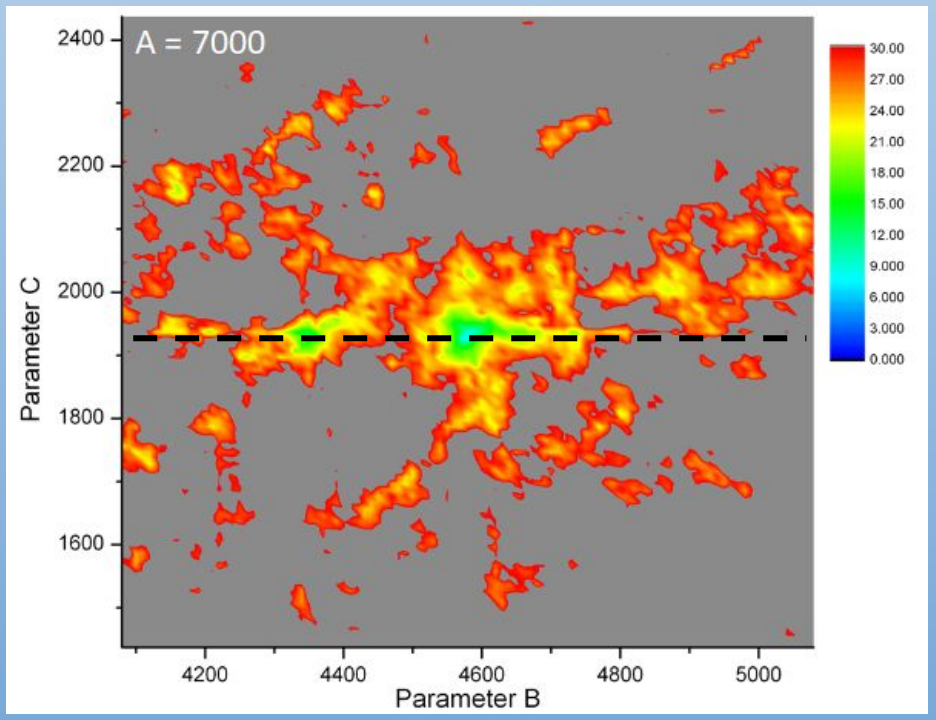
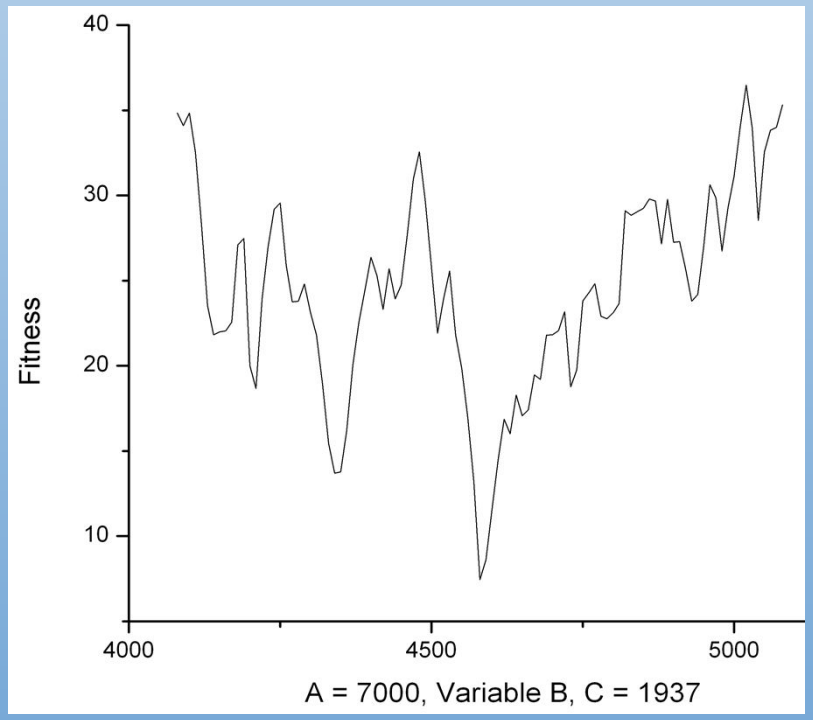
An Overview of Machine Learning in Rotational Spectroscopy

Steven Shipman
New College of Florida
June 2022





Goals of this overview talk

- Introduce some of the intricacies of rotational spectroscopy to people who aren't rotational spectroscopists
- Introduce concepts of machine learning to people who haven't done much with it themselves
- Provide an overview of what's currently out there and discuss strengths and weaknesses
- Identify current obstacles for computational rotational spectroscopy and suggest paths forward

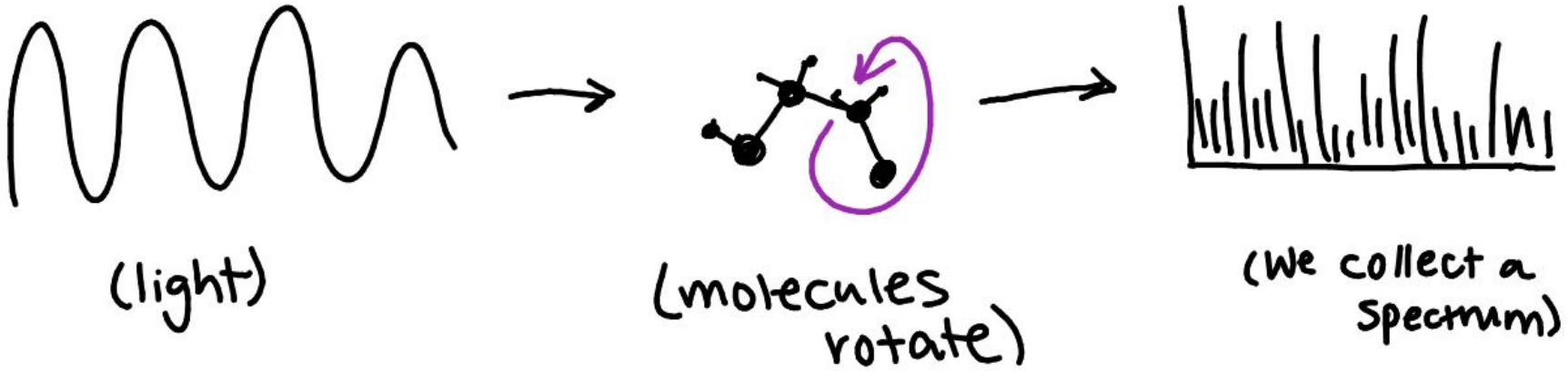
Brief rotational spectroscopy background

Molecules with non-zero dipole moments absorb/emit light in the microwave region of the spectrum.

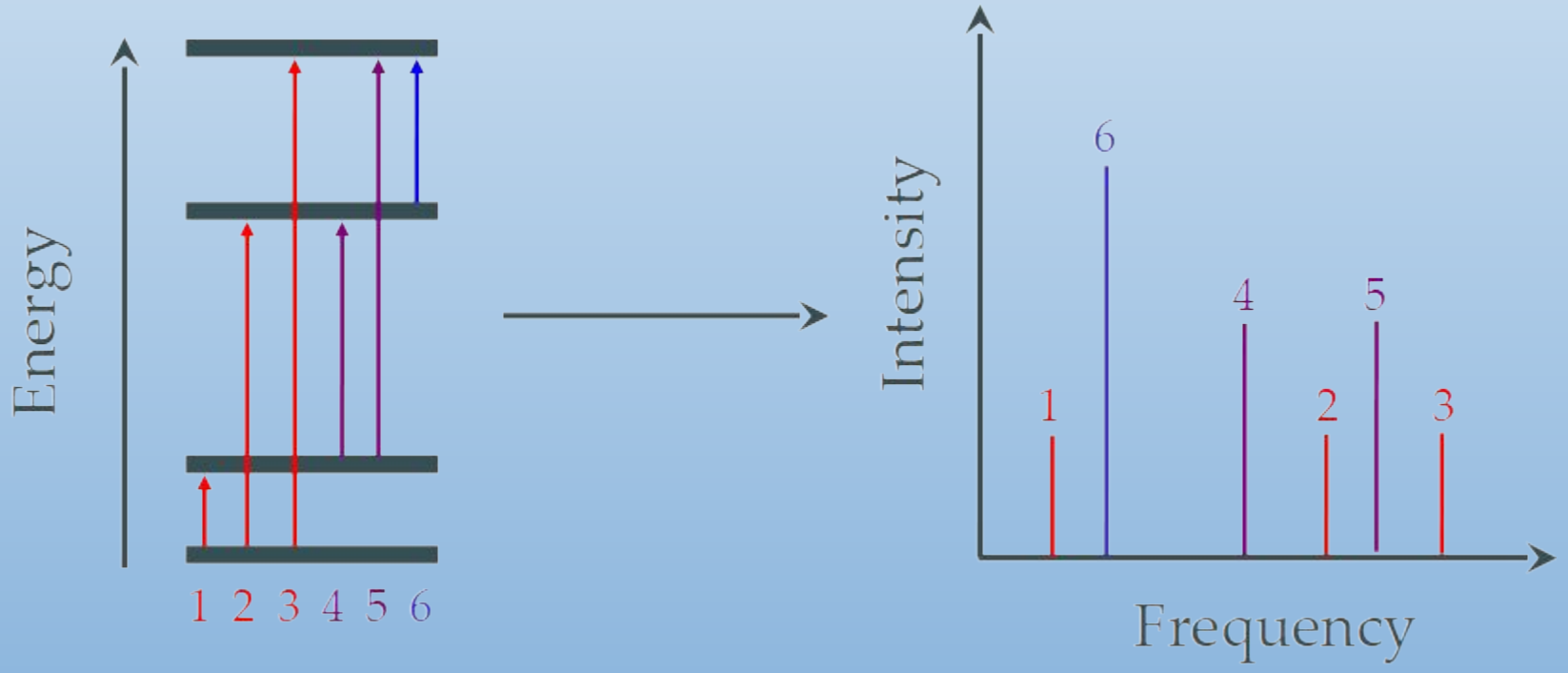
Spectra are high-resolution, allowing us to measure molecule shapes precisely

