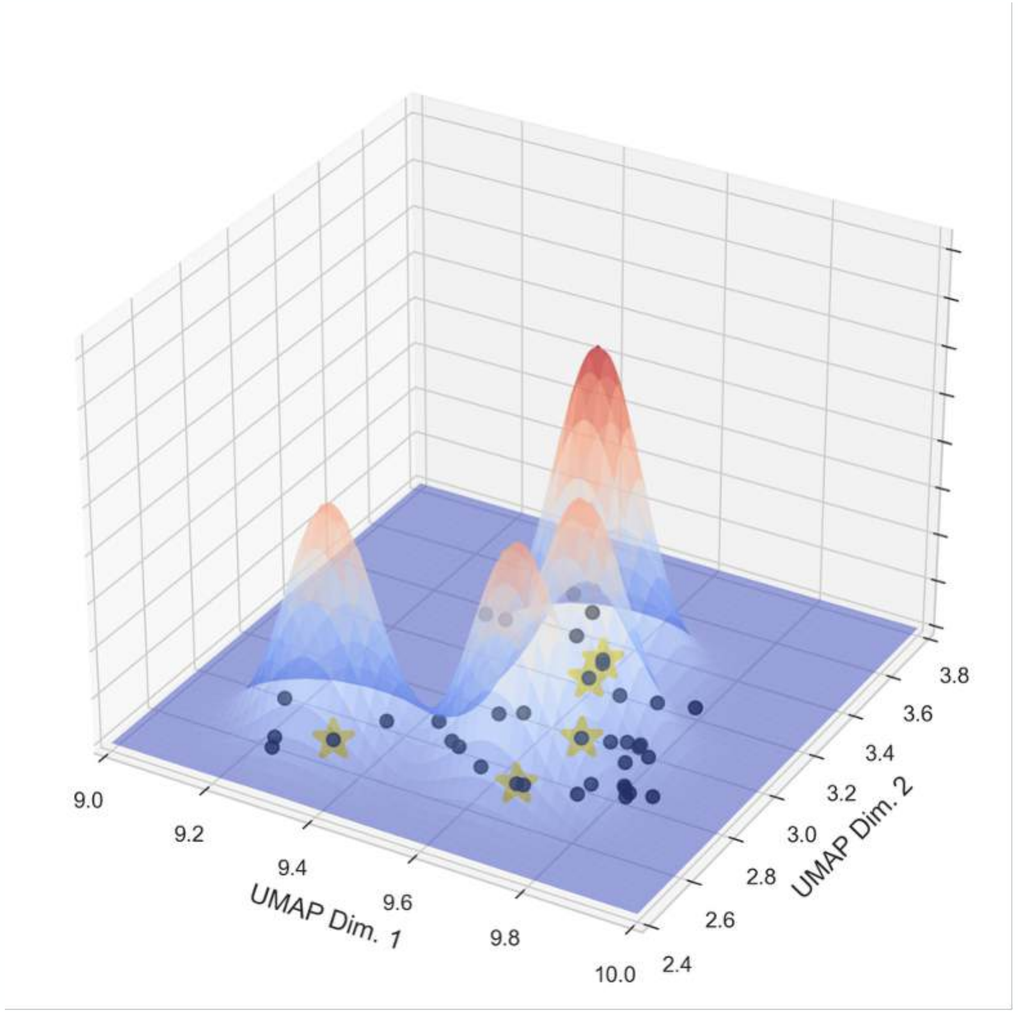


Spectral Surveys - Automated Assignment

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

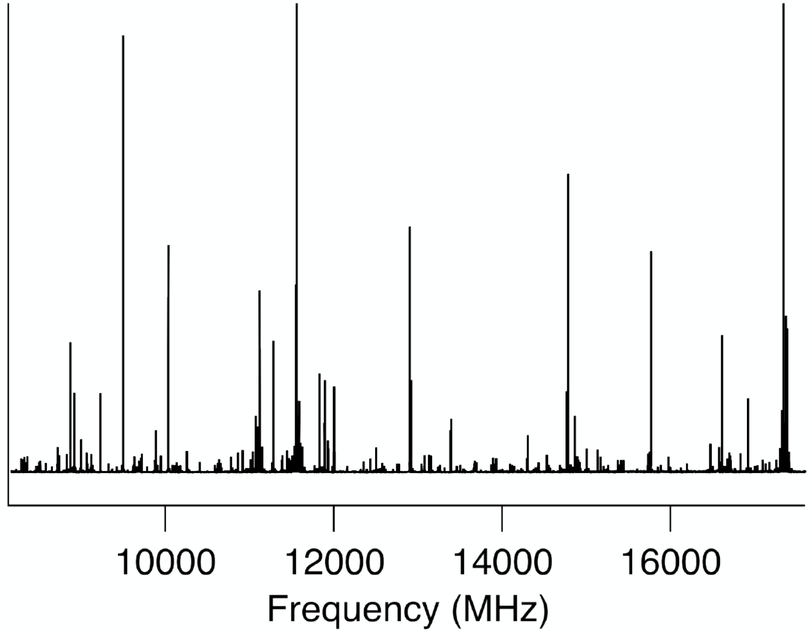
Use the positions of molecules in chemical vector space to upweight nearby (chemically similar) molecules and down-weight distant (chemically dissimilar) ones.





