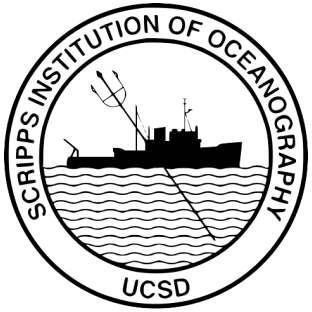
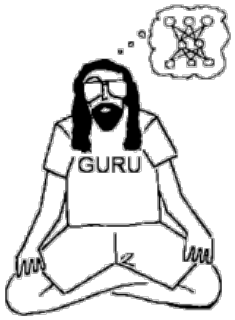




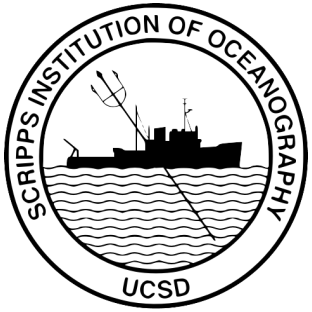
SPECTRE: A Spectral Transformer for Molecule Identification

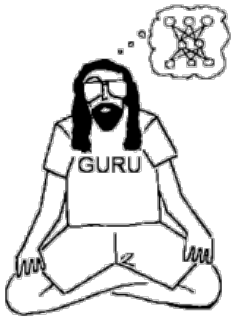
A collaboration between the Gerwick & GURU labs

Gary Cottrell, CSE

With Wangdong Xu (CSE), Byeol Ryu (SIO), Huanru Henry Mao (Calclavia), Hyunwoo Kim (Dongguk U.), James Zhao (CSE), Chen Zhang (SIO), Anthony Tong (CSE), Yiran Xu (CSE).

and Bill Gerwick (SIO)





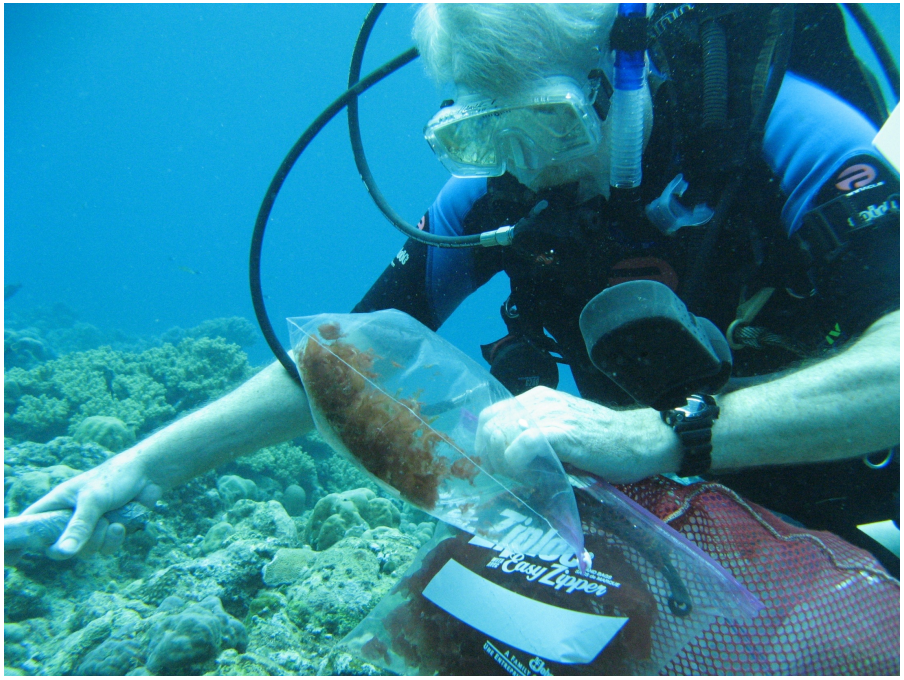
SPECTRE: A Spectral Transformer for Molecule Identification

A collaboration between the Gerwick & GURU labs

Gary Cottrell, CSE

With **Wangdong Xu** (CSE), **Byeol Ryu** (SIO), Huanru Henry Mao (Calclavia), Hyunwoo Kim (Dongguk U.), James Zhao (CSE), Chen Zhang (SIO), Anthony Tong (CSE), Yiran Xu (CSE).

and **Bill Gerwick** (SIO)



Bill Gerwick collecting a natural product

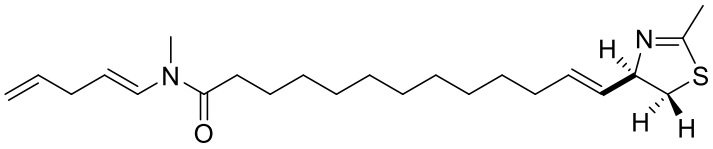
Outline

- **Motivation**
- SMART/DeepSAT
- SPECTRE
  - Methods
  - Results
- Conclustion

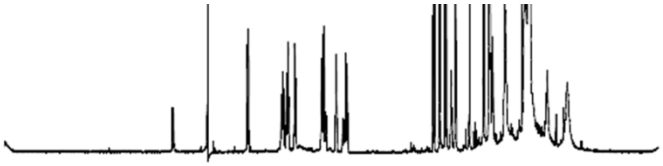
Motivation

- Natural products (NPs) comprise somewhere between 30-50% of drugs on the market
- The pipeline for novel NPs (after collection) starts with purification and structure determination
- 2D HSQC NMR is the preferred method for initial molecule structure elucidation
- But this interpretation step requires a great deal of human expertise, i.e., Bill Gerwick!

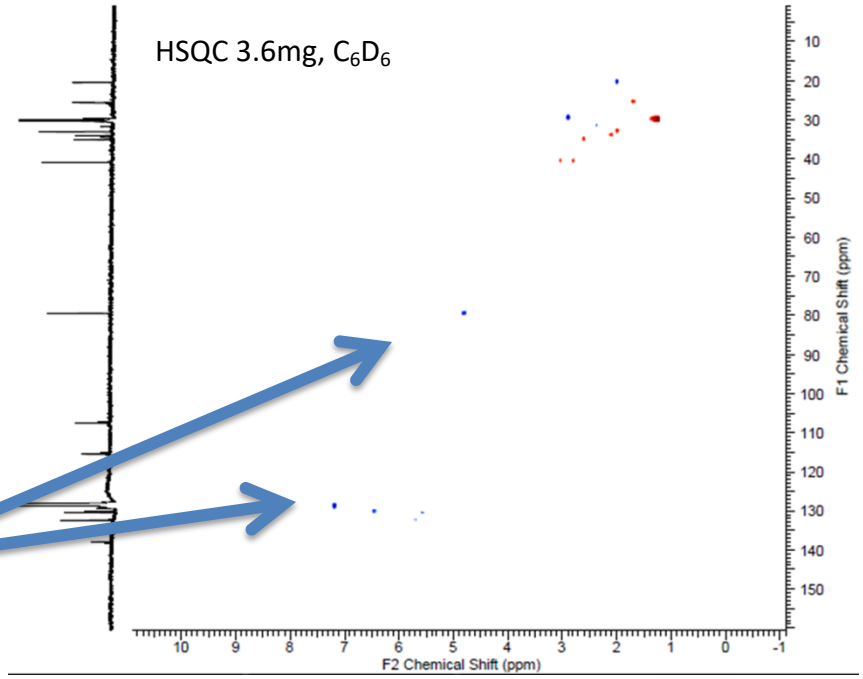
2D NMR is an Indicator of Compound Structure