“Molecular fingerprinting”: identifying molecules in complicated mixtures

- Atmospheric chemistry
- Chemistry of space



Inverse problem: Forward prediction is easy

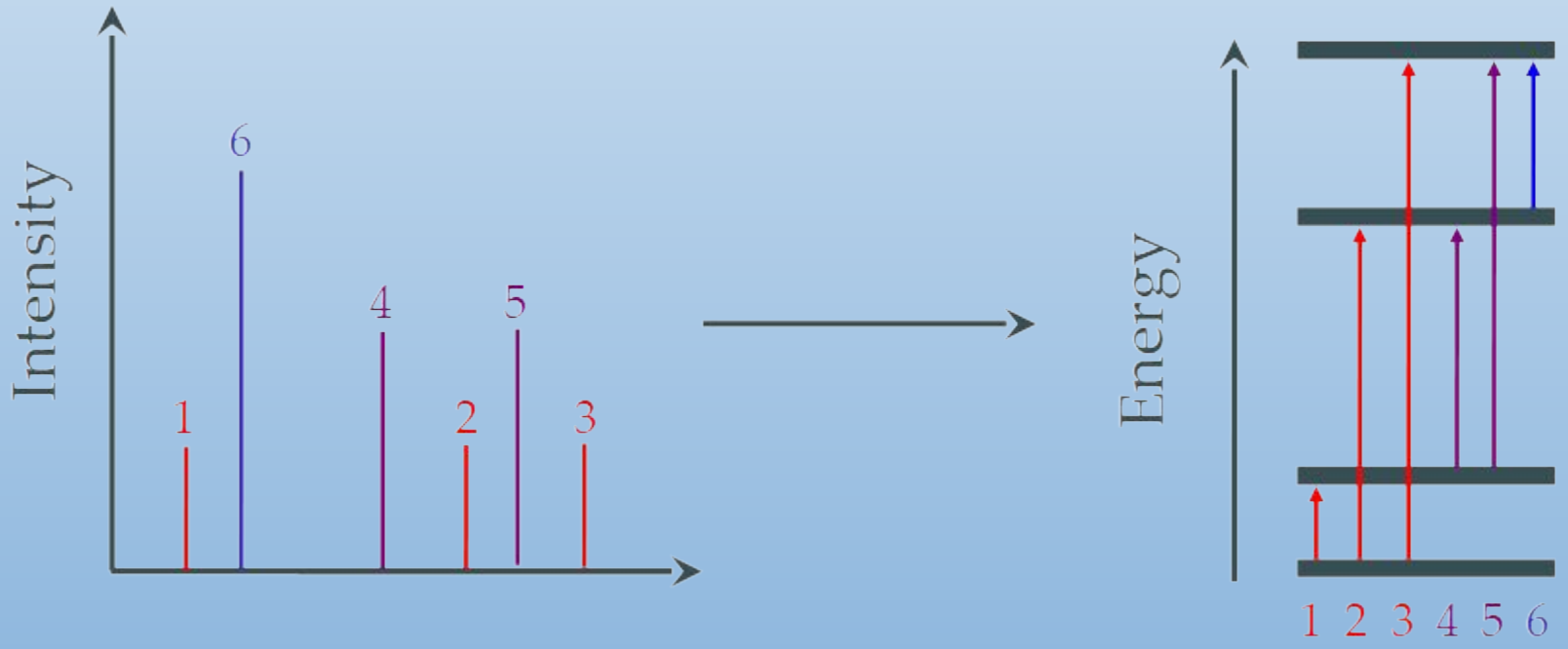


Procedure:

- Construct H and TD matrix from molecular parameters.
- Diagonalize H to get energy levels (eigenvalues).
- Use eigenvectors of H to transform TD.
- $|\text{TD element}|^2$ = prob. of transition (peak height), ΔE gives peak position.

Inverse problem: Assignment is hard!

Assignment: labeling a peak with upper and lower state quantum numbers

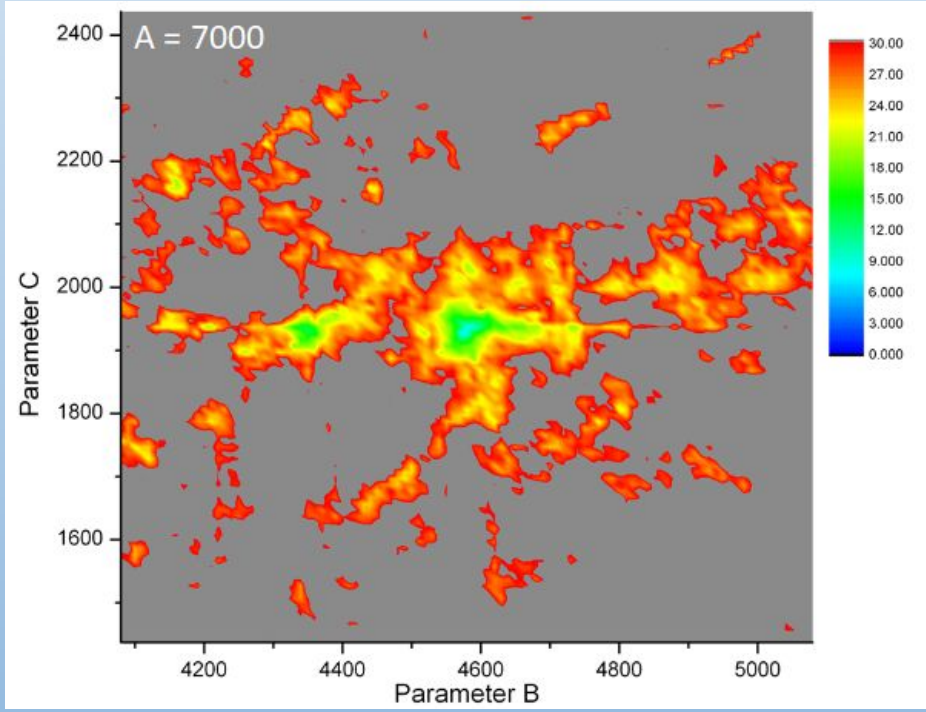


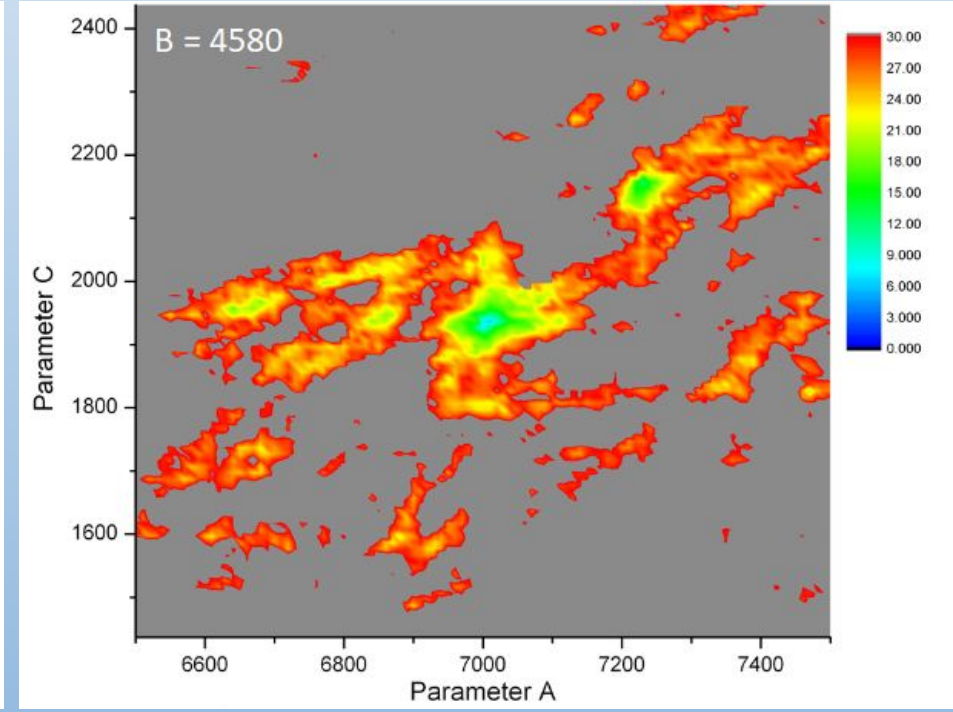
Severely over-determined – 8 parameters determine position of ~ 10^3 lines.