AUTOMATED MIXTURE ANALYSIS VIA STRUCTURAL EVALUATION

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ*, in prep

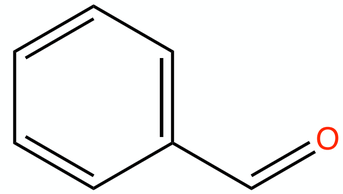


Line Survey

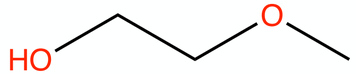
Peak Finder

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

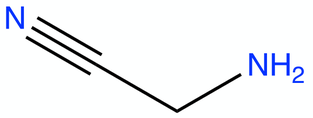
Database Query



8314.92 MHz

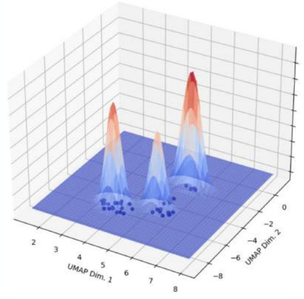


8135.07 MHz



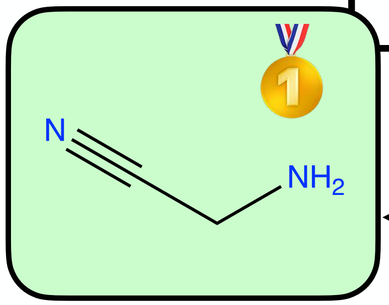
8135.14 MHz

Chemical Intuition Calculation



Scale Based on Frequency Match

Scale Based on Intensity Match



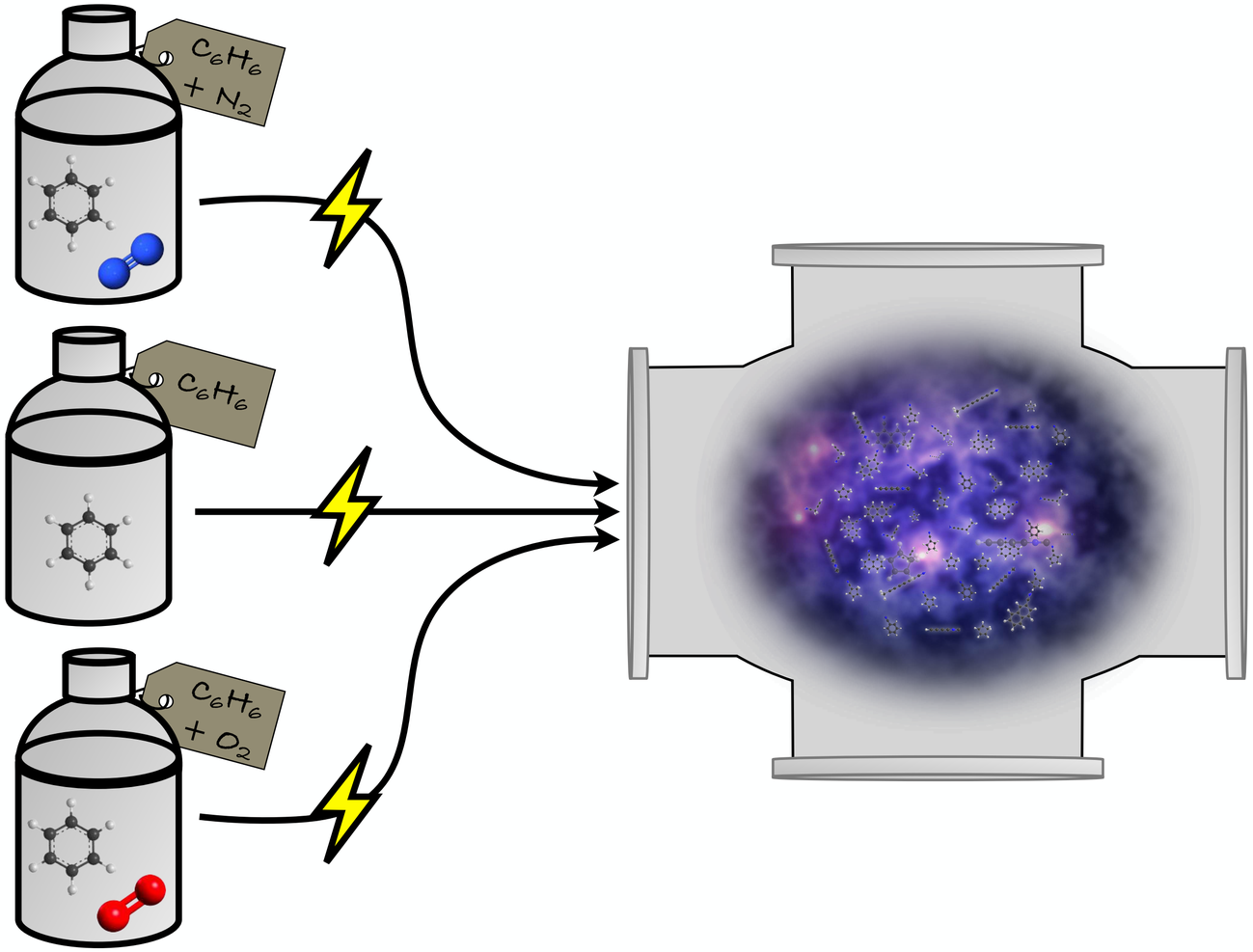
Assign Molecule with Top Score

Update Surface

Next Line

Mixture Analysis Was(?) Hard Work

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep



Testing on laboratory data …

~98% Accuracy vs Human

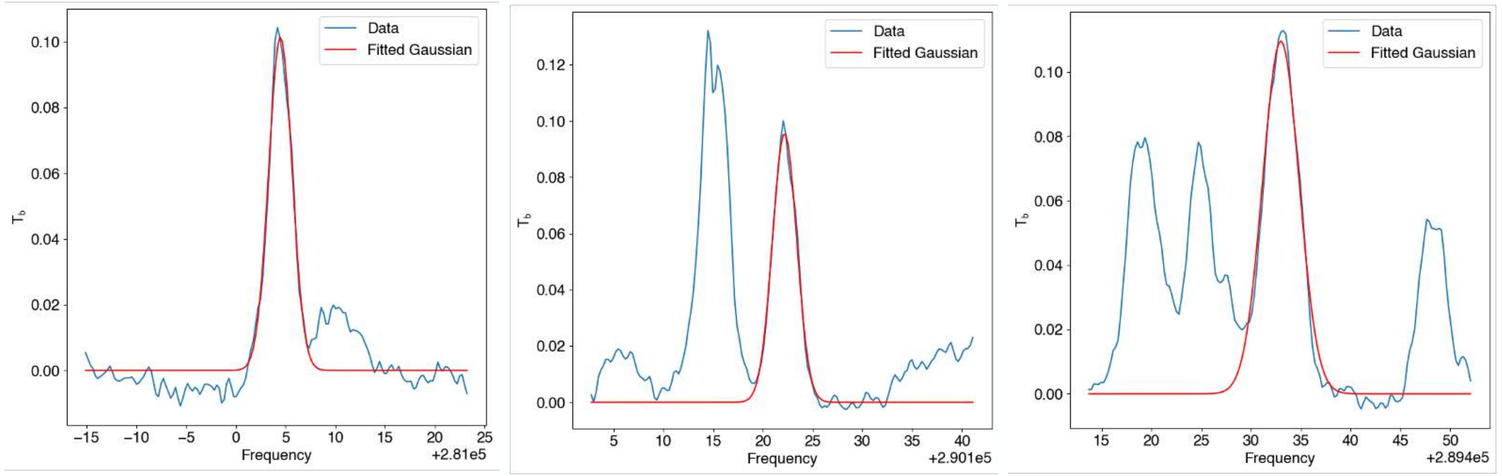
~15 minutes vs many days

Astronomical Data Present Challenges

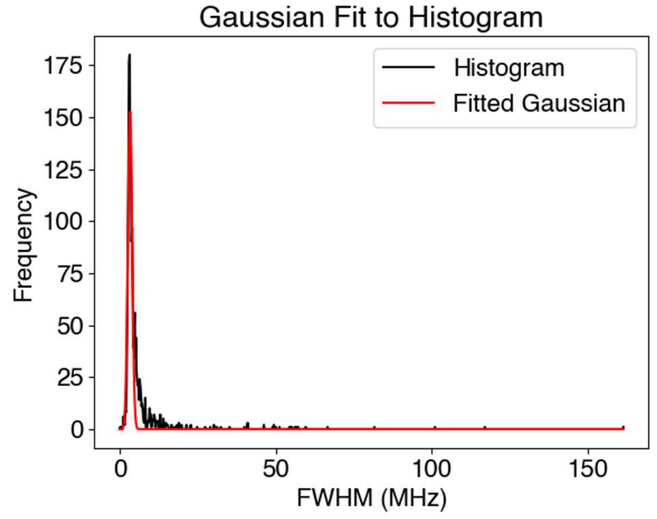
Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