laucysteinamide A

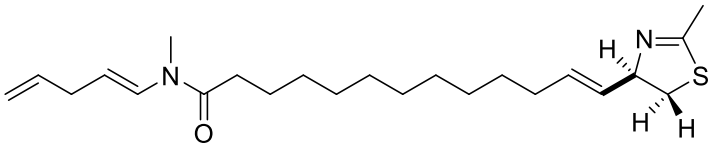


Caldora penicillata

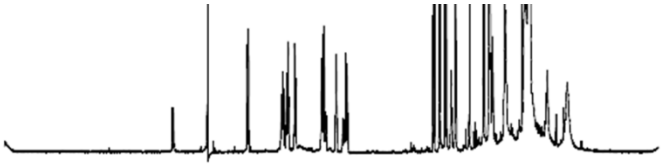


Each "dot" here corresponds to a bond between a hydrogen and a carbon

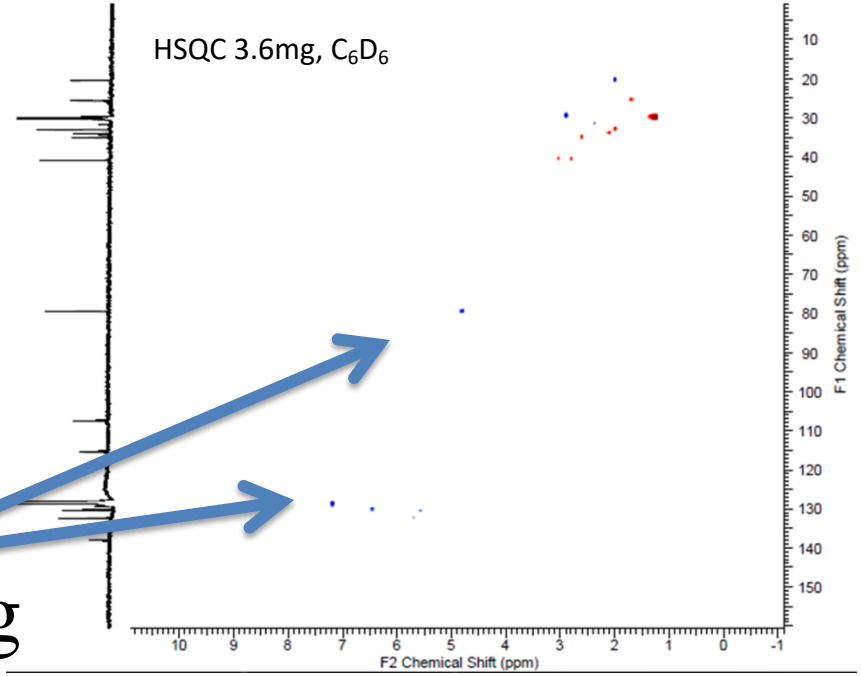
2D NMR is an Indicator of Compound Structure



laucysteinamide A

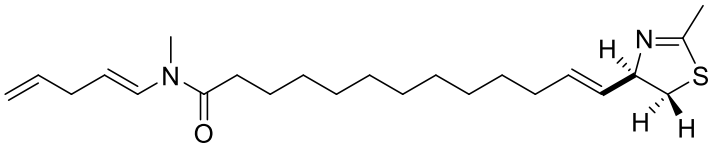


Caldora penicillata



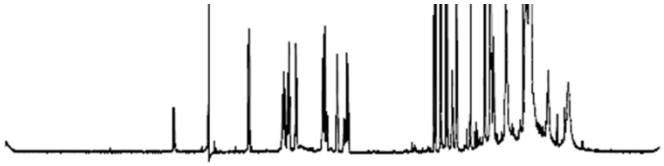
Where the dots appear depend on the neighboring atoms – this is called the "chemical shift"

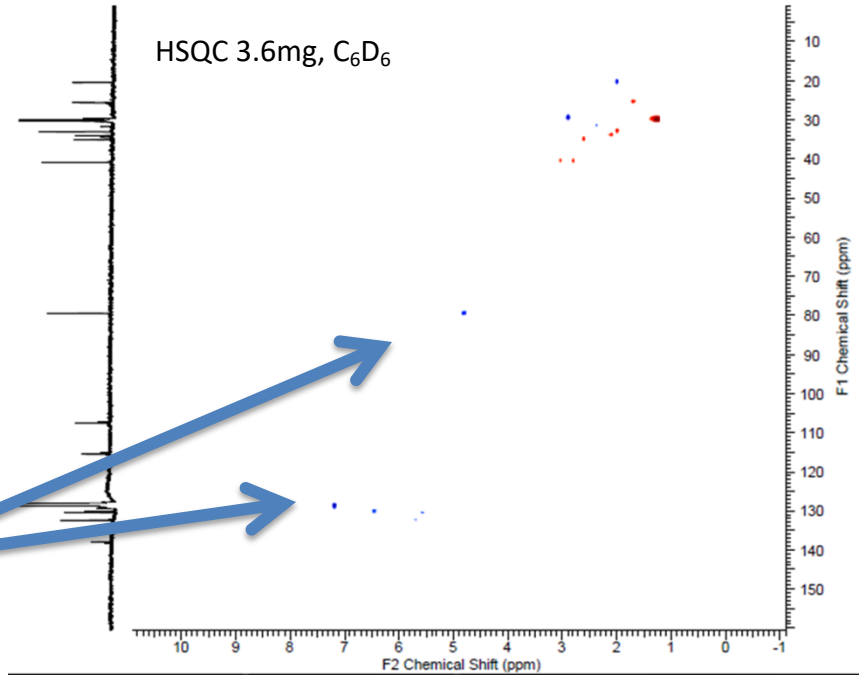
2D NMR is an Indicator of Compound Structure



laucysteinamide A

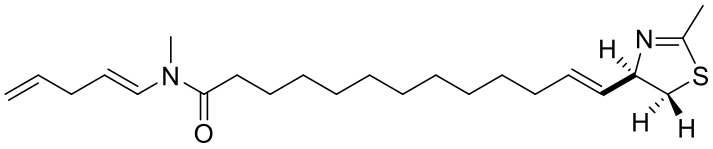
Caldora penicillata



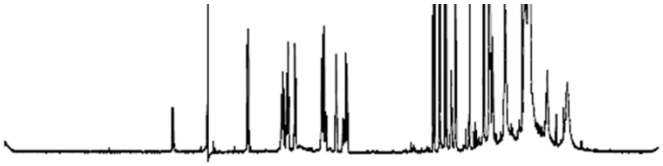


And now I've told you everything I know about NMR!!! This is why it's good to have collaborators!