Acceptable fits require convergence of some parameters to ~ 1 part in 10^5.

Very rough landscape; minute changes completely change predictions.

Very Rough Fitness Landscapes

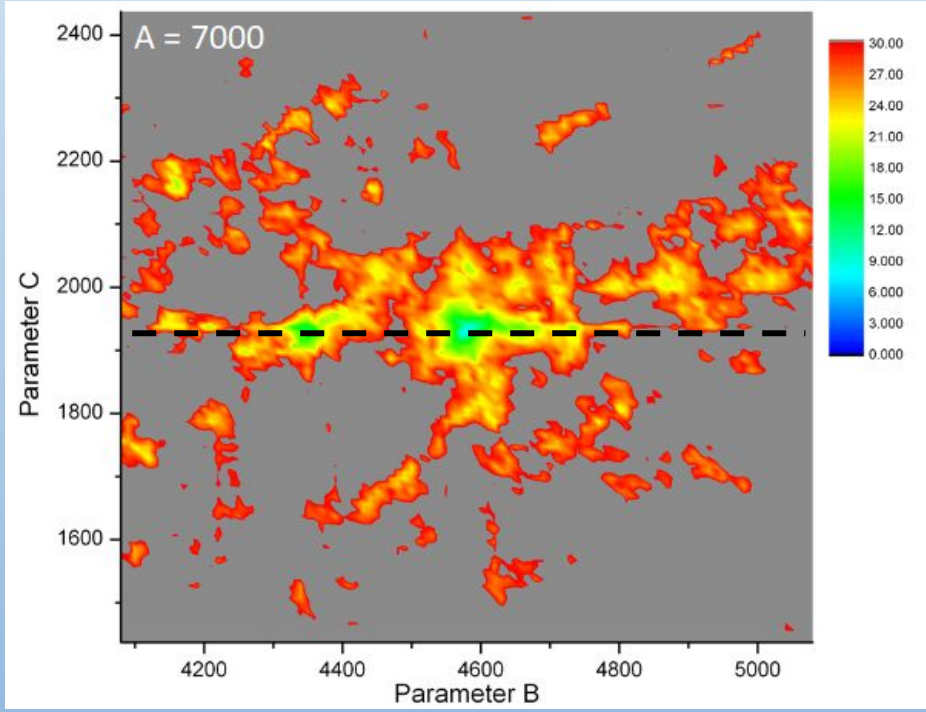


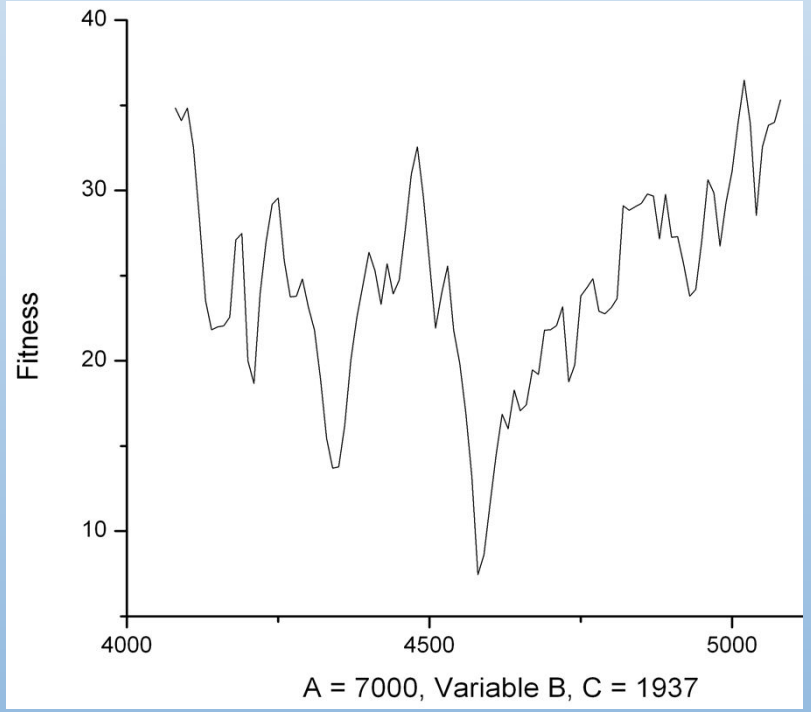


From: Katherine Ervin, “Automated Spectroscopic Analysis Using the Particle Swarm Optimization Algorithm: Implementing a Guided Search Algorithm to Autofit,” New College Undergraduate Thesis, 2017.

Similar plots can be found in Carroll, P.B., Lee, K.L.K., McCarthy, M.C., *J. Mol. Spec.* **379**, 111467 (2021).

Very Rough Fitness Landscapes

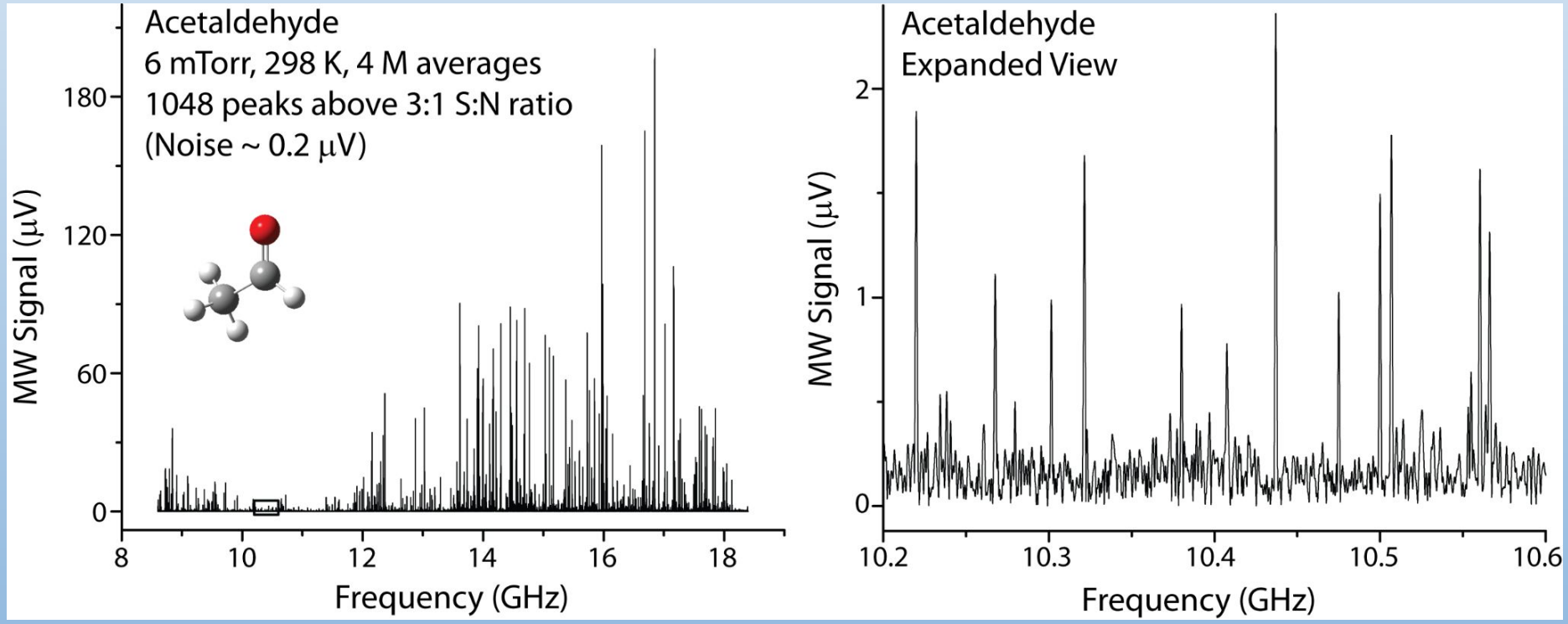




From: Katherine Ervin, "Automated Spectroscopic Analysis Using the Particle Swarm Optimization Algorithm: Implementing a Guided Search Algorithm to Autofit," New College Undergraduate Thesis, 2017.