Unknown *a priori* Linewidths

Fit Gaussians to All Lines



Fit Gaussian to Histogram of ΔVs



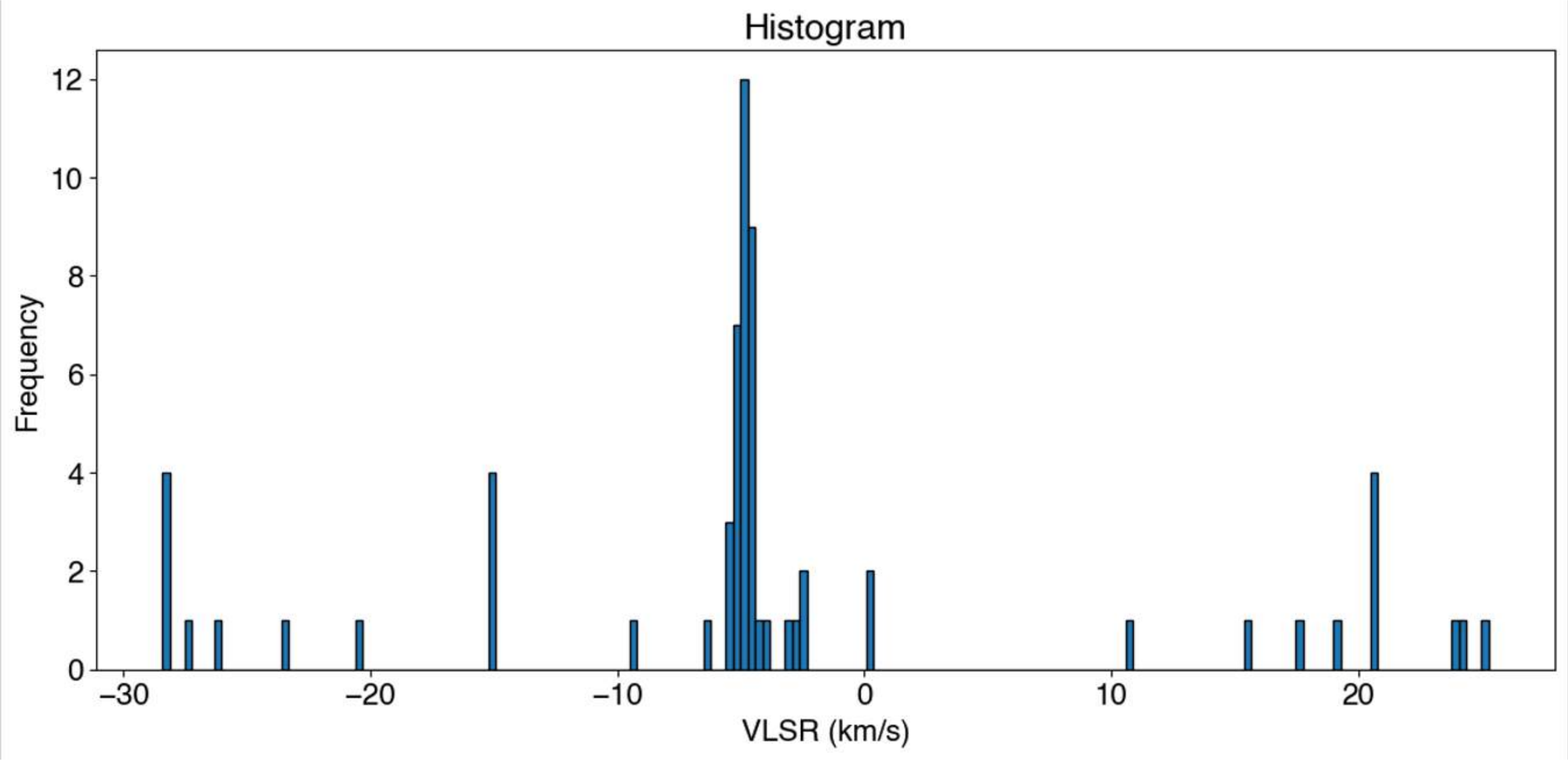
Astronomical Data Present Challenges

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

Unknown *a priori* v_{lsr}

Take a list of the 20 or so most common interstellar molecules

Determine v_{lsr} that would fit strong lines in observation

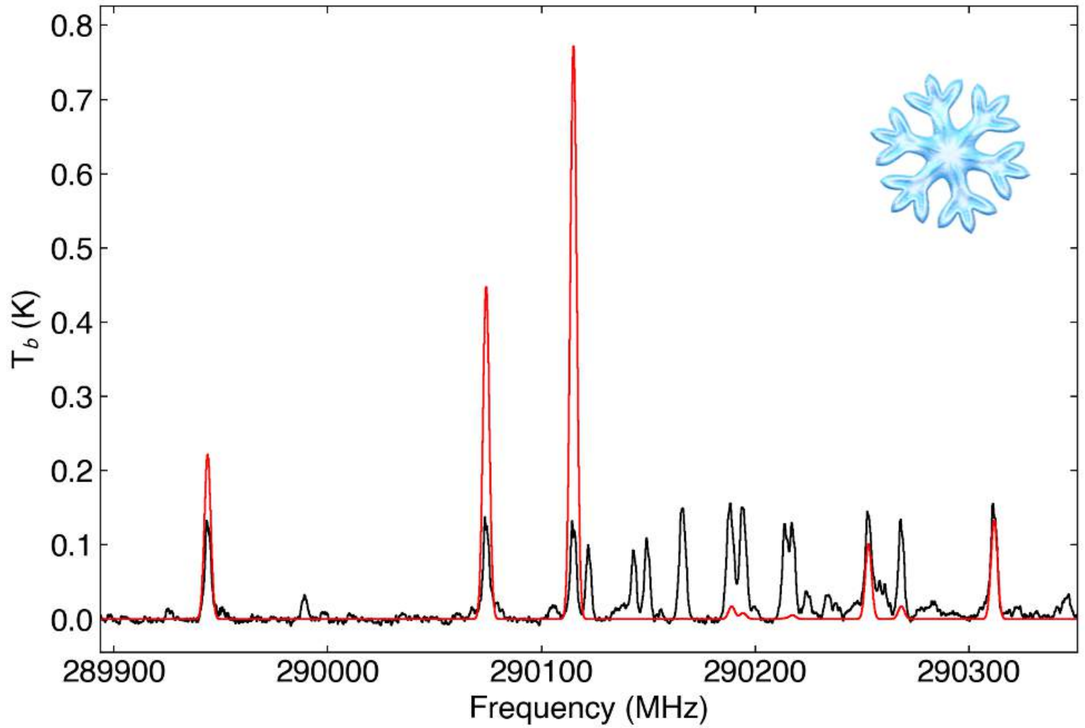


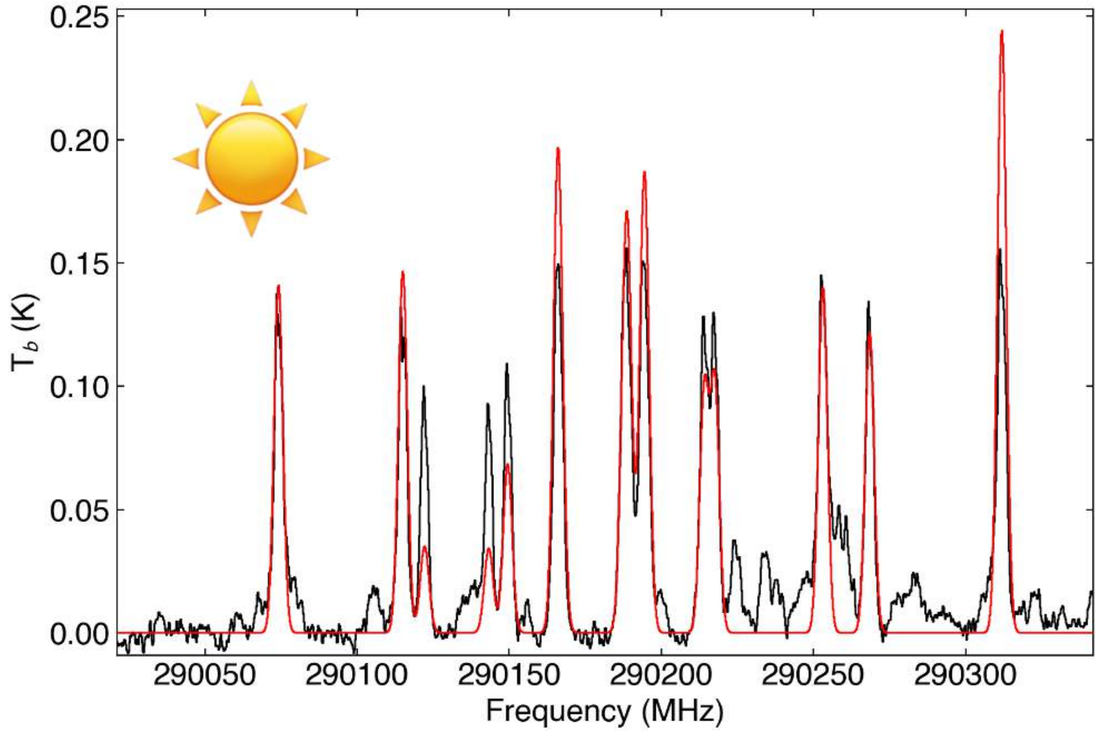
Astronomical Data Present Challenges

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

Unknown *a priori* Temperature

Grid search from 10-500 K for molecule with most lines in the v_{lsr} finder





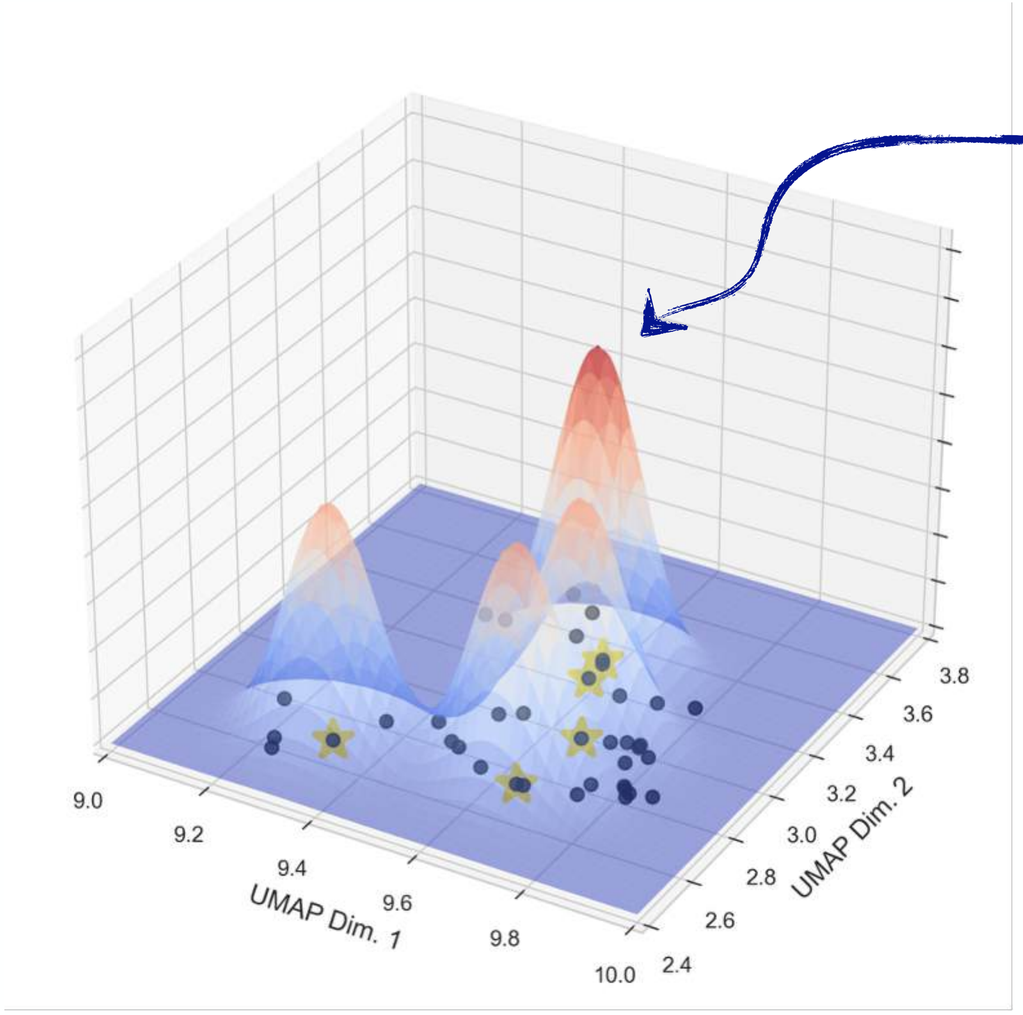
Spectral Survey Assignment

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

Live Demo!

Next-Best Target Predictions

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

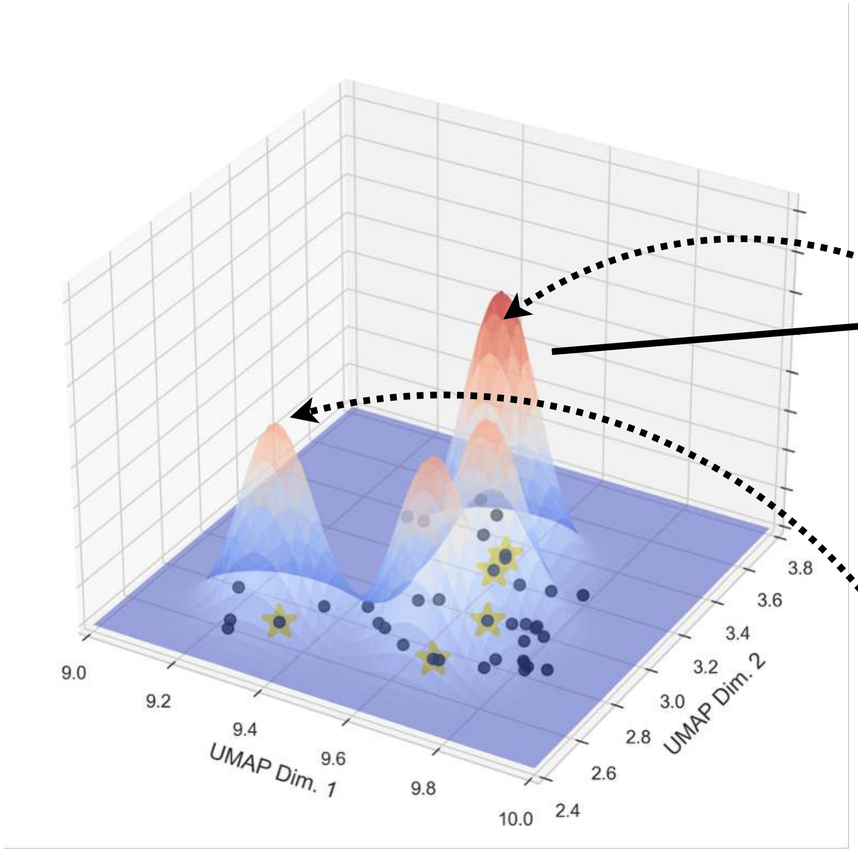


Highly weighted regions should contain good targets for study.

... but there isn't a direct pathway from vector -> chemical formula

NEXT-BEST TARGET PREDICTIONS

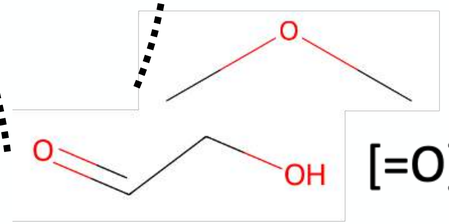
Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep



Determine likely chemical features

$\begin{bmatrix} 0.43 \\ 11.56 \end{bmatrix}$

$\begin{bmatrix} 1.0 \\ 1.0 \\ 0.0 \\ 0.0 \\ 1.0 \\ \vdots \end{bmatrix}$ $\begin{matrix} [C] \\ [\#N] \\ [O] \\ [=O] \\ [\#C] \end{matrix}$



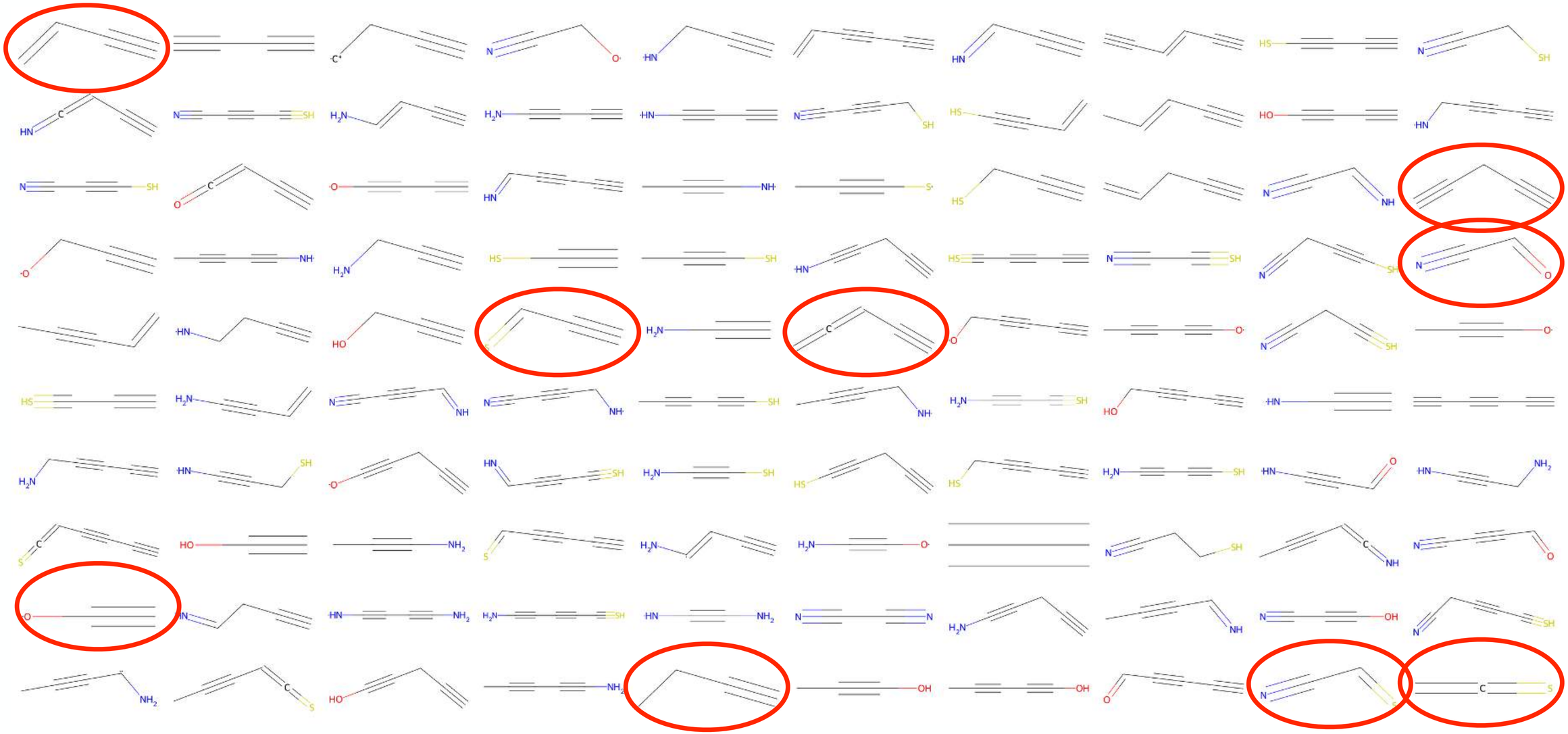
[C][O][C]

[=O][C][C][O]

Generate molecules by probabilistically sampling features

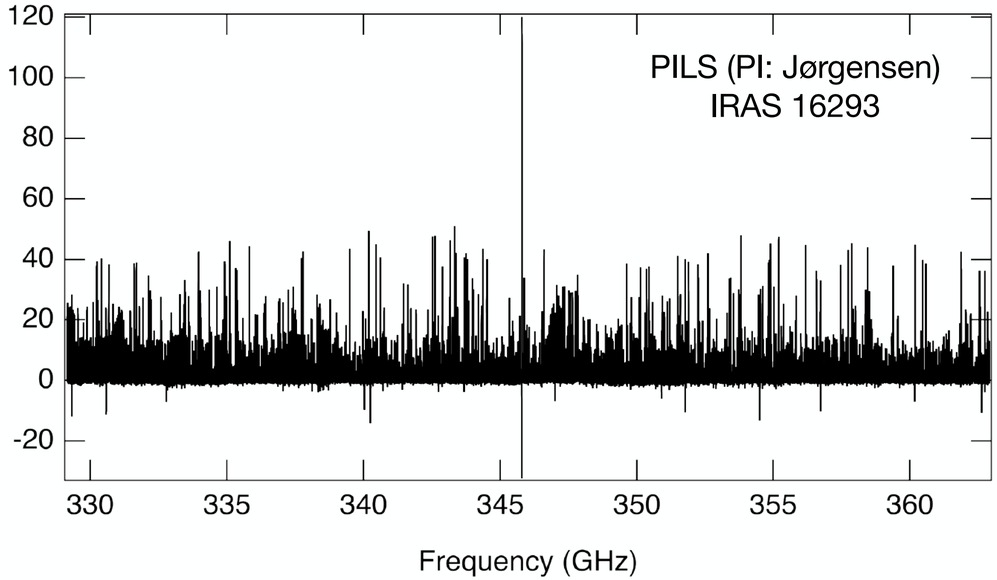
NEXT-BEST TARGET PREDICTIONS - TMC-1

Fried & McGuire 2024 *JPCA* 128, 8254
Fried et al. 2025 *ApJ,* in prep

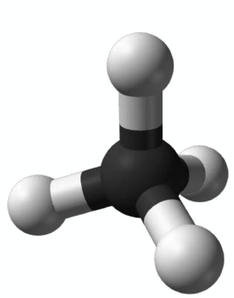


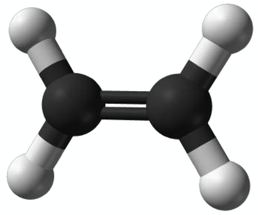
GENERATIVE RECOMMENDATIONS

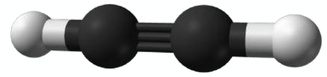
Fried et al. 2025 *ApJ*, in prep

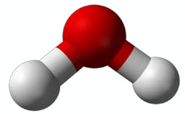


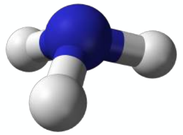
Using the published inventory of molecules in this survey, these other known ISM species are generated and highly ranked (Top 100)!