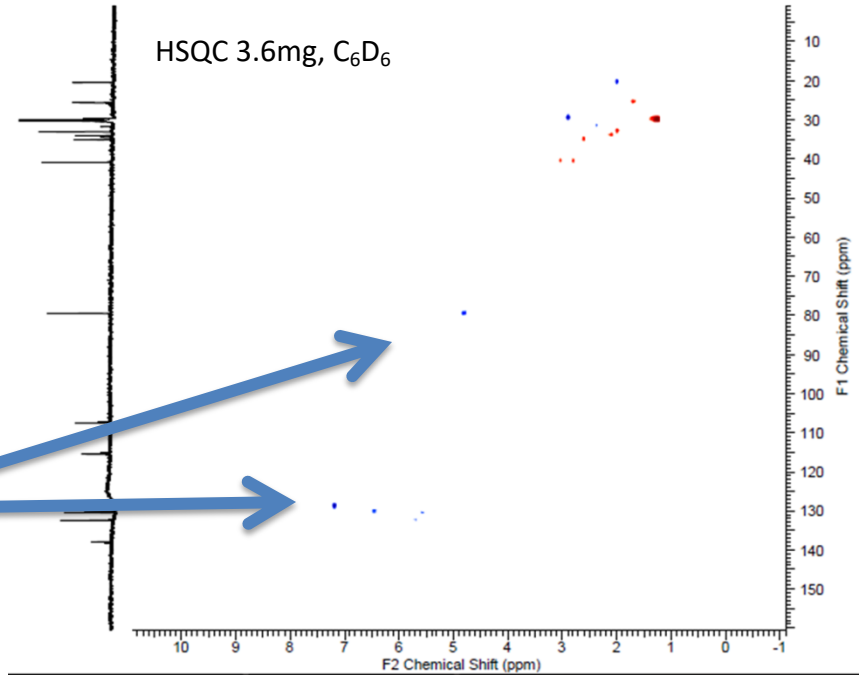
2D NMR is an Indicator of Compound Structure



laucysteinamide A



Caldora penicillata



An expert (e.g., Bill Gerwick) can look at this and say "Ok, looks like we have a methyl group"

We think of this as "the face of the molecule"

Outline

- Motivation
- **SMART/DeepSAT**
- SPECTRE
  - Methods
  - Results
- Conclusion

Enter Deep Learning

- Our student Chen Zhang had the idea that what Bill was doing was like face recognition – so he came to me.
- We started the SMART project – treating the 2D NMR as an image, and using Convolutional Neural Networks to map that image to a cluster space where similar compounds had similar locations in the space.
- Given a new compound, nearby points in the space suggested possible structures

We created a sequence of better and better models…

SCIENTIFIC REPORTS

JOURNAL OF NATURAL PRODUCTS

Article

Cite This: J. Nat. Prod. 2020, 83, 617–625

pubs.acs.org/jnp

Received: 2
Accepted: 2
Published o

Pa
Ch
Ne

Yue
Che
Tho

Kim et al. Journal of Cheminformatics (2023) 15:71
https://doi.org/10.1186/s13321-023-00738-4

Journal of Cheminformatics

SOFTWARE

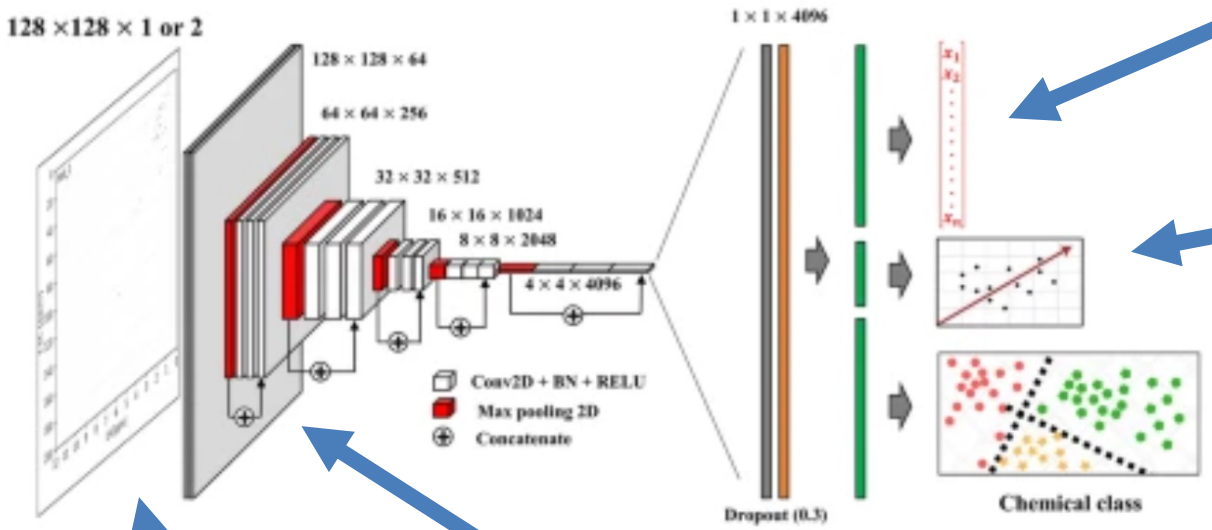
Open Access

Check for updates

DeepSAT: Learning Molecular Structures from Nuclear Magnetic Resonance Data

Hyun Woo Kim[1,2], Chen Zhang[1,3], Raphael Reher[1,4], Mingxun Wang[5,6,7], Kelsey L. Alexander[1,8], Louis-Félix Nothias[9], Yoo Kyong Han[10], Hyeji Shin[10], Ki Yong Lee[1,10], Kyu Hyeong Lee[2], Myeong Ji Kim[2], Pieter C. Dorrestein[5], William H. Gerwick[1,5*] and Garrison W. Cottrell[3*]