Similar plots can be found in Carroll, P.B., Lee, K.L.K., McCarthy, M.C., *J. Mol. Spec.* **379**, 111467 (2021).

Room-temperature cm-wave data



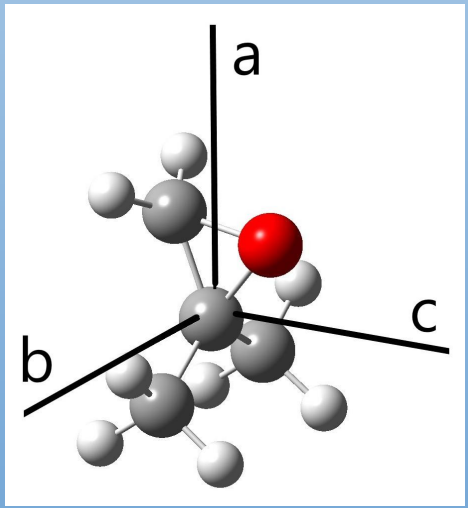
Spectra at 298 K are dramatically more crowded than at 2 K (higher-energy conformers, vibrational partition function, higher J)

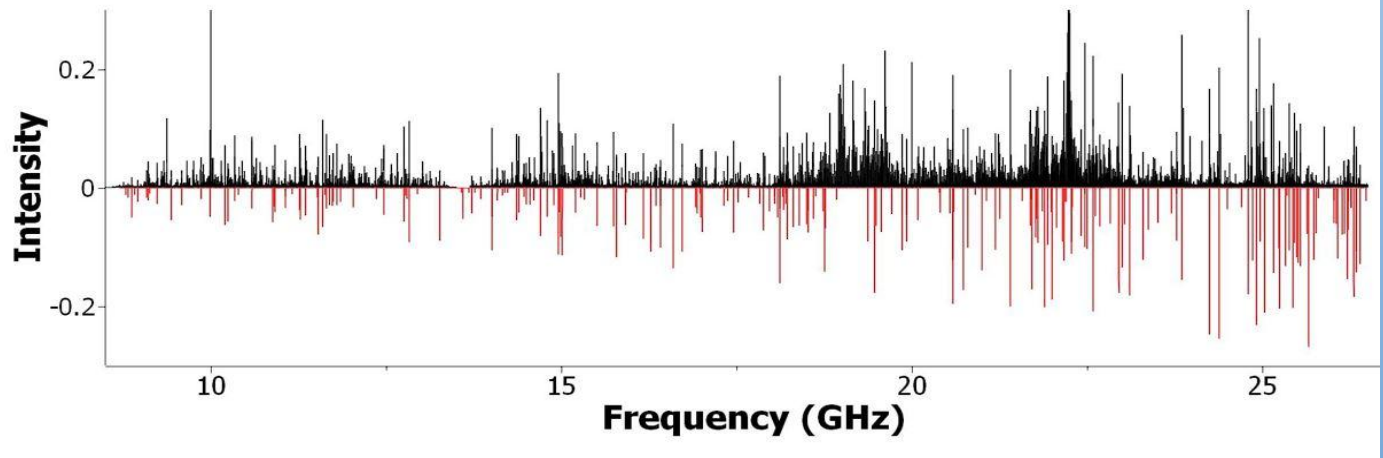
Poses a particular challenge at a PUI; spectra are very challenging for undergraduate students to pick apart and assign!

Typical Approach to Assignment

1) Get initial guess from *ab initio* methods
2) Tentatively assign a few lines based on pattern recognition
3) Non-linear least squares fit on selected parameters
4) Forward prediction with new parameters. Better or worse?
5) Loop until satisfied.

Severe issues with line density (conformers, excited vibrational states).
Few "close" fits – very frustrating (easy to spend hours with no progress).





Non-ML Approaches

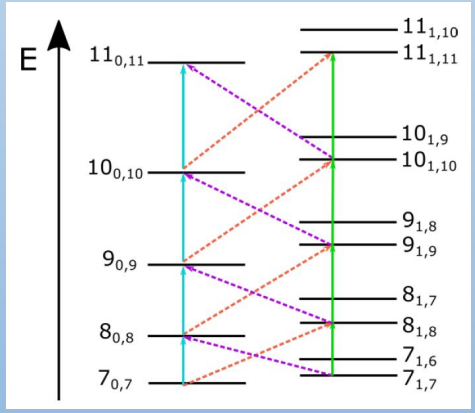
DAPPERS

Love, N. *et al.*, *J. Mol. Spec.*, **370**, 111294 (2020)

Image from DAPPERS documentation, available at:
http://kleopold.dl.umn.edu/content/dappers

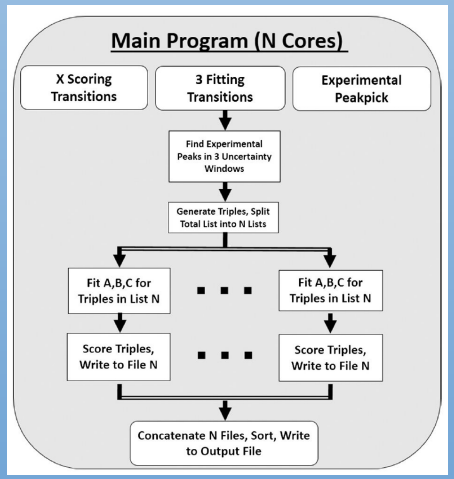
RAARR:

Yeh, L. *et al.*, *J. Chem. Phys.*, **150**, 204122 (2019)



AUTOFIT

Seifert, N.A. *et al.*, *J. Mol. Spec.*, **312**, 13 (2015)



What's Meant by Machine Learning?

Approaches that solve problems not via predefined algorithms but by using statistical properties of known data sets to guide future behavior. Involves an element of "training by example".

Typical kinds of problems:

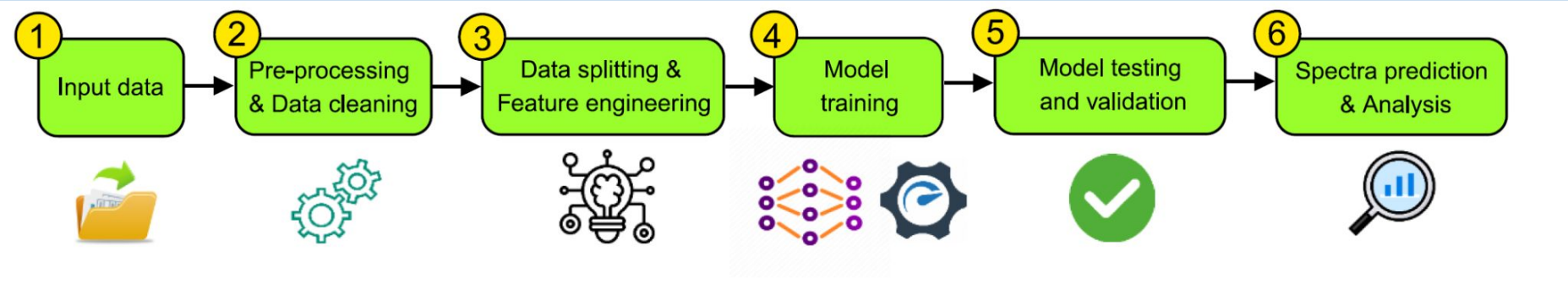
Classification Problems: Spam filtering
Anomaly Detection: Credit card fraud detection
Dynamic Decision-making: Self-driving vehicles
Forecasting: Weather forecasting

Need to be very clear about problem to solve:

- I have a spectrum - what rotational constants generate it?
- I have a very complicated spectrum - how many different molecules are present?
- I have rotational constants - what molecule has those constants?

Images generated by DALL-E mini: huggingface.co/spaces/dalle-mini/dalle-mini

ML Approaches and Architectures



From: Han, R., Ketkaew, R., Luber, S., *J. Phys. Chem. A* **126**, 801 - 812 (2022).

Supervised / Unsupervised / Reinforcement learning - how are you telling your system what you want it to do?