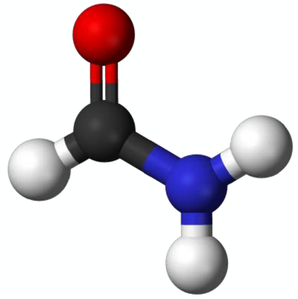

methane

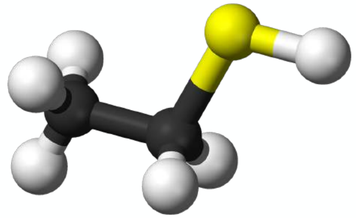

ethylene

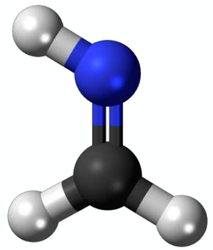

acetylene

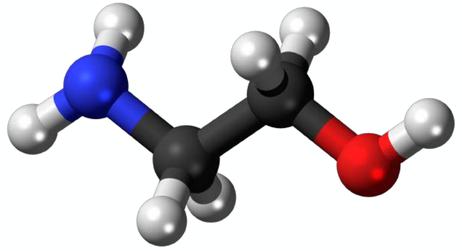

water

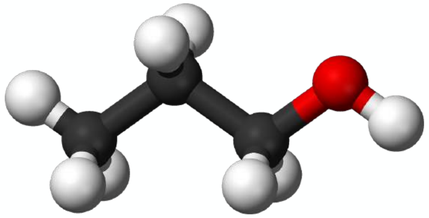

ammonia


formamide


ethanethiol


methanimine


ethanolamine


propanol



Part III

Physical Properties

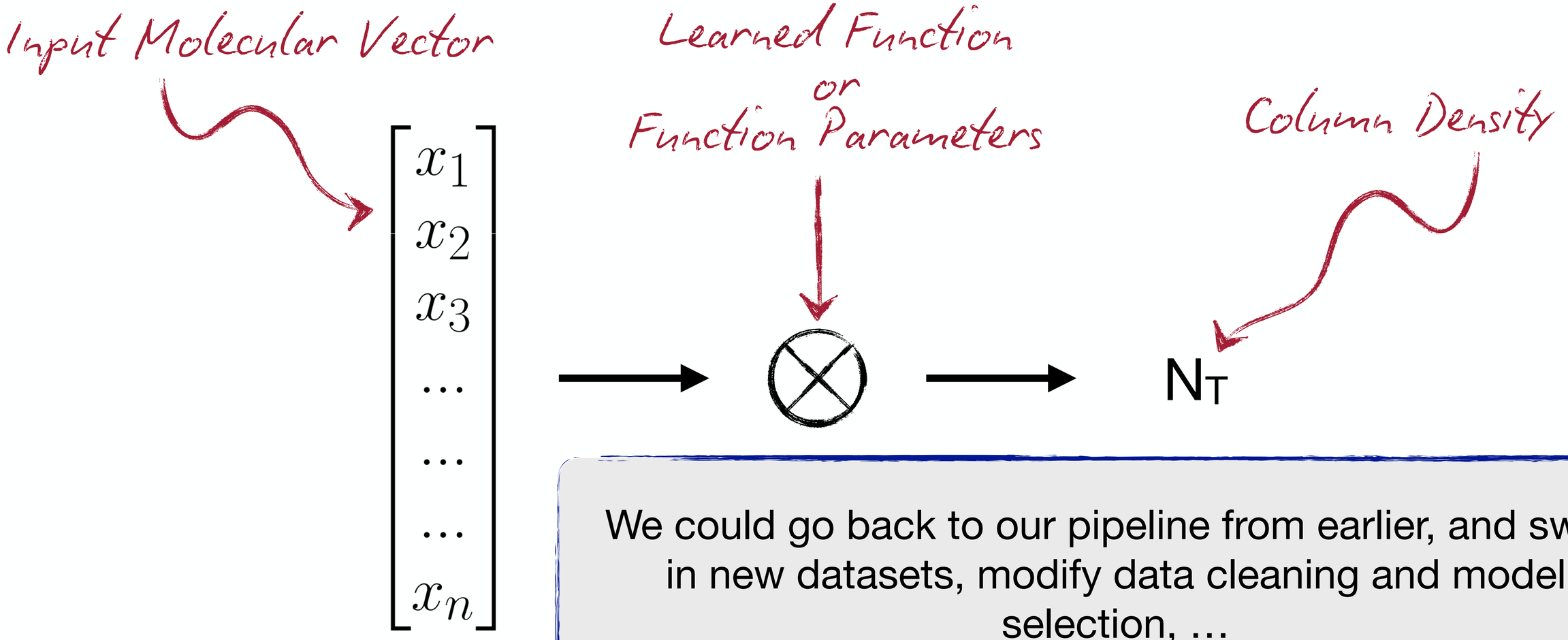


Machine Learning Pipeline for Molecular Property Prediction Using ChemXploreML

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424

Physical Property Predictions

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424



We could go back to our pipeline from earlier, and swap in new datasets, modify data cleaning and model selection, …

It would be great to have a more modular approach.

ChemXploreML

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424



Objective
(e.g. classification, regression)



Data Collection
Gather relevant raw data from diverse sources

Data Preprocessing



Data Exploration
Understand data characteristics and issues



Data Cleaning
Handle missing values, duplicates, and outliers



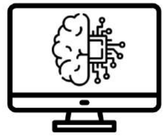
Data Transformation
Scale/normalize data, encode categorical features



Feature Engineering
Create new, more informative features from existing ones



Dimensionality Reduction
(e.g., PCA)

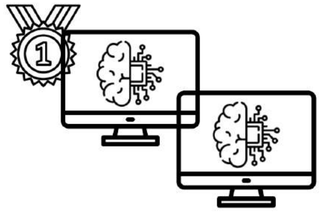


Model Selection
Linear Regression, GBR, CatBoost, XGBoost, LightGBM, etc.



Model Training
Train the model using the training data

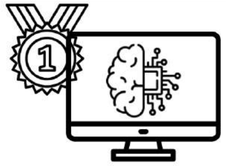
Model Evaluation



Hyperparameter Tuning
Optimize model parameters for best performing model



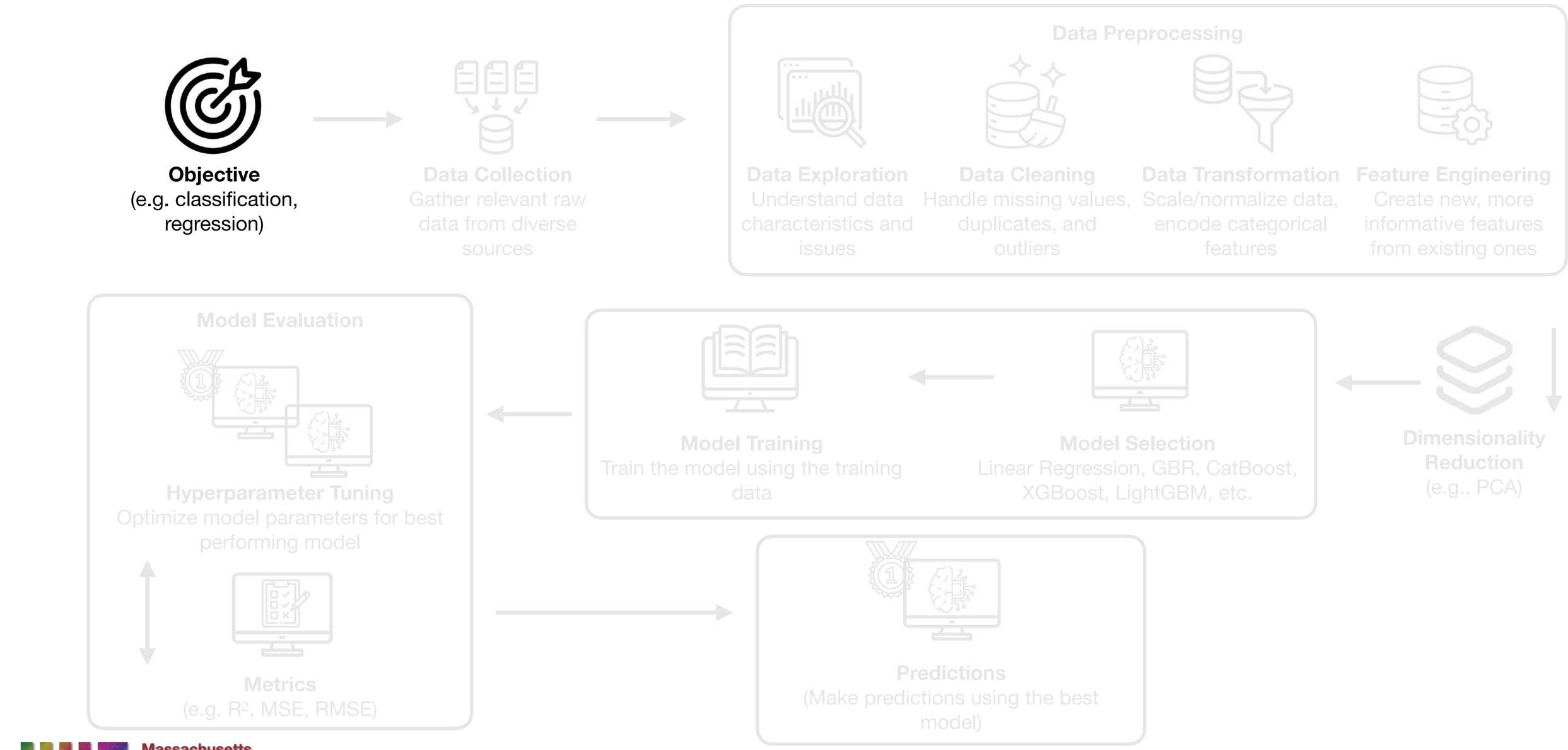
Metrics
(e.g. R^2, MSE, RMSE)



Predictions
(Make predictions using the best model)

CHEMXPLOREML

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424



CHEMXPLOREML

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424

Objective
(e.g. classification, regression)



Data Collection
Gather relevant raw data from diverse sources

Data Preprocessing

Data Exploration
Understand data characteristics and issues

Data Cleaning
Handle missing values, duplicates, and outliers

Data Transformation
Scale/normalize data, encode categorical features

Feature Engineering
Create new, more informative features from existing ones

Model Training
Train the model using the training data

Model Selection
Linear Regression, GBR, CatBoost, XGBoost, LightGBM, etc.

Dimensionality Reduction
(e.g., PCA)

Model Evaluation

Hyperparameter Tuning
Optimize model parameters for best performing model

Metrics
(e.g. R^2, MSE, RMSE)

Predictions
(Make predictions using the best model)

CHEMXPLOREML

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424

Objective
(e.g. classification, regression)

Data Collection
Gather relevant raw data from diverse sources

Data Preprocessing



Data Exploration
Understand data characteristics and issues



Data Cleaning
Handle missing values, duplicates, and outliers



Data Transformation
Scale/normalize data, encode categorical features



Feature Engineering
Create new, more informative features from existing ones

Model Training
Train the model using the training data

Model Selection
Linear Regression, GBR, CatBoost, XGBoost, LightGBM, etc.