DeepSAT Network architecture: Supervised multi-task CNN



Morgan Fingerprint

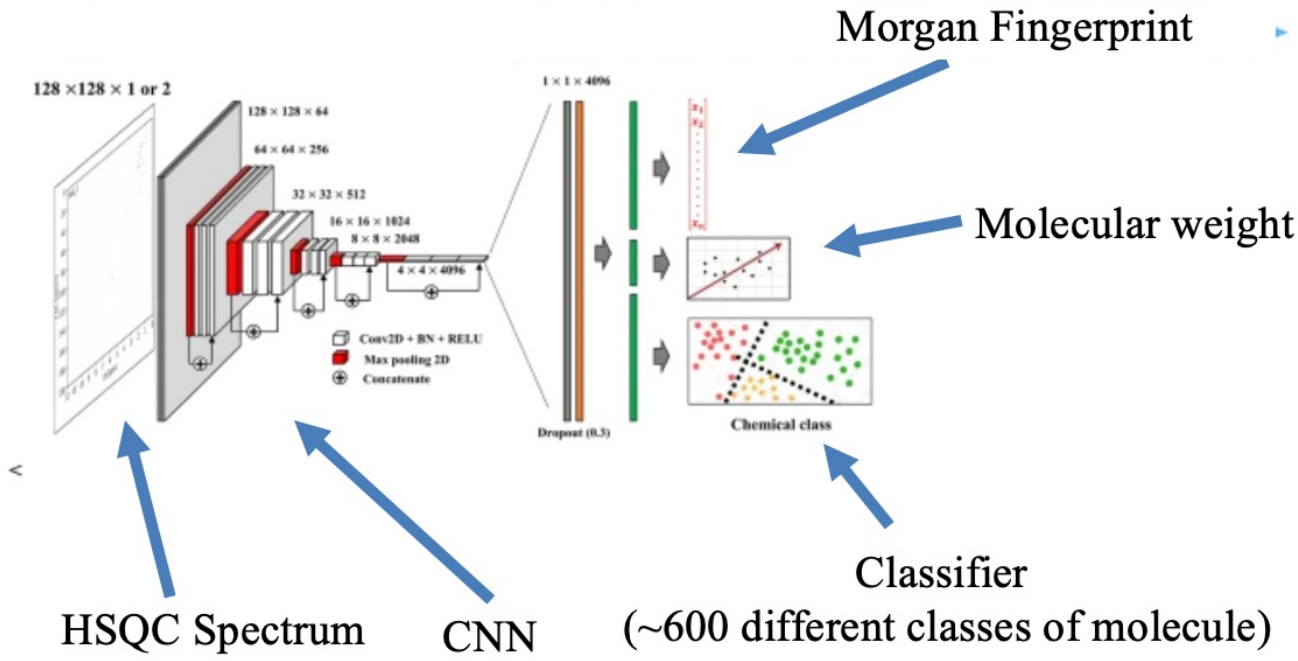
Molecular weight

HSQC Spectrum

Convolutional Neural Network

Classifier (~600 different classes of molecule)

DeepSAT Network architecture: Supervised multi-task CNN

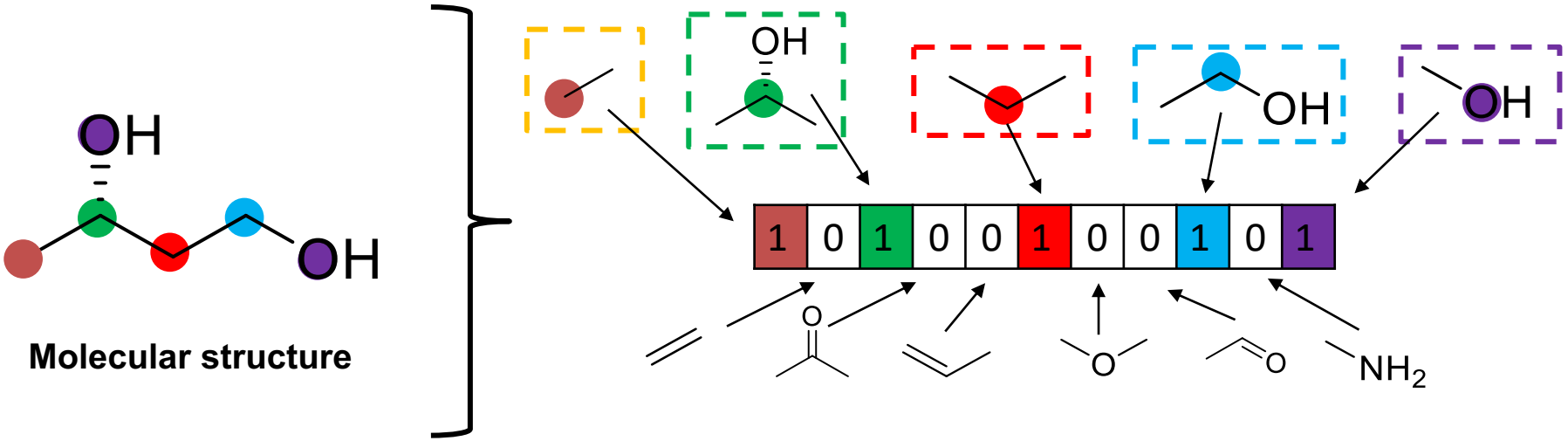


Main output: Morgan Fingerprints, a vector-based representation of molecular structure

These are compared to a database of MFs, and a list of similar molecules are returned