Neural networks / decision trees / support vector machines / many others

Important advantage for rotational spectroscopy - relatively easy to generate training sets since we have Hamiltonians and can simulate large numbers of spectra

Limitations / Concerns

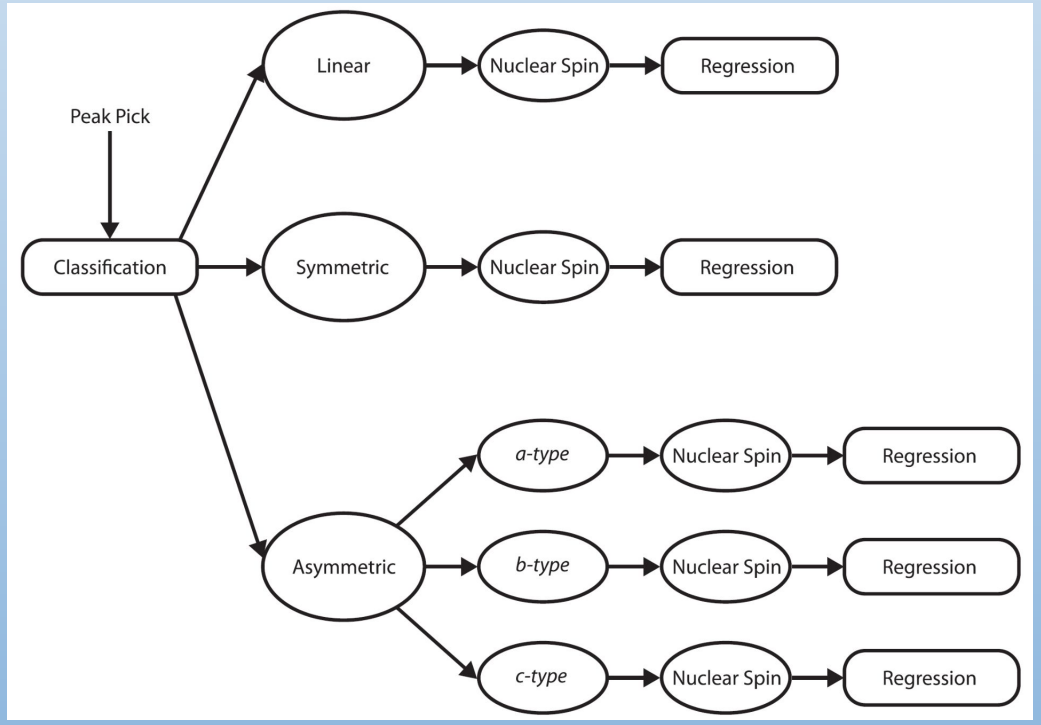
Edge / boundary problems - how sensitive is the system? Can be cases where small amounts of noise can lead to radically different outcomes.

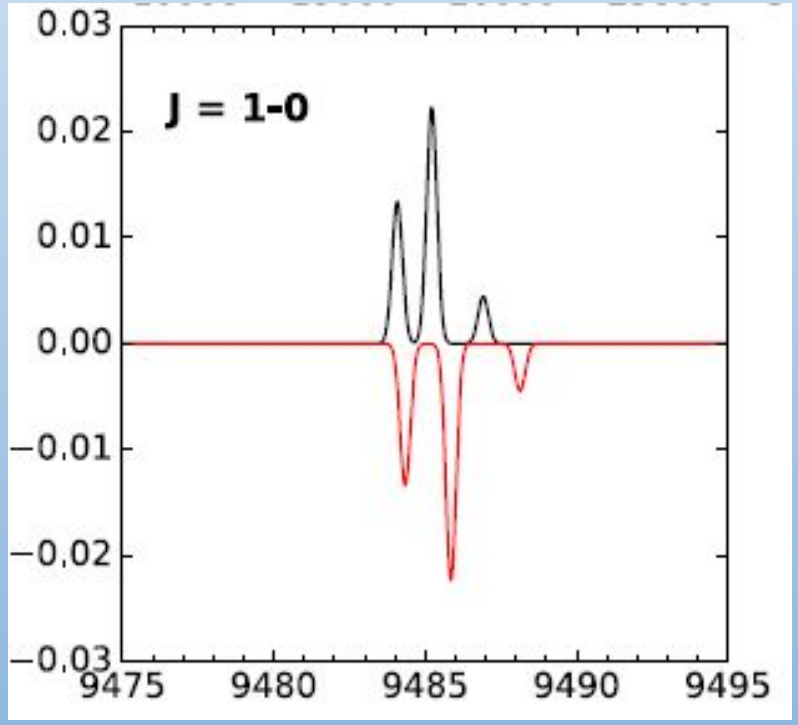
Overfitting - possible to “overtrain” a system so it identifies features peculiar to the training set. Less able to successfully handle new data that wasn’t in the training set.

Bias in training sets - typically generated by people and so can have sampling biases (ex: no van der Waals complexes in training set). Can unintentionally limit applicability.

Generality vs Specificity - truly general algorithm can only do so much; won’t outperform specialized algorithms within particular domains. Difficulty is figuring out in practice how general to be.

RAINet
(Rotational Assignment and Identification Network)



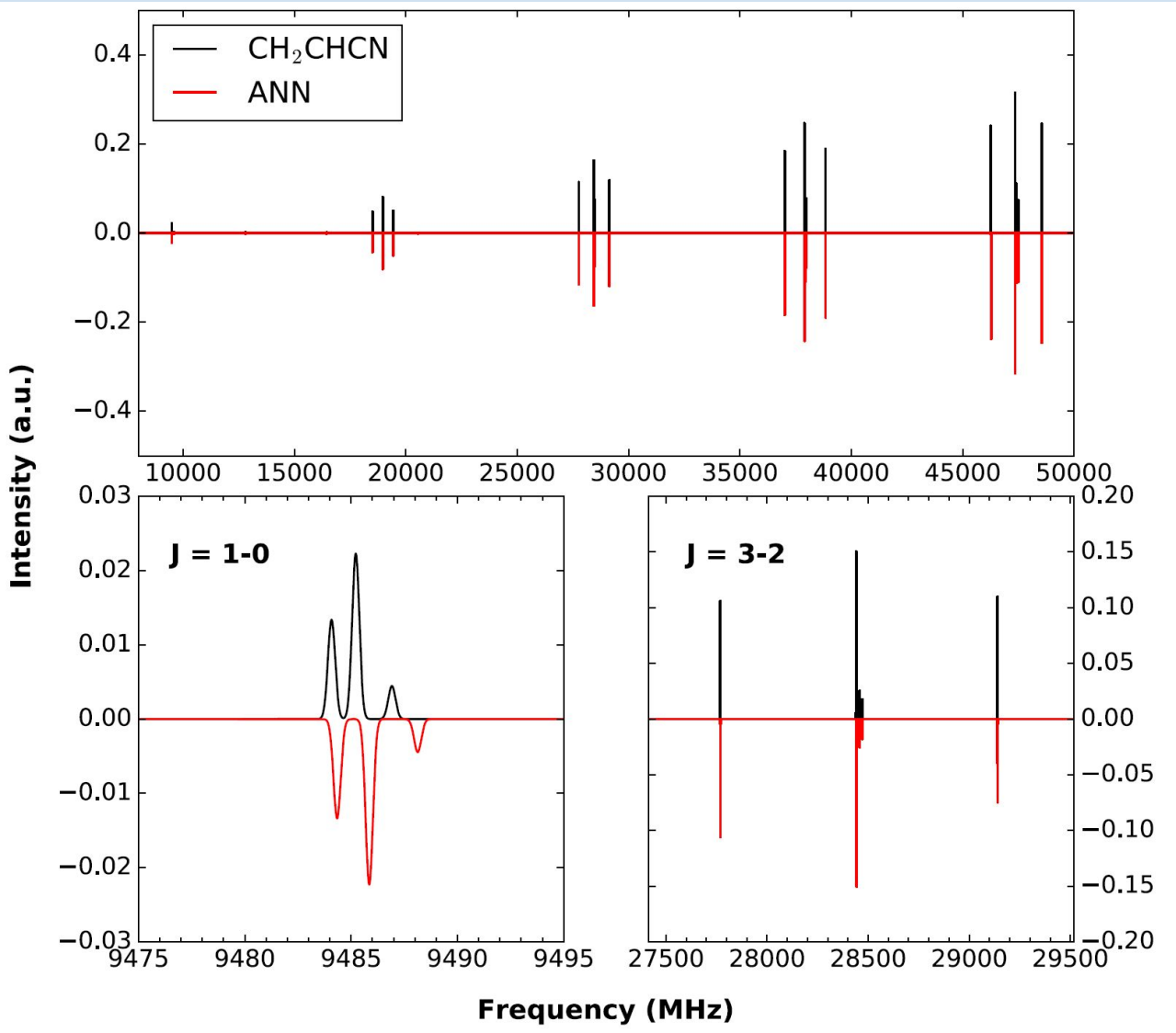


From: Zaleski, D.P. and Prozument, K., *J. Chem. Phys.* **149**, 104106 (2018)

Network of networks: classification, then regression networks for constants.