Dimensionality Reduction
(e.g., PCA)

Model Evaluation

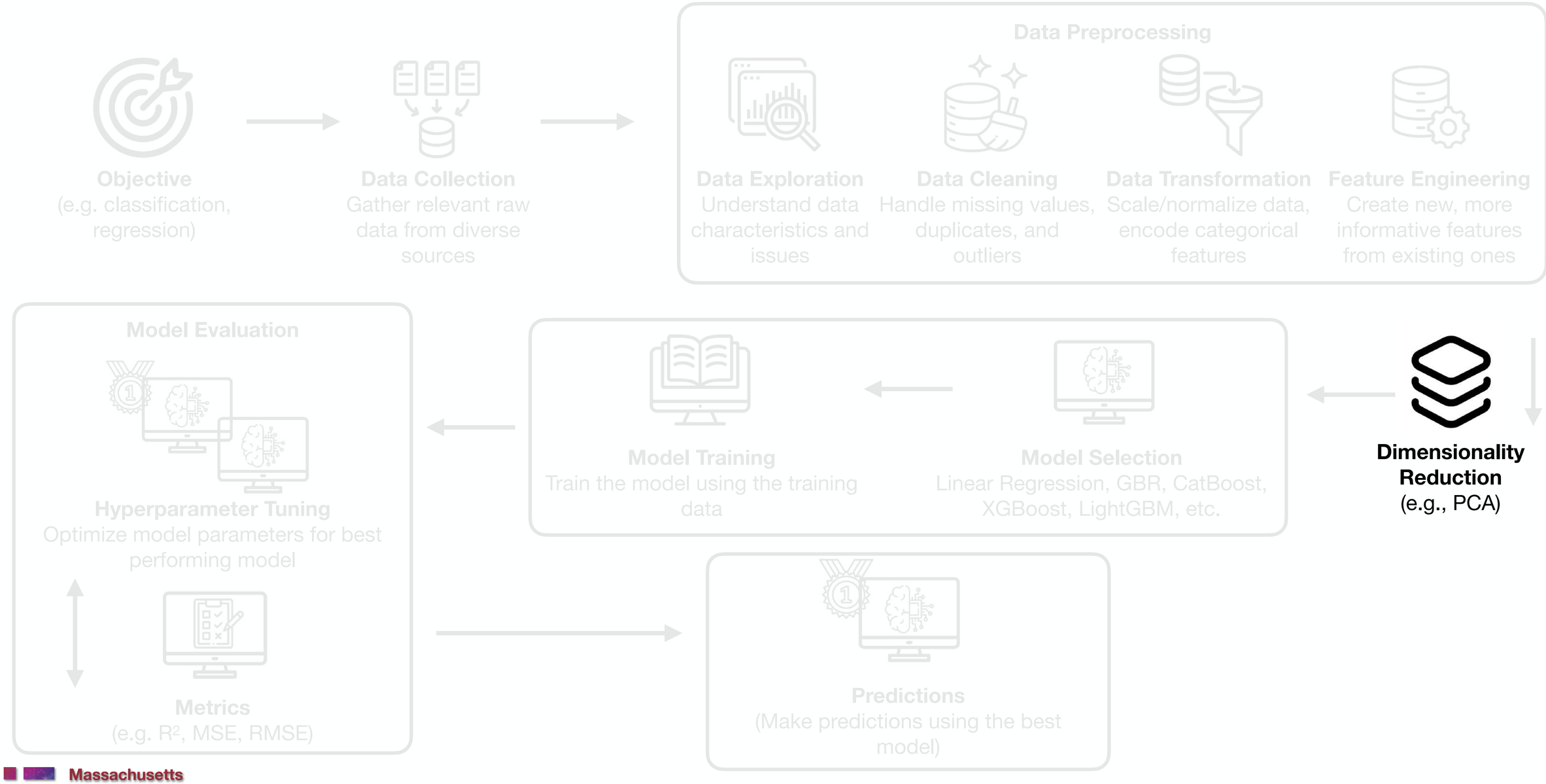
Hyperparameter Tuning
Optimize model parameters for best performing model

Metrics
(e.g. R^2, MSE, RMSE)

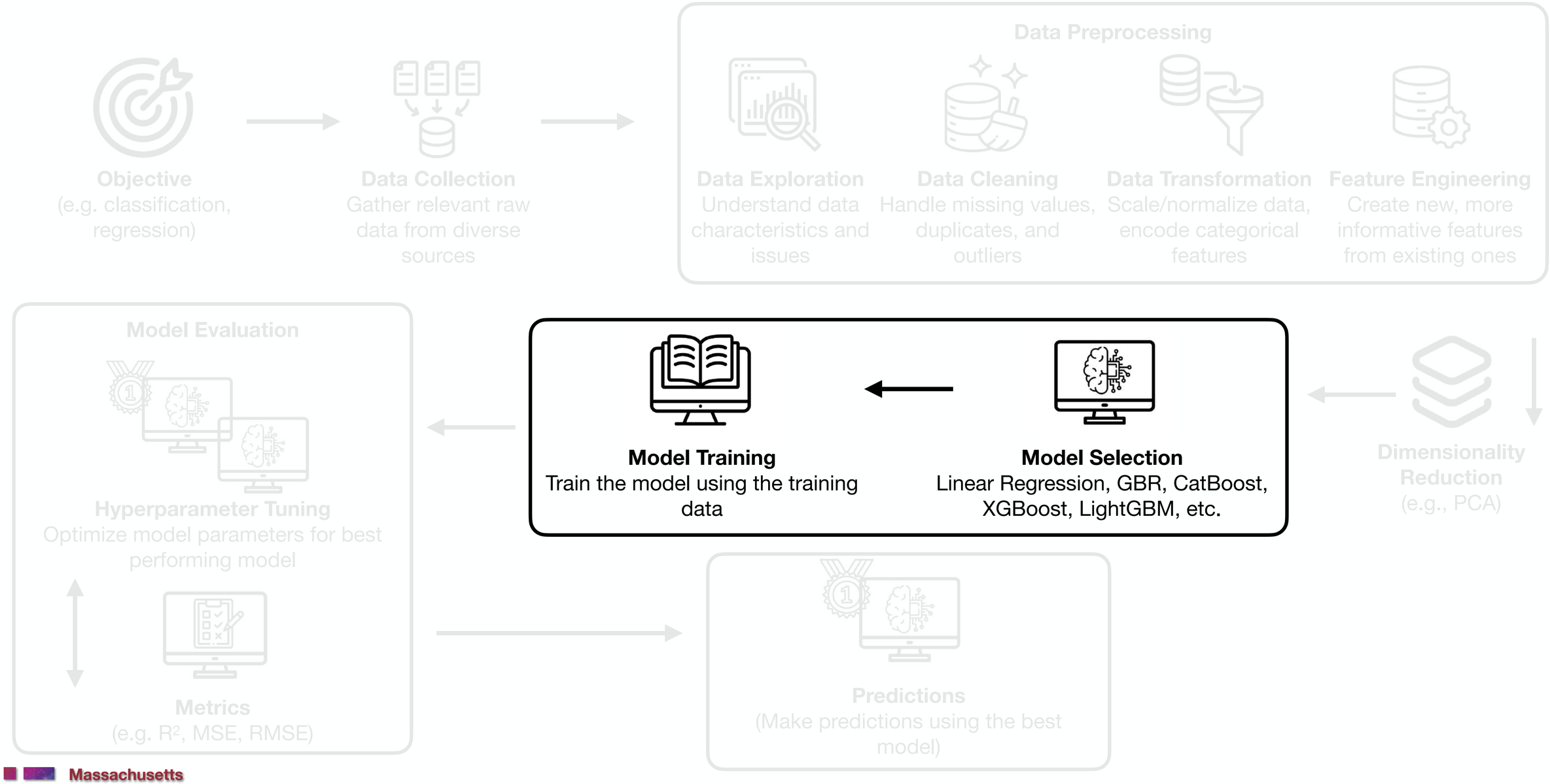
Predictions
(Make predictions using the best model)

ChemXploreML

Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424



ChemXploreML



CHEMXPLOREML

Objective
(e.g. classification, regression)

Data Collection
Gather relevant raw data from diverse sources

Data Preprocessing

Data Exploration
Understand data characteristics and issues

Data Cleaning
Handle missing values, duplicates, and outliers

Data Transformation
Scale/normalize data, encode categorical features

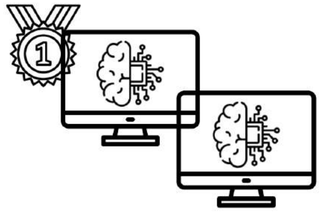
Feature Engineering
Create new, more informative features from existing ones

Dimensionality Reduction
(e.g., PCA)

Model Selection
Linear Regression, GBR, CatBoost, XGBoost, LightGBM, etc.

Model Training
Train the model using the training data

Model Evaluation



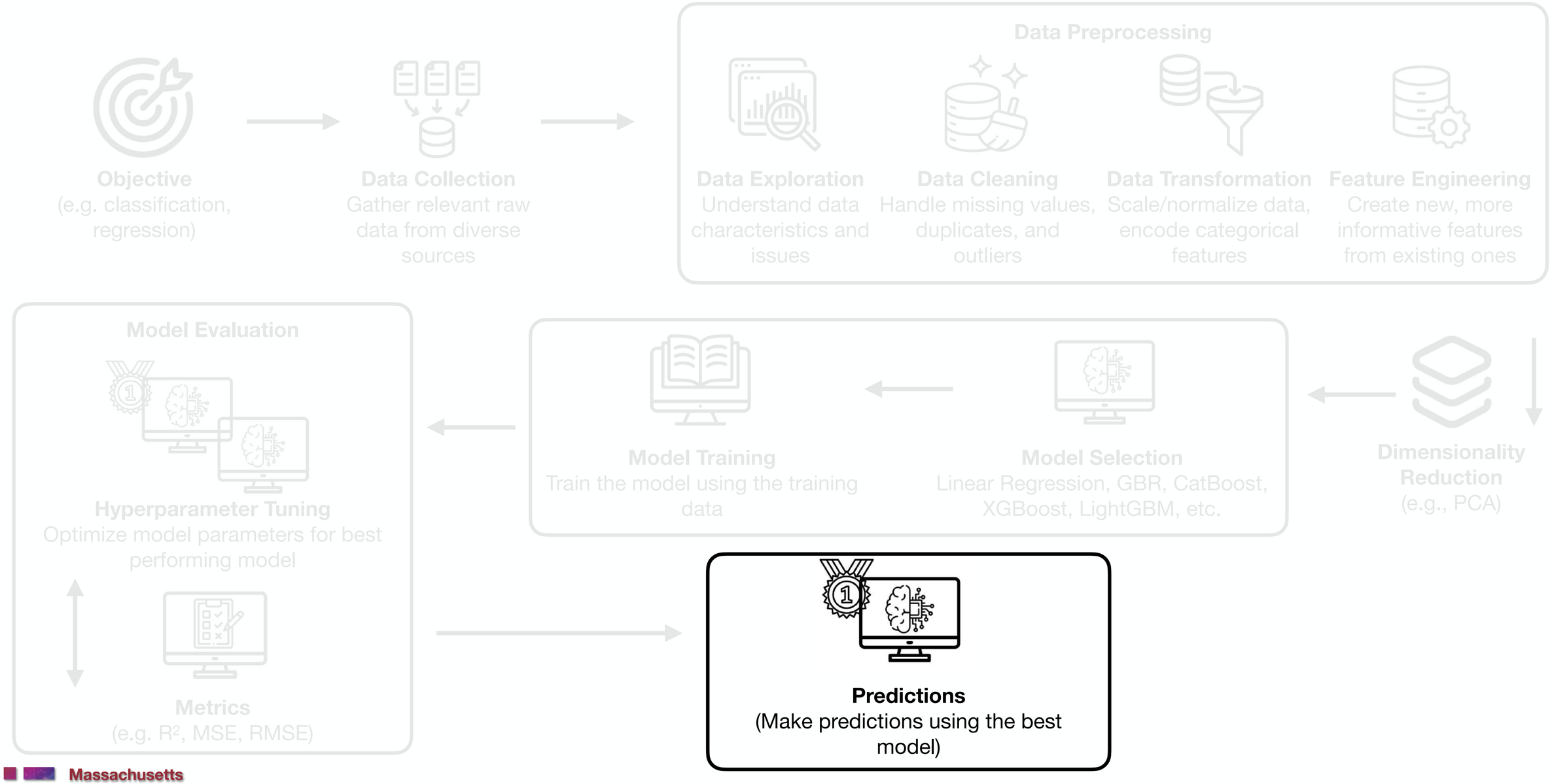
Hyperparameter Tuning
Optimize model parameters for best performing model