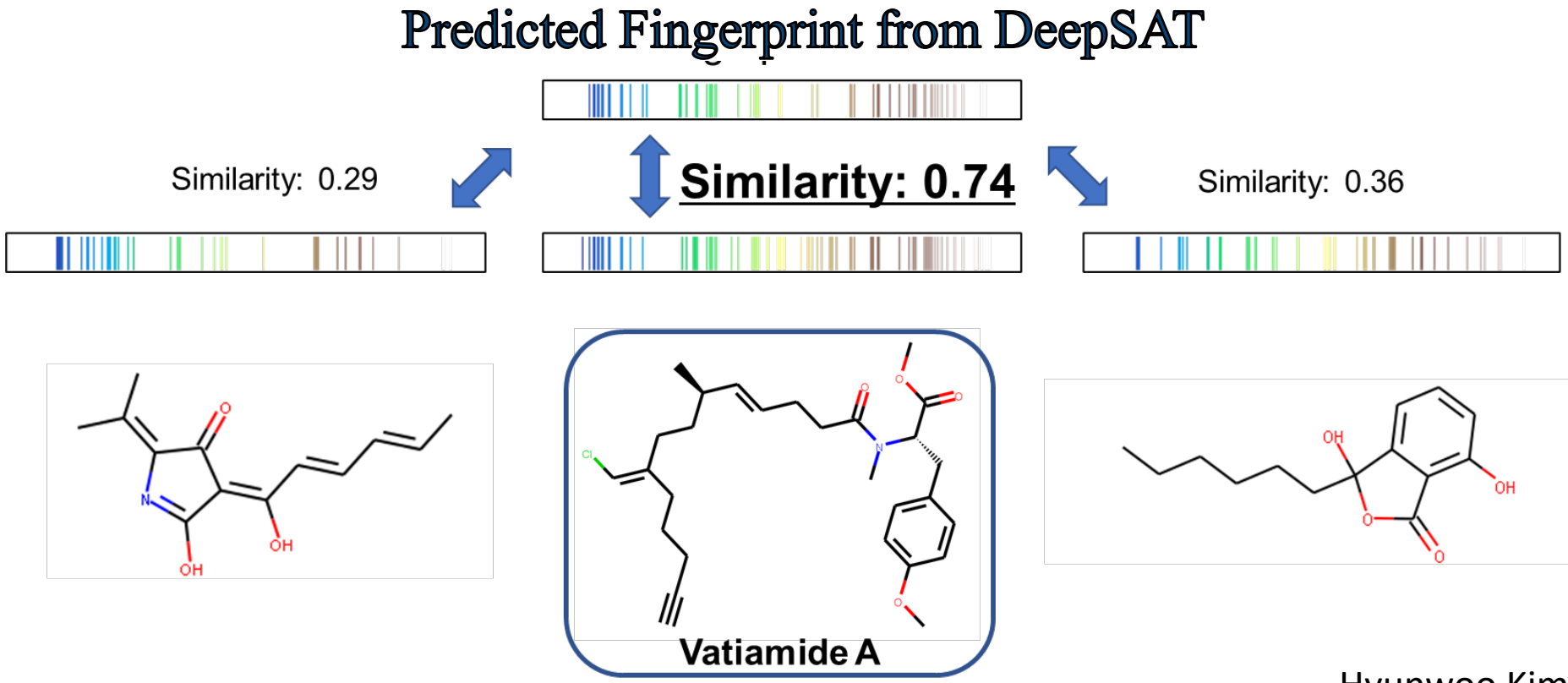
Morgan fingerprints

- Morgan Fingerprints are a vector-based representation of molecular structure used in many computational tools for cheminformatics



- DeepSAT uses a 6,144 bit vector, with each position associated with a specific partial structure
- A “1” means that this substructure is in the molecule,
- A “0” means it’s absent

DeepSAT – Using Morgan-type Fingerprints to Determine Chemical Similarity



Hyunwoo Kim

This result gives the researcher clues to the chemical structure of a novel compound – speeding structure identification

Outline

- Motivation
- SMART/DeepSAT
- **SPECTRE**
  - Methods
  - Results
- Conclusion

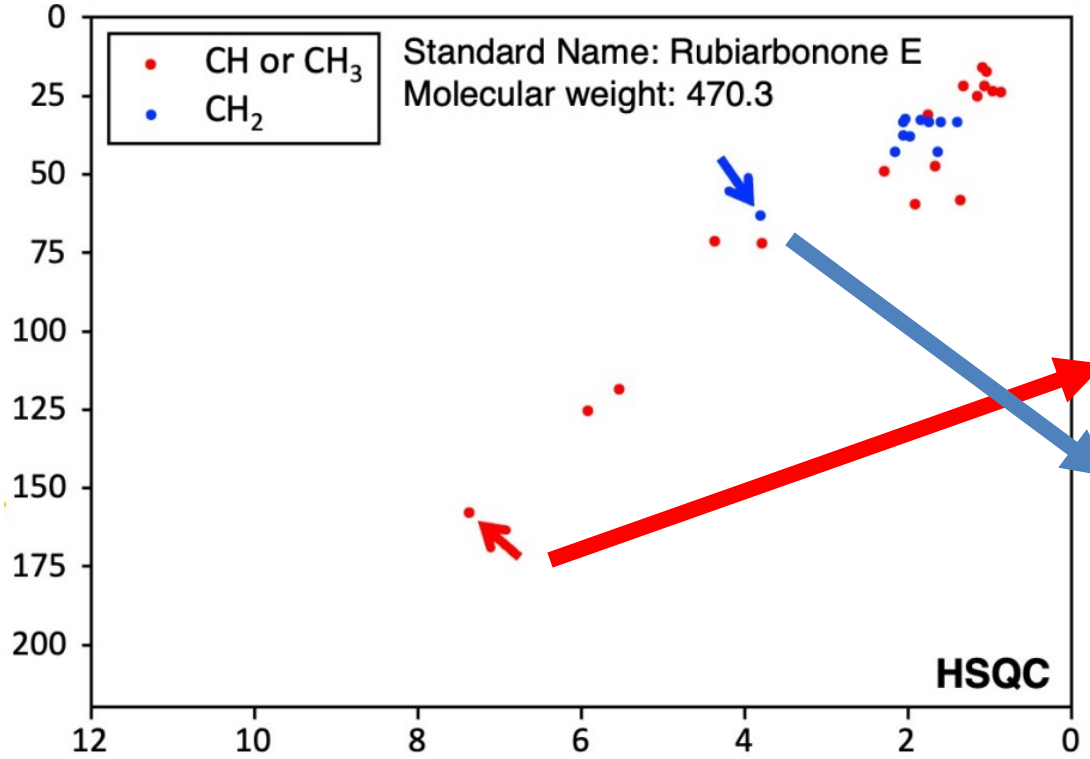
What's wrong with this picture?

- These models have over 8k users and over 450k queries.
- But they're limited and inefficient for three reasons:
  - They only take 2D HSQC NMR as inputs
    - Inflexible
  - Over 99% of the pixels are zero!
    - A lot of wasted compute.
  - The target, Morgan Fingerprints, are a hash table
    - Collisions: Locations in the table are *ambiguous*

The First Main Idea of this talk: Transformers

- A transformer is what underlies ChatGPT:
  - It takes words as input
  - It processes those words in the context of other words to extract meaning
- Instead of words, we give it the only the (x,y) locations of the peaks
  - much more efficient (no wasted computation)
  - It processes the peaks in the context of the other peaks
- Just like DeepSAT, we train it to produce Morgan Fingerprints from thousands of examples

The First Main Idea of this talk: Transformers



Input format to Transformer

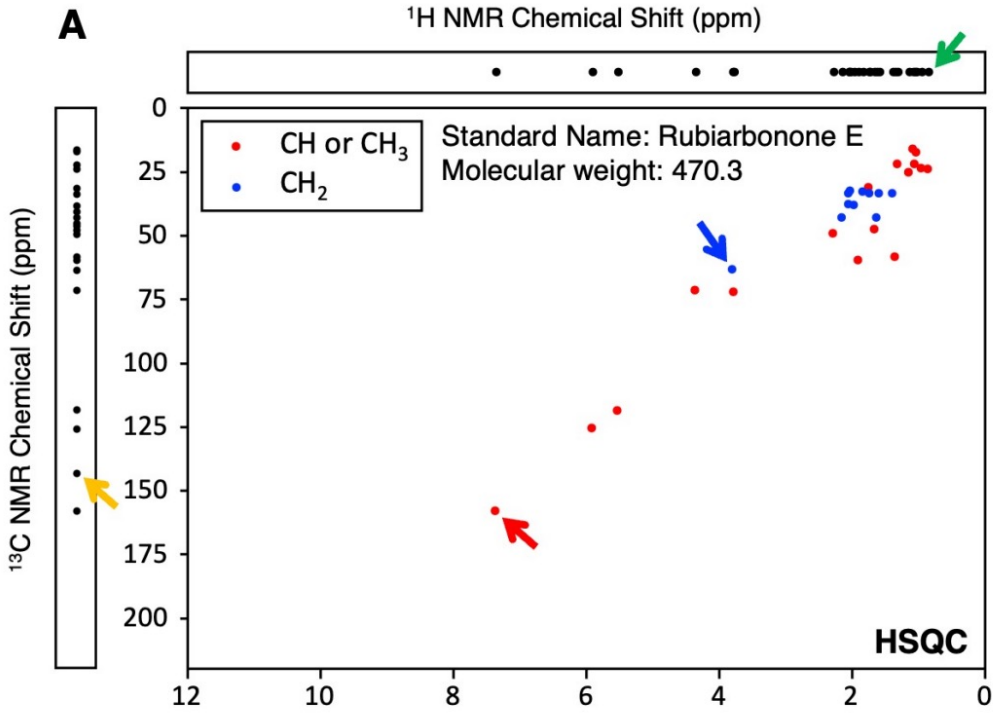
Peak	type
[157.7, 7.3, 1],	HSQC
…	…
[63.4, 3.0, -1],	HSQC