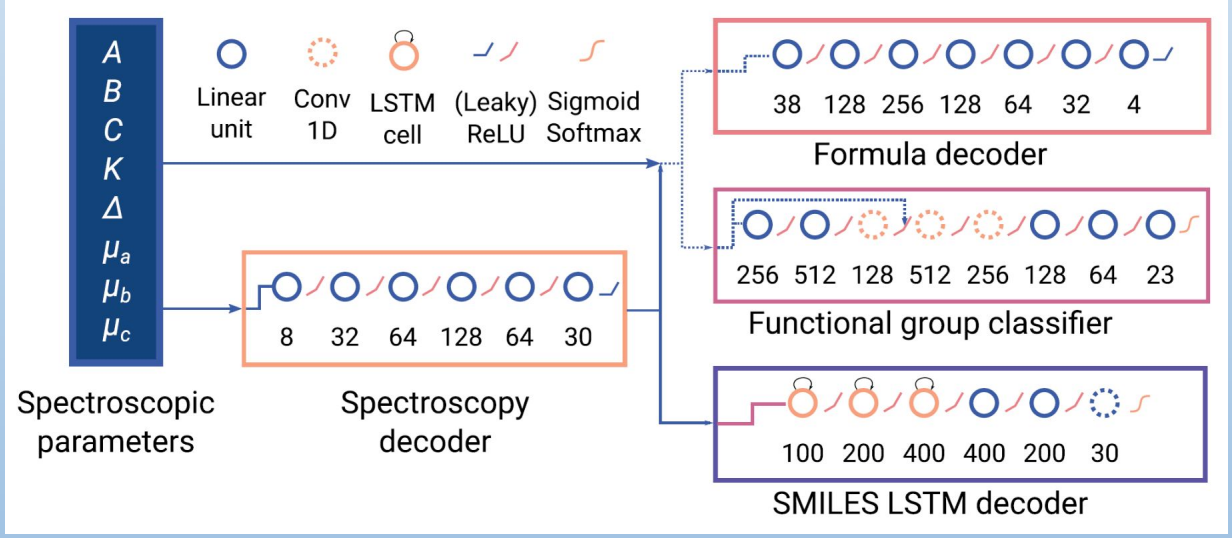
Can handle hyperfine structure, doesn't use intensity information

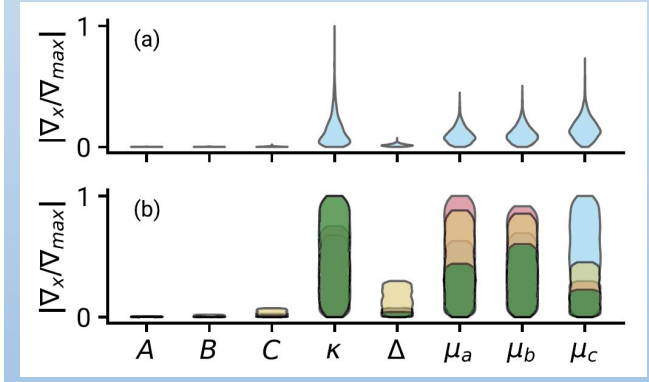
Difficulty with hybrid spectra and data containing more than one component



From: Zaleski, D.P. and Prozument, K., *J. Chem. Phys.* **149**, 104106 (2018)

Molecular Identity from Rotational Constants





From: McCarthy, M.C. and Lee, K.L.K., *J. Phys. Chem. A* **124**, 3002 - 3017 (2020).

Turns spectroscopic parameters into Coulomb matrix eigenvalues, then uses those to assess number and types of atoms, predict functional groups, generate SMILES strings.

SMILES strings turn out to be hard to get right, but number of heavy atoms and presence/absence of heteroatoms is easier to determine.

Other Challenges to Deal With

Incomplete or missing experimental information (partial spectral coverage, limited resolution)

Many approaches don't fully incorporate intensity or linewidth information

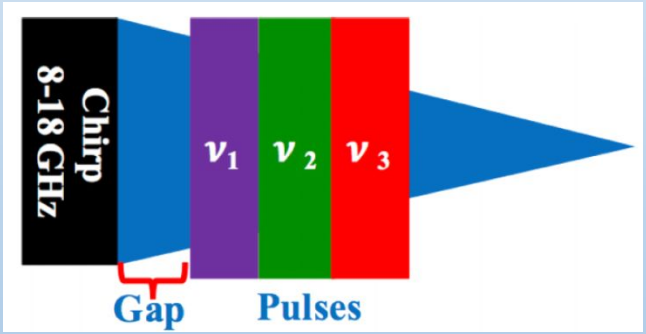
Need efficient ways to represent molecules

SPCAT is comparatively slow (for doing millions of calculations)

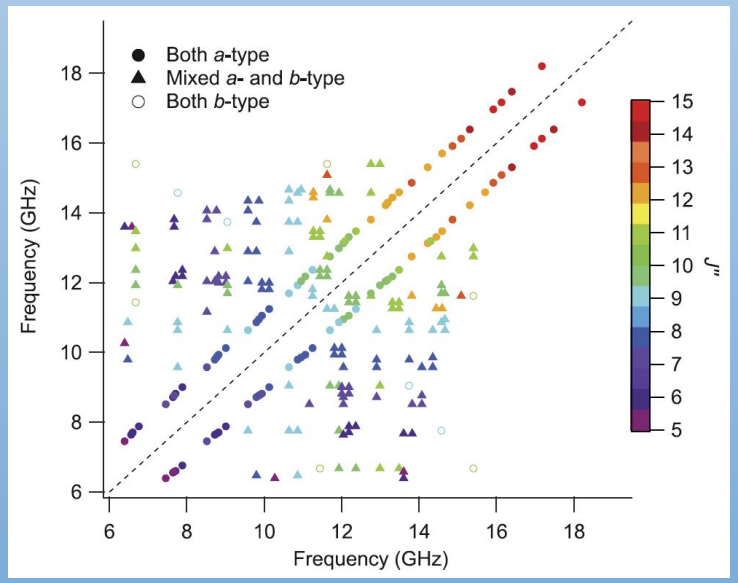
General usability/accessibility hurdles

Acquisition of Additional Information

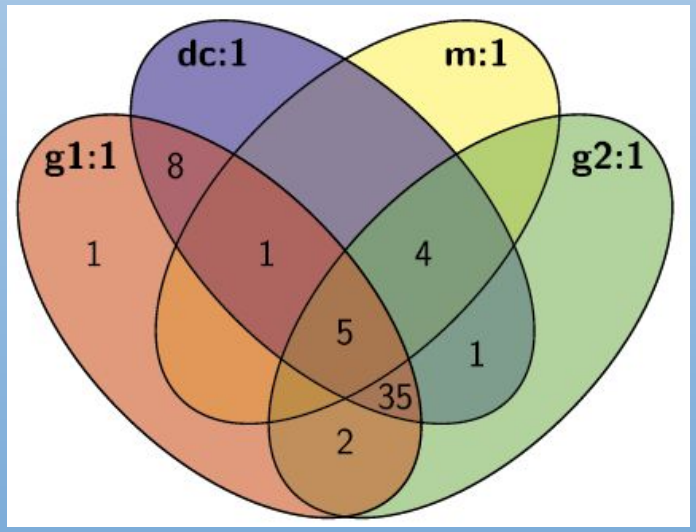
- DR / AMDOR
- Spectral Taxonomy
- Strong-Field Coherence Breaking



From: Hernandez-Castillo, A.O. *et al.*, *J. Chem. Phys.* **145**, 114203 (2016)