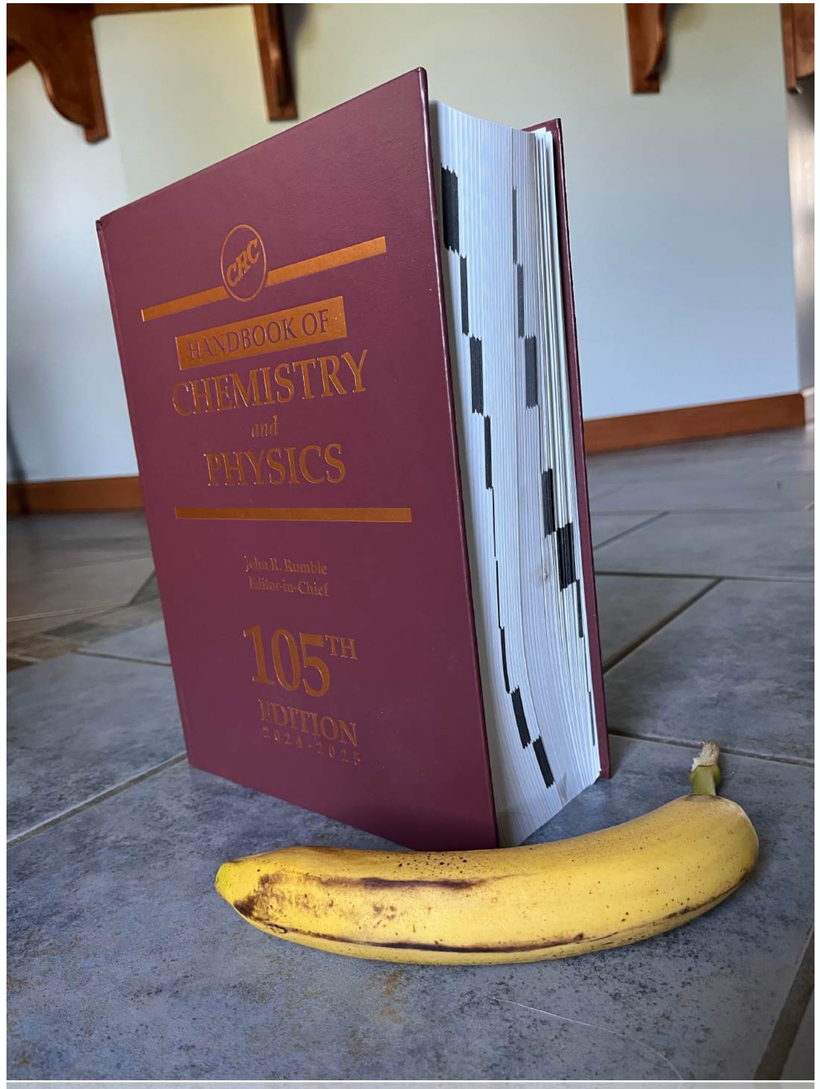
Metrics
(e.g. R^2, MSE, RMSE)

Predictions
(Make predictions using the best model)

CHEMXPLOREML



ChemXploreML - Properties



Melting Point

Boiling Point

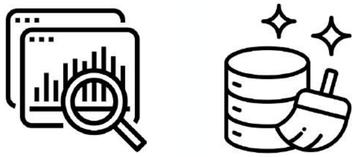
Critical Pressure

Critical Temperature

Vapor Pressure

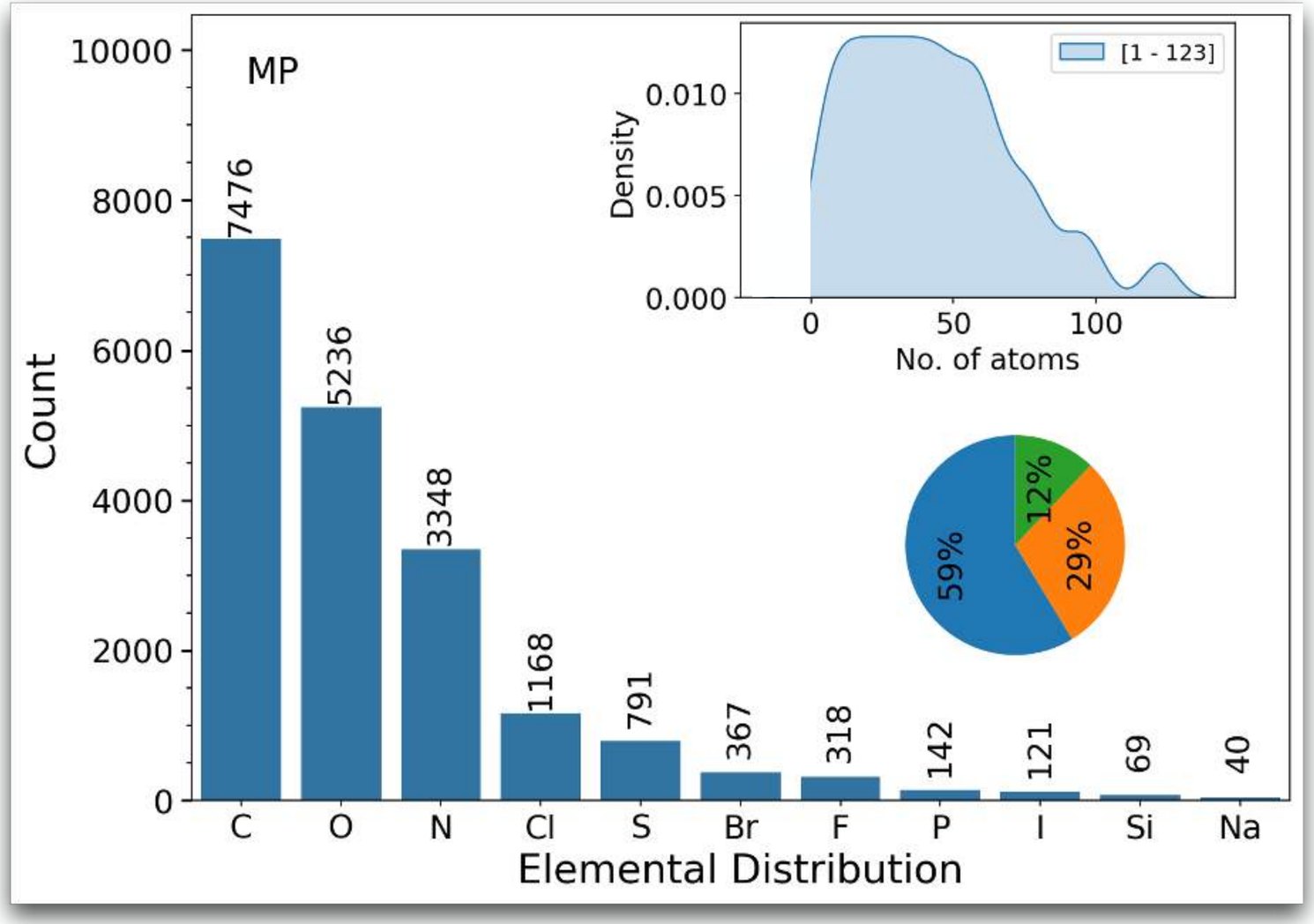
ChemXploreML - Properties

Preprocessing



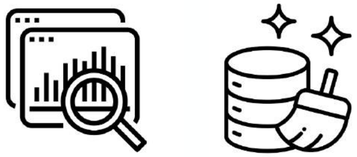
Question

What are some takeaways from this visualization of the data?



ChemXploreML - Properties

Preprocessing

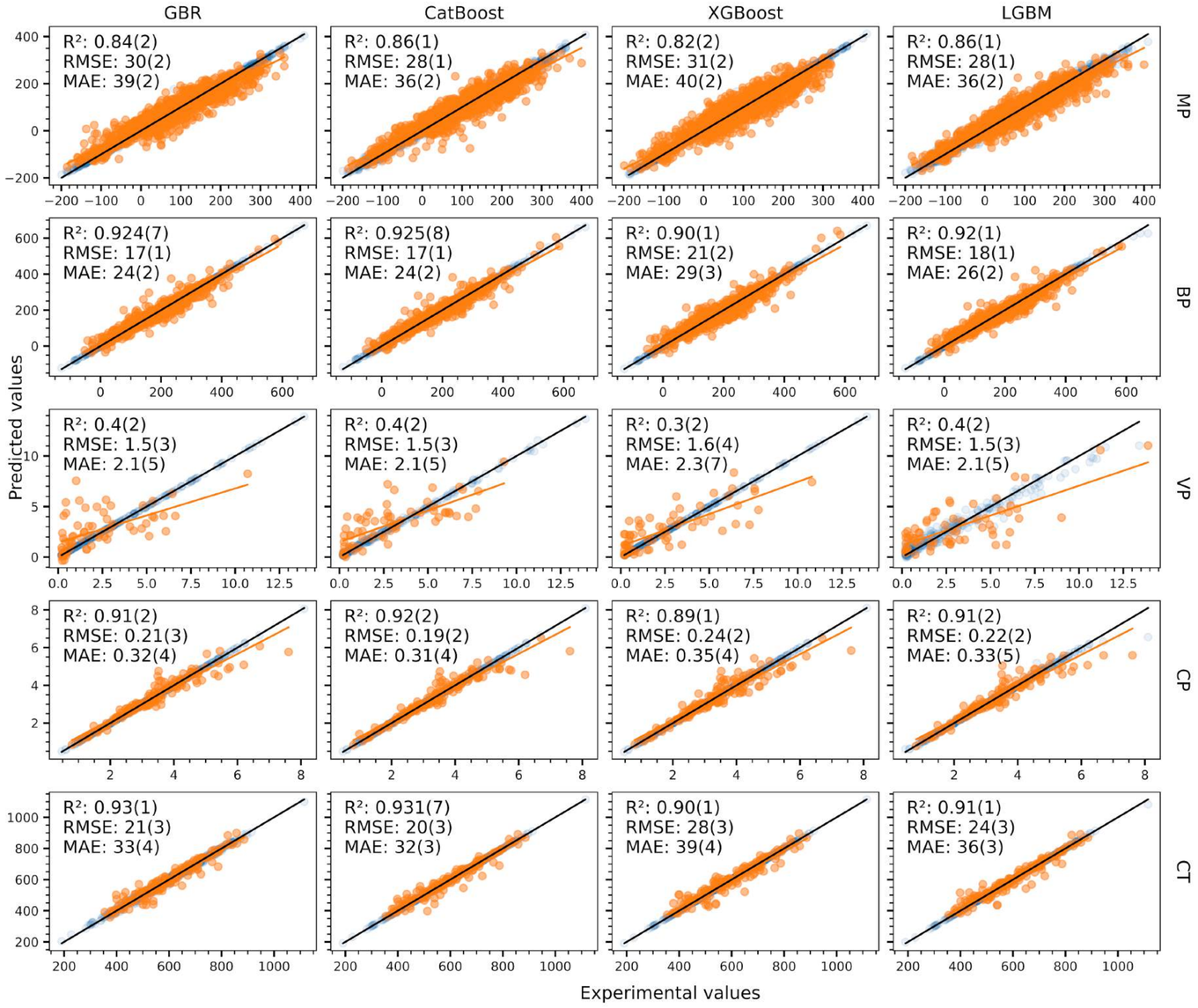


Question

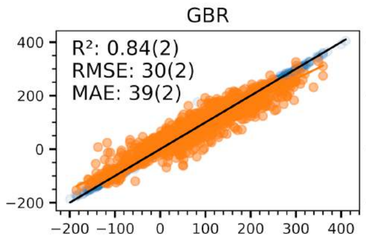
What should we be cautious about here?