Peaks are encoded by their chemical shift as well as their multiplicity - odd or even # of bonds

The Second Main Idea of this talk: Flexible Inputs

- A transformer is what underlies ChatGPT:
  - It takes words as input
  - It processes those words in the context of other words to extract meaning
- Instead of words, we give it the (x,y) locations of the peaks
  - much more efficient!
- We can *also* give it other data, tagged by its type:
  - 1D ^{13}C NMR peaks: (C,x_1), (C,x_2), (C,x_3),… (C,x_{NC})
  - 1D ^{1}H NMR peaks: (H,x_1), (H,x_2), (H,x_3),…, (H,x_{NH})
  - Molecular weight: (MW,470.3)
- We train it by *randomly choosing what data types to give it as input* - over many training trials, *it learns to use whatever data is available*

The Second Main Idea of this talk: Flexible Inputs



Input format to Transformer

With 1D NMR
When 1D NMR is available

Peak	type
[157.7, 7.3, 1],	HSQC
…	…
[63.4, 3.0, -1],	HSQC
…	…
[143.7, 0, 0],	¹³C NMR
…	…
[0, 0.92, 0],	¹H NMR
…	…
[470.3, 0, 0],	MW

Without 1D NMR
When 1D NMR is NOT available

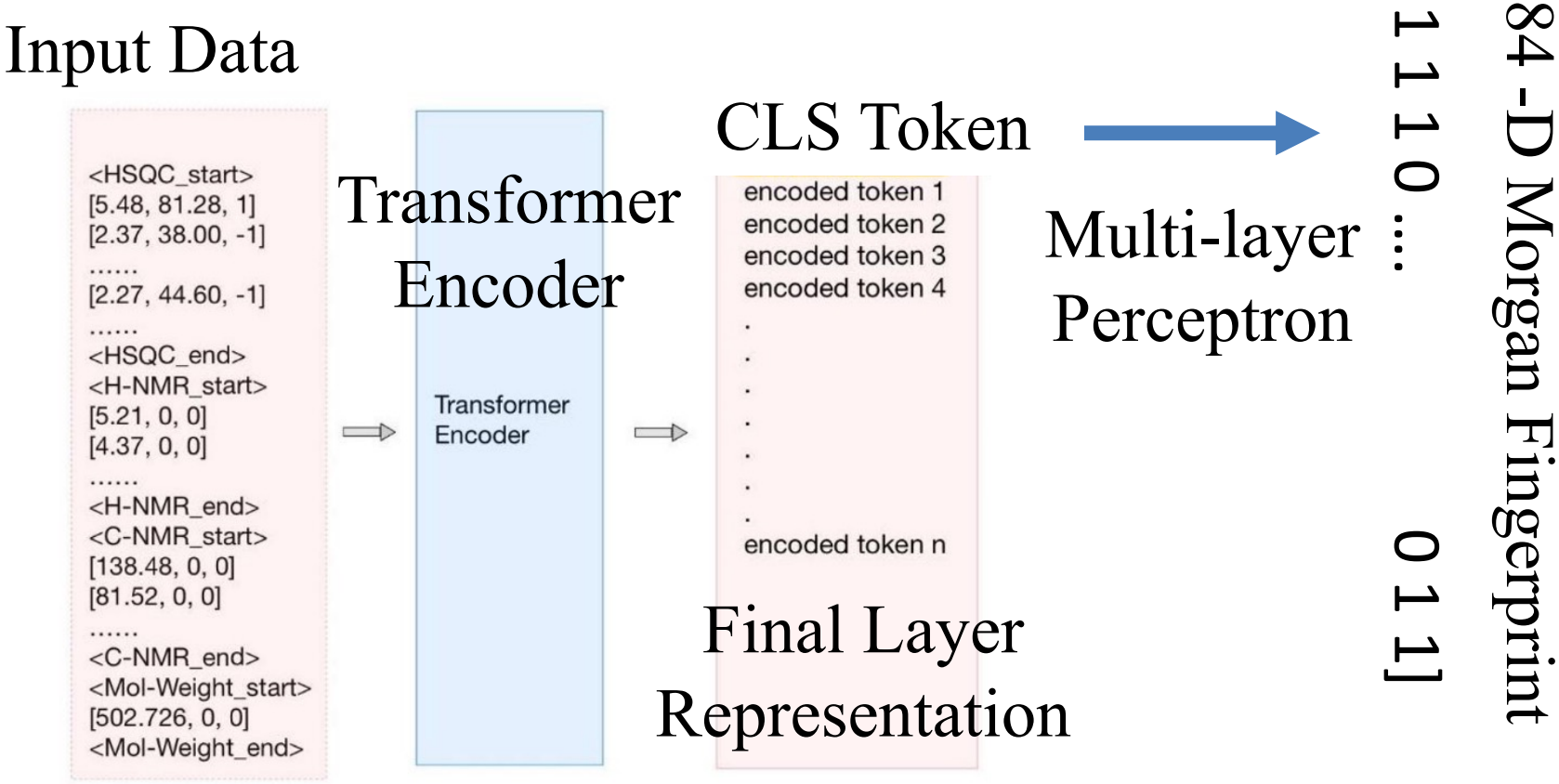
Peak	type
[157.7, 7.3, 1],	HSQC
…	…
[63.4, 3.0, -1],	HSQC
…	…
[470.3, 0, 0],	MW

- Now we can give it *multiple data types* – 2D NMR, 1D NMR, Molecular weight
- We train it by randomly giving it different types of data for each example – one time, it might just have 2D HSQC and 1D Carbon NMR, other times, just 1D Carbon, etc.