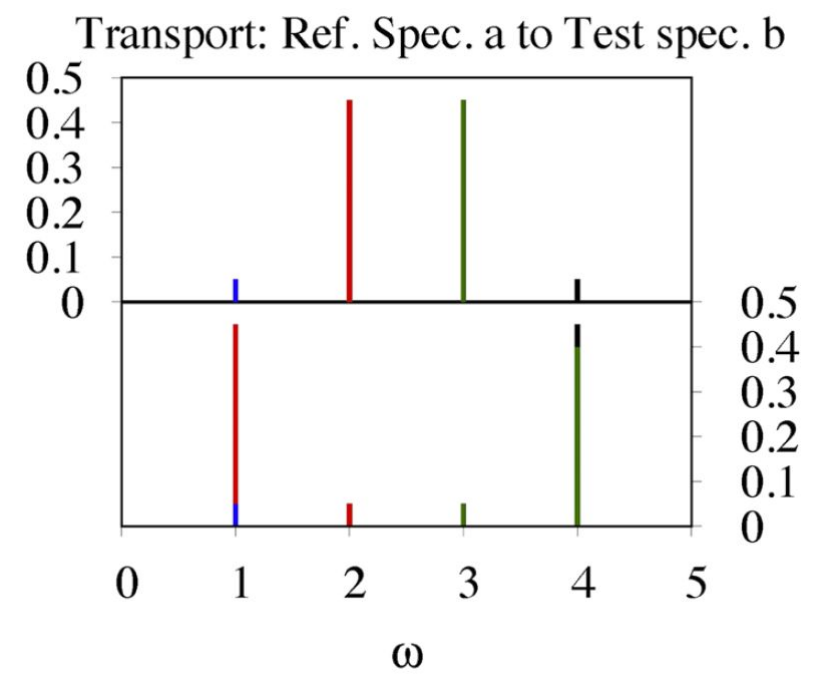


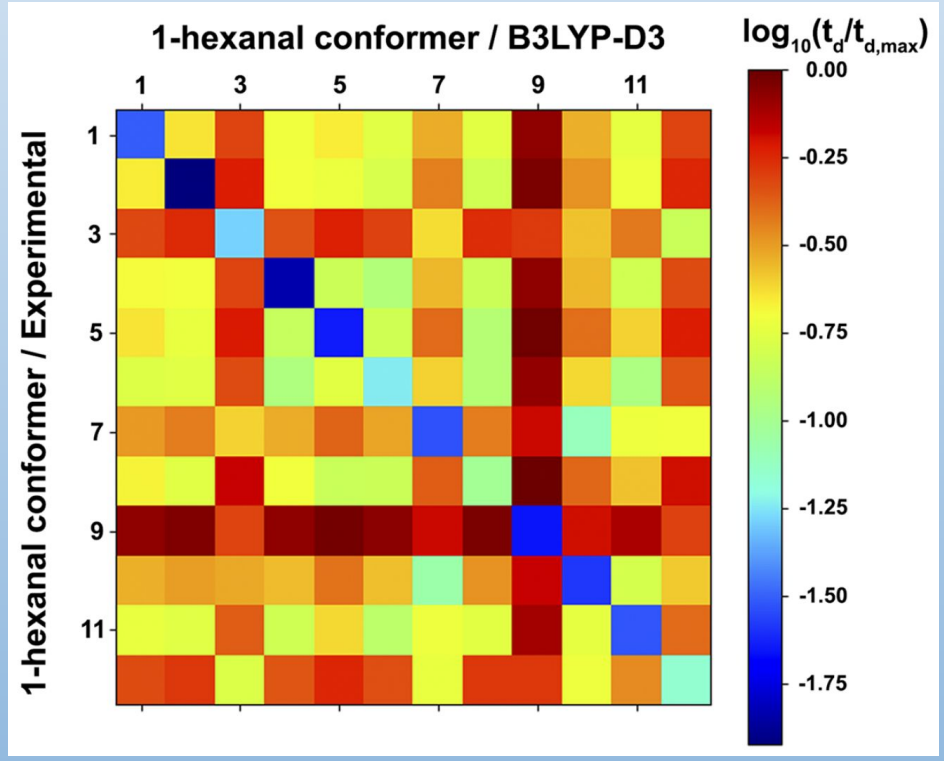
From: Martin-Drumel, M.A. *et al.*, *J. Chem. Phys.* **144**, 124202 (2016)



From: Crabtree, K.N. *et al.*, *J. Chem. Phys.* **144**, 124201 (2016)

Computational Optimal Transport



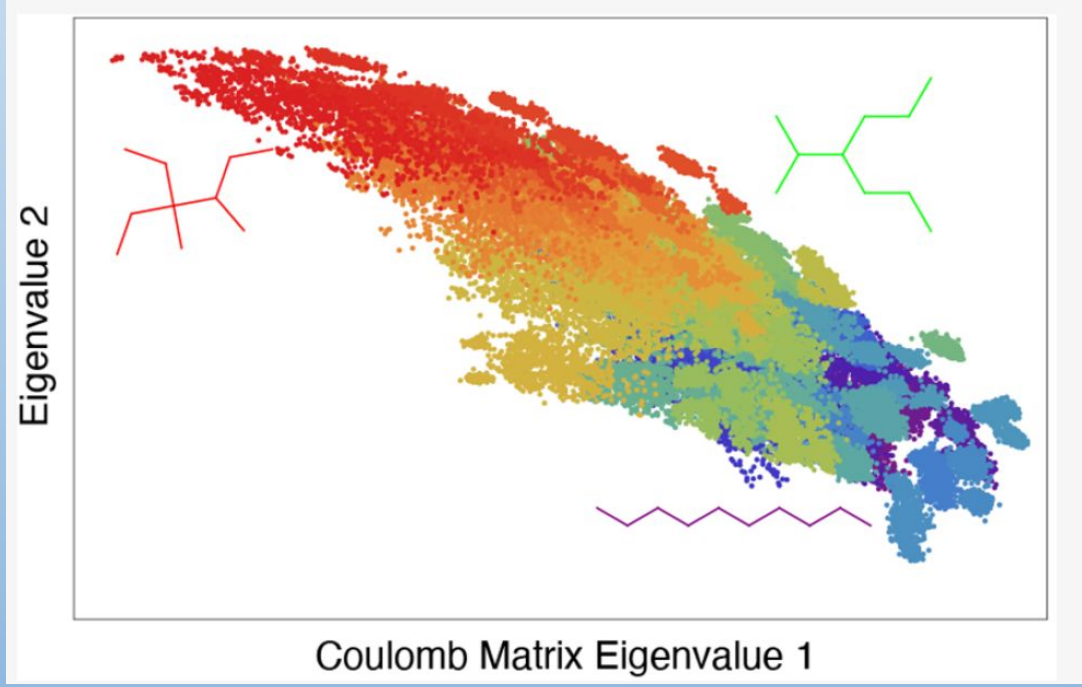


From: Seifert, N.A. *et al.*, *J. Chem. Phys.*, **155**, 184101 (2021).
See also: Seifert, N.A. *et al.*, *J. Chem. Phys.*, **156**, 134117 (2022).

Papers referenced describe a way to estimate the “distances” between two spectra, used for gauging similarity. Typically we’ve relied on frequency mapping while disregarding intensity - this is throwing away information!

Molecule Representations: Coulomb Matrix Eigenvalues

$$M_{ij} = \begin{cases} 0.5Z_i^{2.4} \text{ for } i = j \\ Z_i Z_j / R_{ij} \text{ for } i \neq j \end{cases}$$

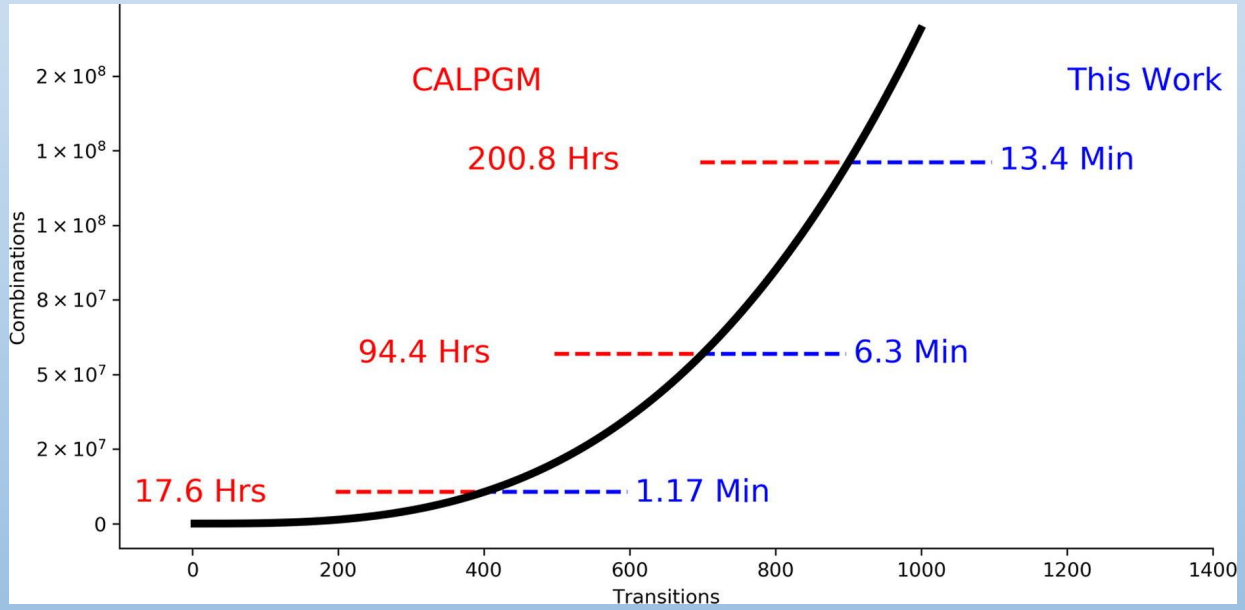


Coulomb Matrix introduced in: Rupp, M. *et al.*, *Phys. Rev. Lett.* **108**, 058301 (2012).

Figures from: Schrier, J., *J. Chem. Inf. Model.*, **60**, 3804 (2020).

Way of mapping N-atom molecule to a vector with N elements. Some indication that this will not be easily distinguishable as molecule size increases beyond 10 heavy atoms.