Generate Embeddings

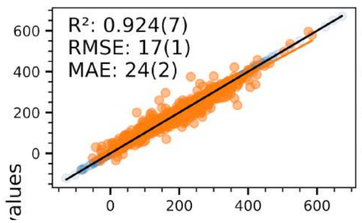
Pre-packaged with pre-trained mol2vec and VICGAE embedders


GBR
CatBoost
XGBoost
LGBM
MP
BP
VP
CP
CT
R²: 0.84(2)
RMSE: 30(2)
MAE: 39(2)
R²: 0.86(1)
RMSE: 28(1)
MAE: 36(2)
R²: 0.82(2)
RMSE: 31(2)
MAE: 40(2)
R²: 0.86(1)
RMSE: 28(1)
MAE: 36(2)
R²: 0.924(7)
RMSE: 17(1)
MAE: 24(2)
R²: 0.925(8)
RMSE: 17(1)
MAE: 24(2)
R²: 0.90(1)
RMSE: 21(2)
MAE: 29(3)
R²: 0.92(1)
RMSE: 18(1)
MAE: 26(2)
R²: 0.4(2)
RMSE: 1.5(3)
MAE: 2.1(5)
R²: 0.4(2)
RMSE: 1.5(3)
MAE: 2.1(5)
R²: 0.3(2)
RMSE: 1.6(4)
MAE: 2.3(7)
R²: 0.4(2)
RMSE: 1.5(3)
MAE: 2.1(5)
R²: 0.91(2)
RMSE: 0.21(3)
MAE: 0.32(4)
R²: 0.92(2)
RMSE: 0.19(2)
MAE: 0.31(4)
R²: 0.89(1)
RMSE: 0.24(2)
MAE: 0.35(4)
R²: 0.91(2)
RMSE: 0.22(2)
MAE: 0.33(5)
R²: 0.93(1)
RMSE: 21(3)
MAE: 33(4)
R²: 0.931(7)
RMSE: 20(3)
MAE: 32(3)
R²: 0.90(1)
RMSE: 28(3)
MAE: 39(4)
R²: 0.91(1)
RMSE: 24(3)
MAE: 36(3)
Predicted values
Experimental values

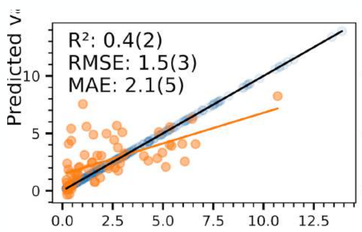
CHEMXPLOREML - PROPERTIES



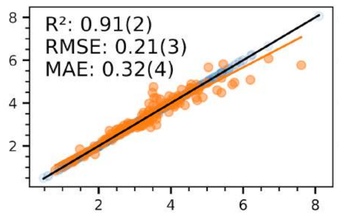
Melting Point



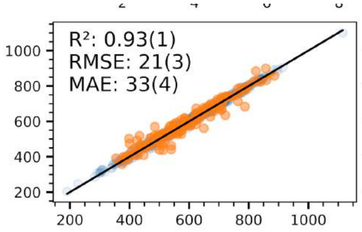
Boiling Point



Vapor Pressure



Critical Pressure



Critical Temperature

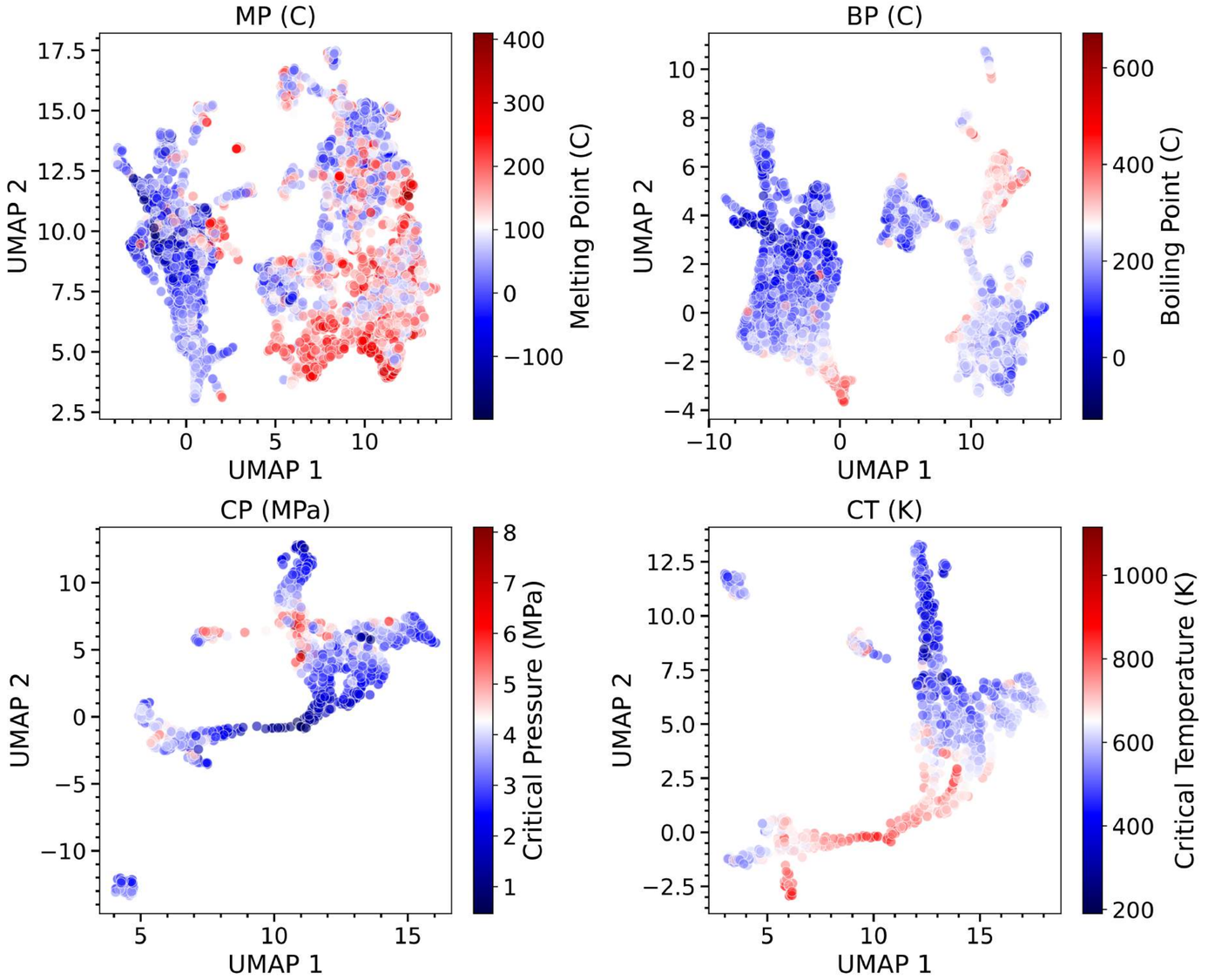
Training Data
Test Data

Question

Why is the vapor pressure fit relatively poor?