The Third Main Idea of this talk: Better Morgan Fingerprints

- We created Morgan Fingerprints up to radius 9 over a large dataset of molecules
- We sort them by their entropy, keeping the bits with the most information and label them with the substructure
- We keep 16,384 of these bits in the vector
- In our hands, these are collision-free: ever bit in the vector corresponds to a unique substructure
  - So we can *label* which parts of the retrieved structures match and which don't

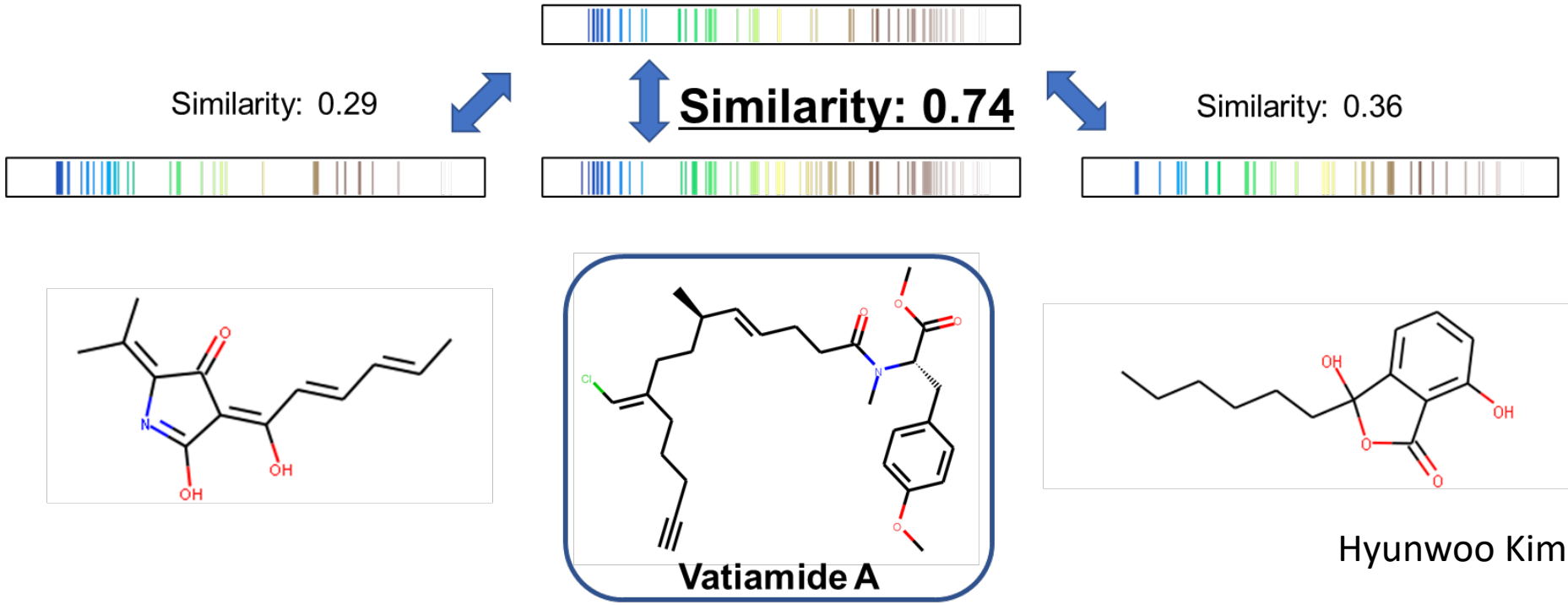
Network Architecture



Instead of predicting the next word, like ChatGPT, it is trained to predict our super-duper Morgan Fingerprint

Like DeepSAT, SPECTRE uses the super-duper Morgan Fingerprint to find similar molecules

Our retrieval set is very large - over **520,000 NP candidates**



Hyunwoo Kim

The result is a list of molecules ordered by their similarity to the predicted fingerprint – speeding structure identification

Data Collection & Training, Test, and Validation Set: We collected a LOT of data from publicly available datasets, as well as *predicting* data to train on

Data Collection

CH-NMR-NP

Literature NMR spectra
29,500 NPs + 6,000 organic compounds

ACD/Labs

Computed NMR spectra
ACD/Labs 113,967

^{1}H-^{13}C HSQC (n = 137,267)

NP-MRD

1D NMR spectra (^{1}H & ^{13}C)
Natural Products 155,815

1D NMR (n = 155,815)

COCONUT

LOTUS

Retrieval DB
526,316 NPs

chemical names, SMILES strings, and molecular weight

Training Set

	SPECTRE (with DTD)
Model Input	occurrence rate during training
All 3 NMR spectra types	7.6%
^{13}C NMR and ^{1}H NMR	12.2%
HSQC and ^{13}C NMR	7.6%
HSQC and ^{1}H NMR	7.6%
Only ^{13}C NMR	12.2%
Only ^{1}H NMR	12.2%
Only HSQC	40.4%

	Specialized Model (w/o DTD)
Model Input	training set size
All 3 NMR spectra types	33,203
^{13}C NMR and ^{1}H NMR	93,370
HSQC and ^{13}C NMR	39,685
HSQC and ^{1}H NMR	33,203
Only ^{13}C NMR	103,409
Only ^{1}H NMR	93,374
Only HSQC	109,694

Test Set
n = 4,096

Validation Set
n = 4,056

Chosen when all three NMR spectra (1H, 13C, and HSQC) were available

Results, quantitative

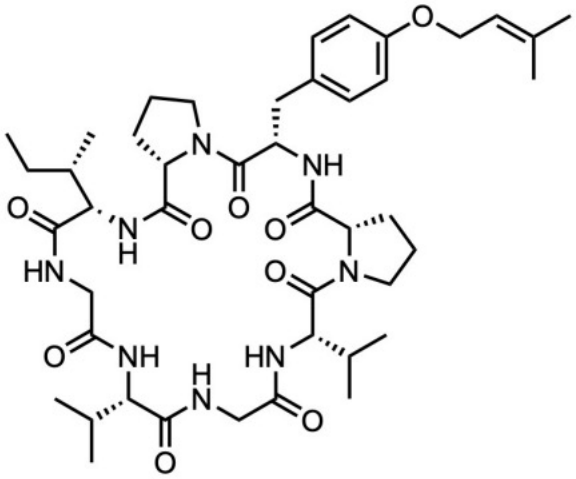
We compared SPECTRE against specialized models trained only on one type of data; here, we're measuring when the correct molecule is the top hit.
(in practice, we can apply the best model in each case)

Model Input Type	Top-1	
	SPECTRE	**Specialized**
ME-HSQC, ^{13}C NMR, and ^{1}H NMR	**79.78%±0.82%**	72.32 %±0.19%
Standard HSQC, ^{13}C NMR, and ^{1}H NMR	**78.41%±0.77%**	69.61 %±0.36%
ME-HSQC and ^{13}C NMR	**79.81%±0.47%**	73.14 %±0.33%
Standard HSQC and ^{13}C NMR	**77.98%±0.76%**	69.57 %±0.42%
ME-HSQC and ^{1}H NMR	**75.65%±0.82%**	65.50 %±1.95%
Standard HSQC and ^{1}H NMR	**72.83%±0.98%**	61.88 %±2.01%
ME-HSQC	74.25%±0.79%	**76.52 % ±0.40%**
Standard HSQC	70.20%±0.90%	**72.83 % ±0.23%**
^{13}C NMR and ^{1}H NMR	59.36%±0.90%	**68.91 %±0.65%**
^{13}C NMR	51.96%±0.73%	**57.59 %±0.26%**
^{1}H NMR	15.38%±0.25%	**19.63 %±0.46%**