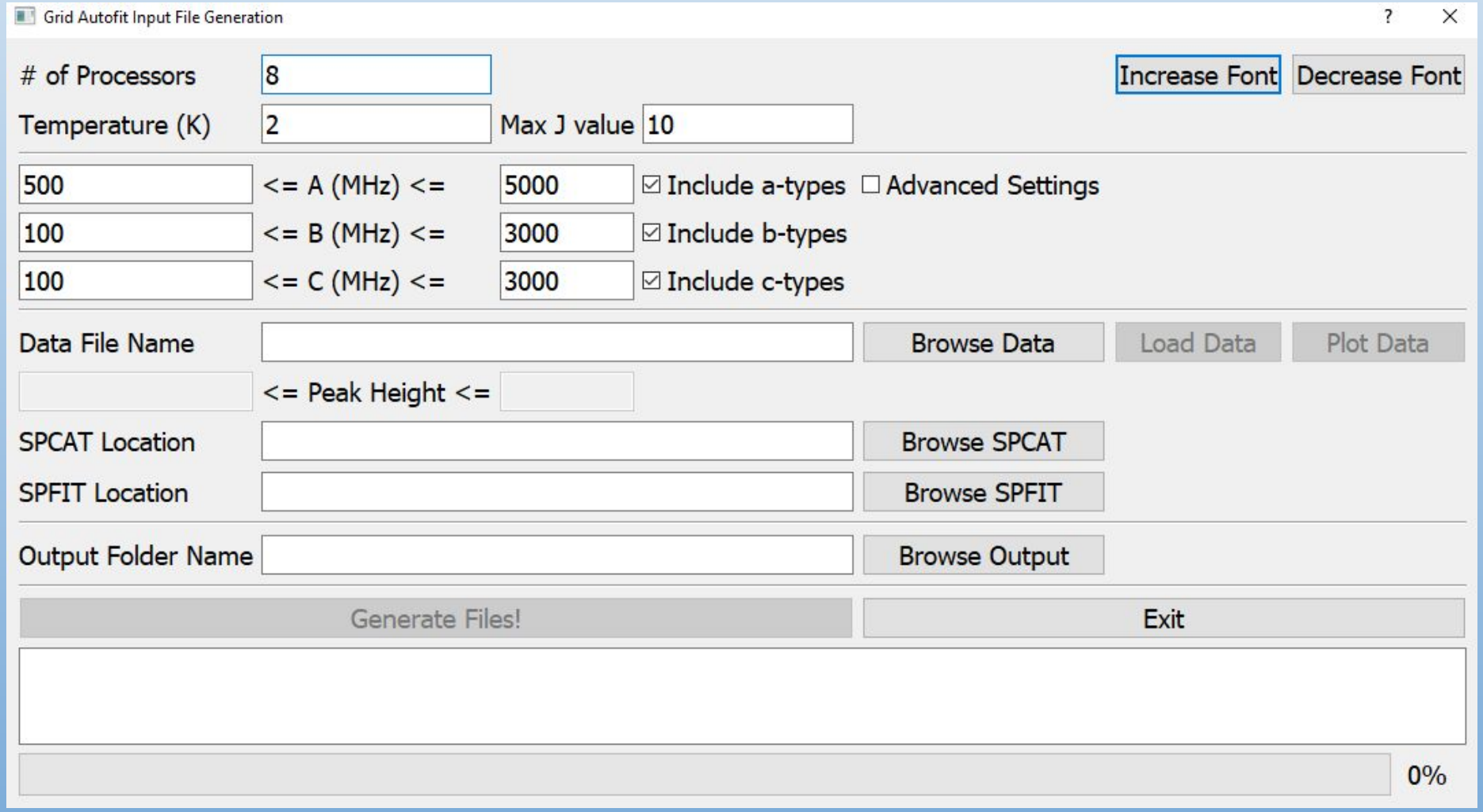
Fast Alternatives to CALPGM



Operation	This work	CALPGM	Speed
	(kHz)	(Hz)	Up
Fitting (100 Lines)	17.9	155.9	114.8
Fitting (10 Lines)	150.4	167.5	897.9
Spectrum Calculation (J = 25)	23.1	46.9	492.5
Spectrum Calculation (J = 15)	63.5	86.9	731.0

Carroll, P. *et al.*, *J. Mol. Spec.*, **379**, 111467 (2021)

Improved Usability / Transparency



Summary

- Primary problem in automated analysis of rotational spectra is solving the inverse problem to appropriate precision
- The general case is extremely challenging since a given data set might contain contributions from many different kinds of molecules, which might each need different Hamiltonians
- Some (but not many) attempts have been made to address this with machine learning, but not ready for general use; more development and testing is needed
- Ultimately, would be ideal to collaborate rather than for each group to develop their own approaches that are hard for others to use - maybe we should form a working group to coordinate?

Acknowledgements

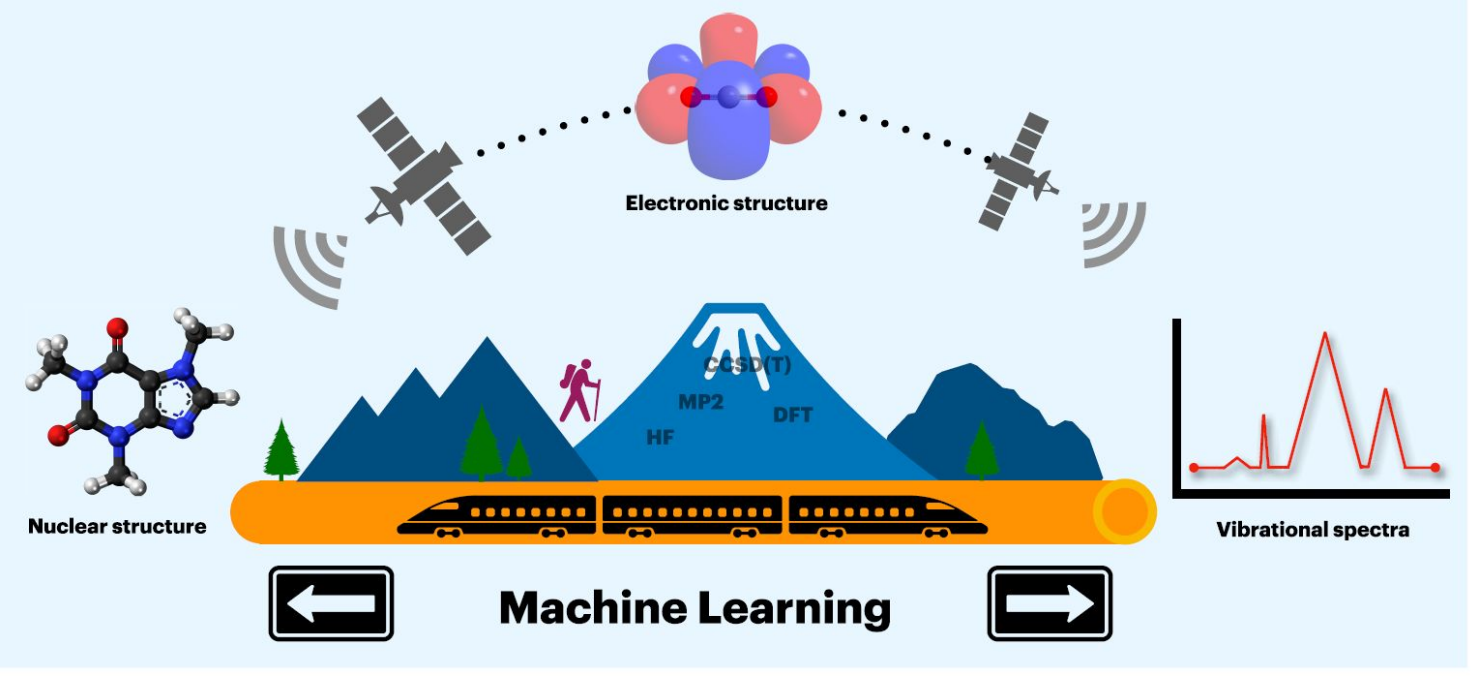
- Brooks Pate, Susanna Widicus Weaver, Gordon Brown, Kyle Crabtree, and Brandon Carroll

- Former students:
  Noah Anderson, Katherine Ervin, Ian Finneran, Erika Johnson, Aaron Olinger, Maria Phillips, Erika Riffe, Wes Westerfield, Andi Wright

Where Might ML Help?

Would love to go from spectra directly to structure, but would still be useful to go from spectra to rotational constants.

Techniques that could yield electronic structure without needing full *ab initio* calculations would also be helpful



From: Han, R., Ketkaew, R., Luber, S., *J. Phys. Chem. A* **126**, 801 - 812 (2022).

How Bad Is The Problem?

Extremely naive brute force approach:
Pre-calculate all spectra and use a lookup table

A: 1000 - 20000 MHz, steps of 5 MHz
B: 500 - 10000 MHz, steps of 5 MHz
C: 500 - 10000 MHz, steps of 5 MHz

Five distortion constants, ΔJ is 0 to +100 kHz, rest are -100 to +100 kHz (steps of 5 kHz)

About 2.76×10^{17} spectra; assume 5 MB .cat file:
"Just" do lookup in a dataset of size 1.35×10^{15} GB
(8.24×10^{10} 16 TB hard drives, about $300 each)

At 1 sec per spectrum calculation, would take 8.7×10^{9} CPU-years

(This doesn't include hyperfine, internal rotation, spin-rotation, etc.)