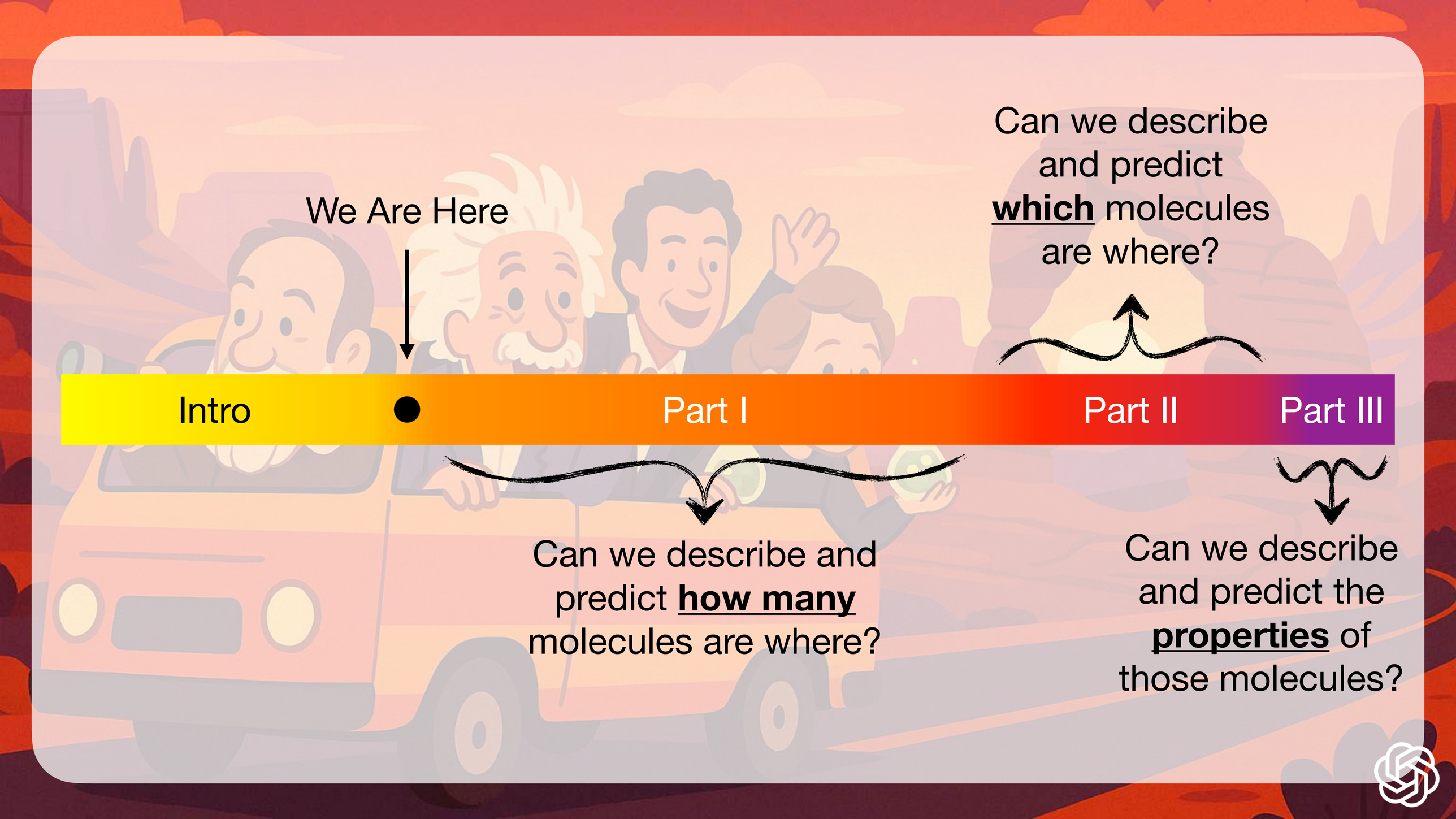
ChemXploreML - Properties

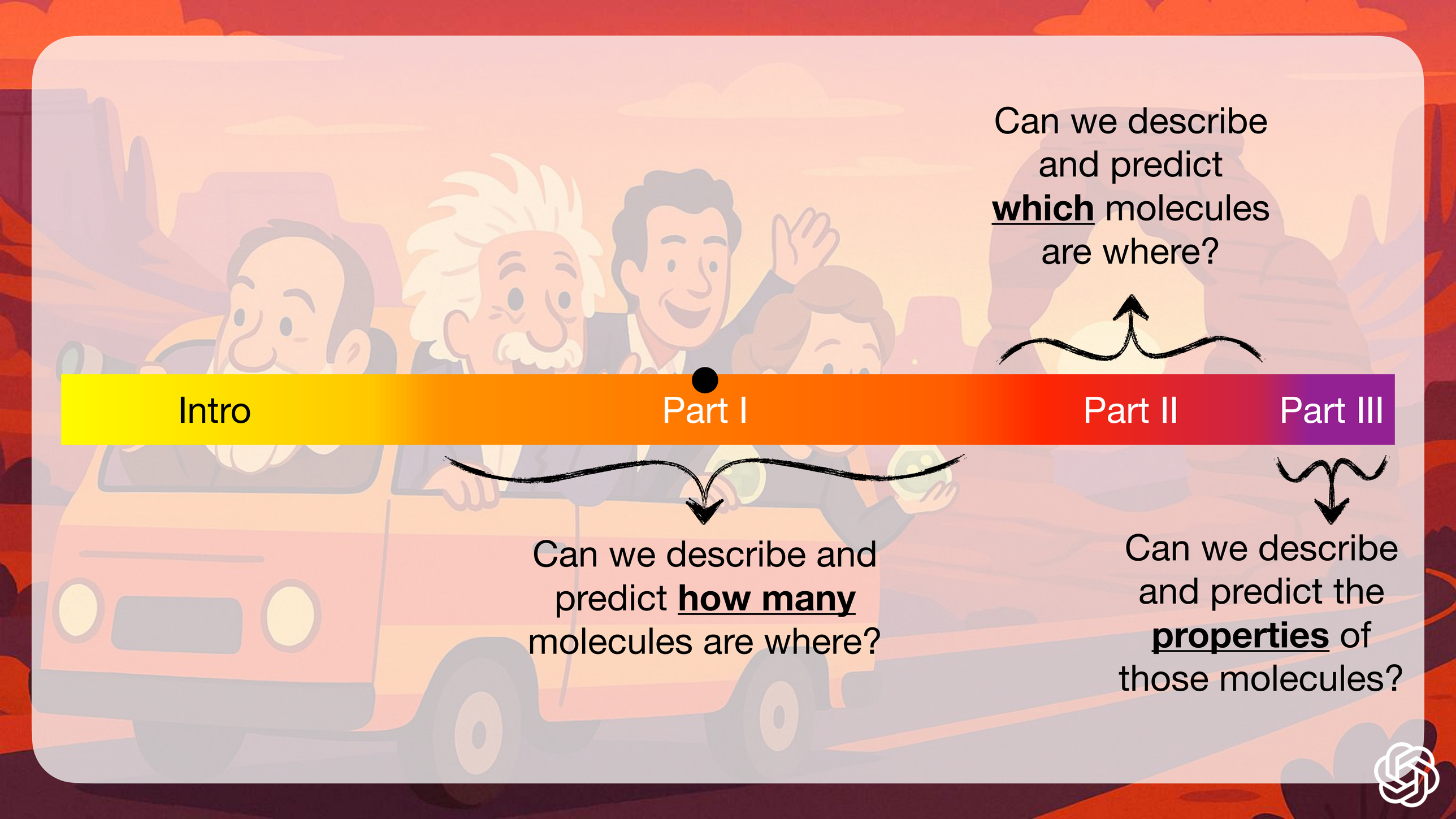
Marimuthu & McGuire 2025 *J. Chem. Inf. Mod.* 65, 5424

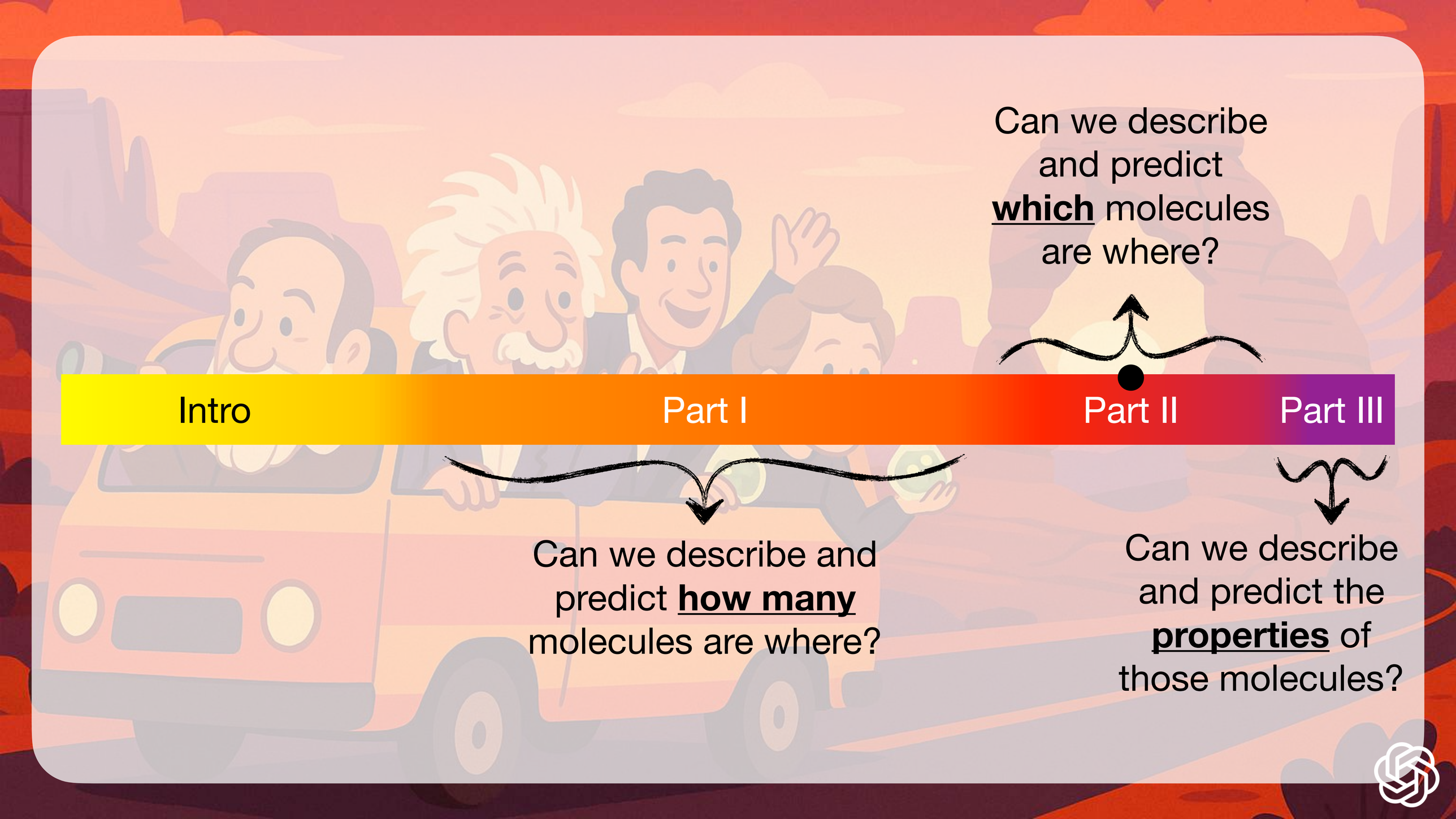


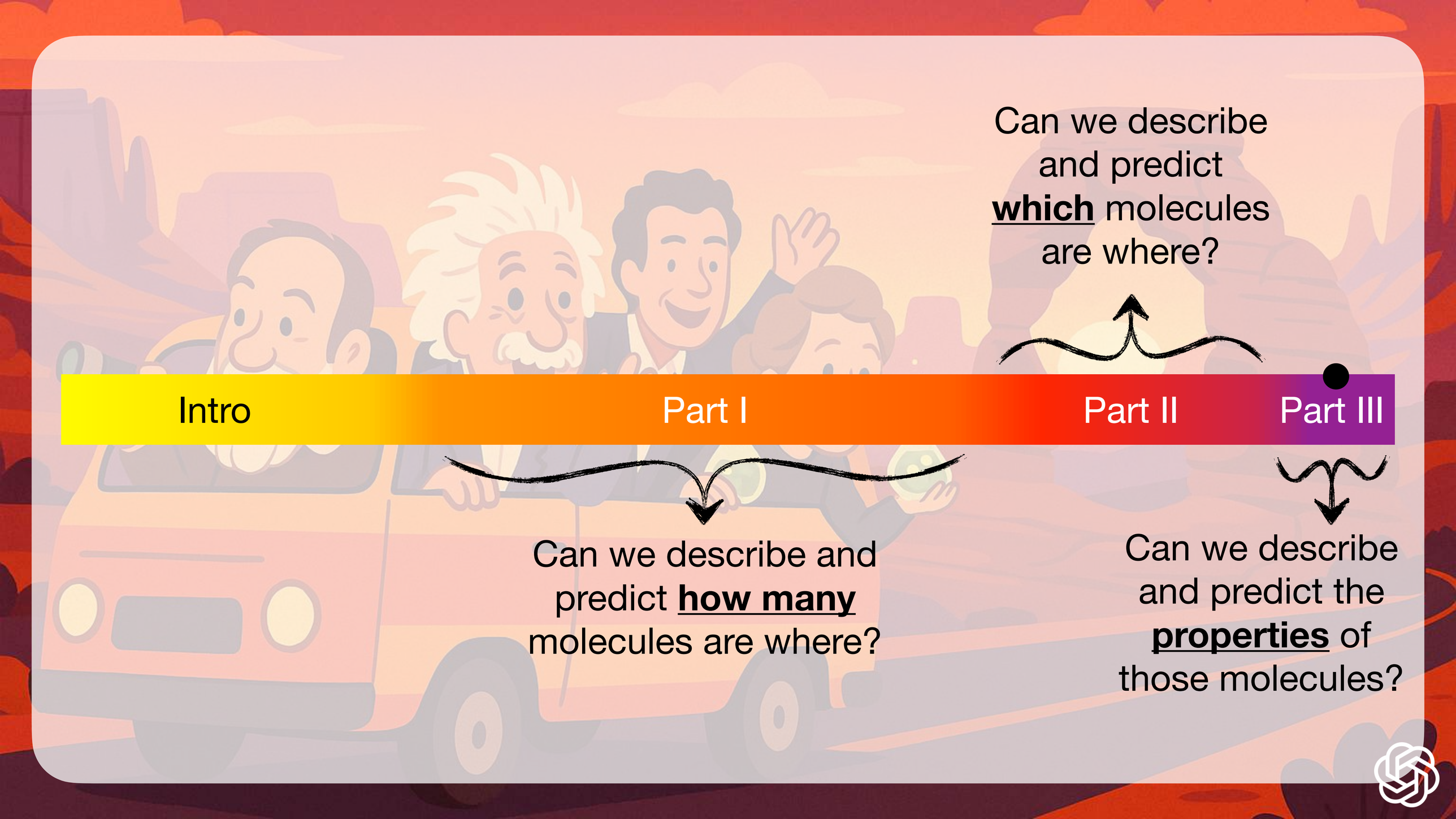
Question

Why is the vapor pressure fit relatively poor?


We Are Here
Can we describe and predict **which** molecules are where?
Intro
Part I
Part II
Part III
Can we describe and predict **how many** molecules are where?
Can we describe and predict the **properties** of those molecules?


Can we describe and predict **which** molecules are where?
Intro
Part I
Part II
Part III
Can we describe and predict **how many** molecules are where?
Can we describe and predict the **properties** of those molecules?


Can we describe and predict **which** molecules are where?
Intro
Part I
Part II
Part III
Can we describe and predict **how many** molecules are where?
Can we describe and predict the **properties** of those molecules?


Can we describe and predict **which** molecules are where?
Intro
Part I
Part II
Part III
Can we describe and predict **how many** molecules are where?
Can we describe and predict the **properties** of those molecules?

Acknowledgements







• This work is based upon work from COST Action CA21126 - Carbon molecular nanostructures in space (NanoSpace), supported by COST (European Cooperation in Science & Technology).

• COST (European Cooperation in Science & Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

• www.cost.eu
• https://www.cost.eu/actions/CA21126/
• https://research.iac.es/proyecto/nanospace/