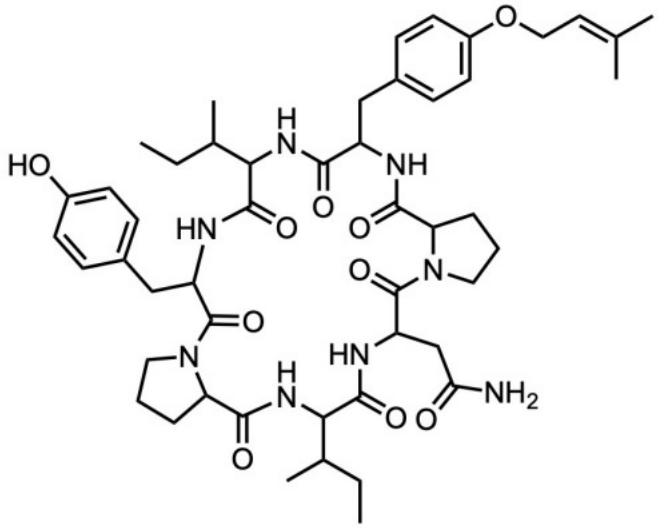
Results, qualitative

Here, using Multiplicity-Edited HSQC as input

A. Multiplicity-Edited HSQC



Monchicamide I (input)

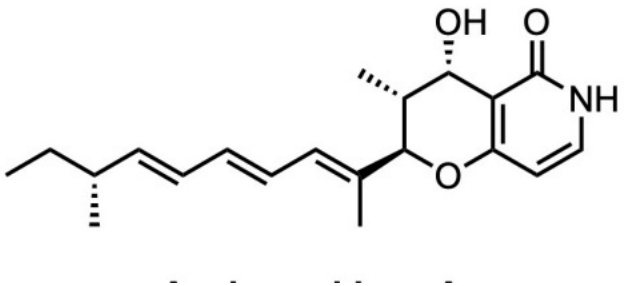


Closest retrieved molecule

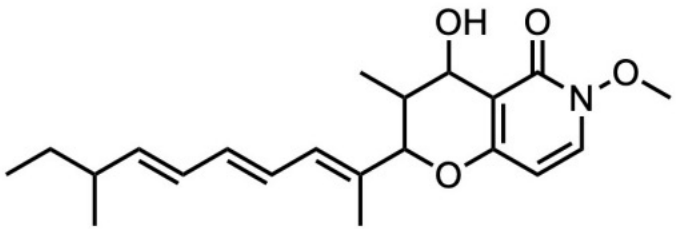
Results, qualitative

Using only Standard HSQC as input

B. Standard HSQC



Aculeapuridone A (input)

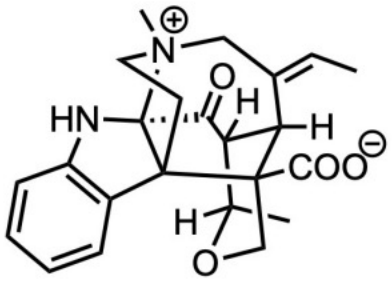


Closest retrieved molecule

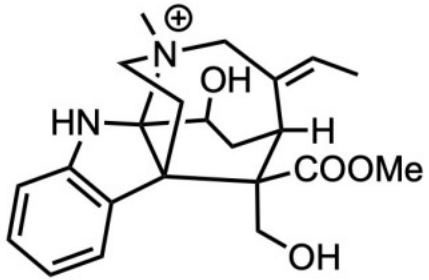
Results, qualitative

Here, using only ^{13}C NMR as input

C. ^{13}C NMR



Alstolarsine A (input)

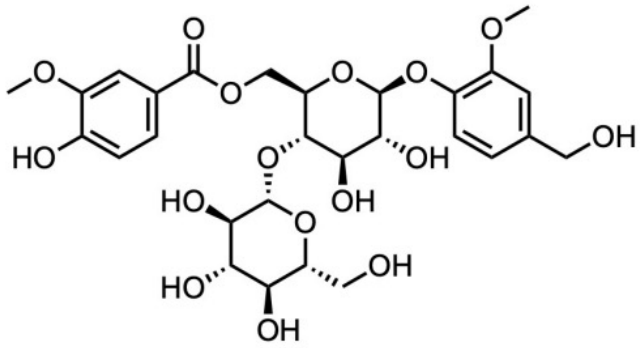


Closest retrieved molecule

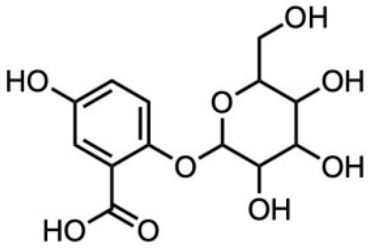
Results, qualitative

Here, using only ^{1}H NMR as input
(doesn't work well)

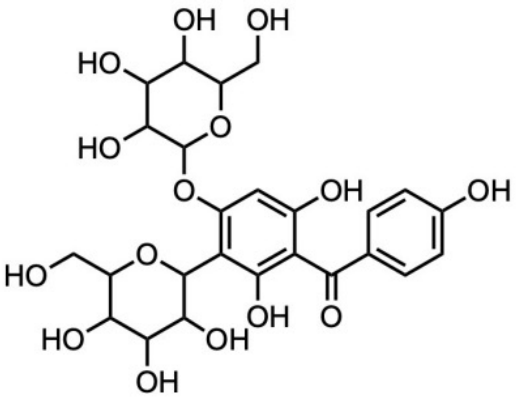
D. ^{1}H NMR



Wrightioside A
(input)



1st Closest
retrieved
molecule

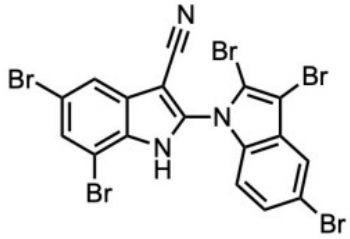


2nd Closest
retrieved
molecule

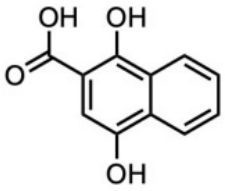
Results, qualitative

Proton-deficient compounds:

The power of multiple input data types



Aetokthonotoxin (input)

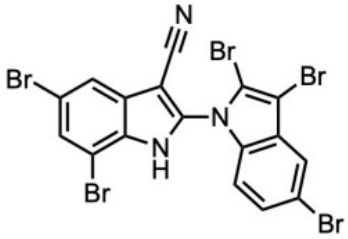


Using only ME-HSQC

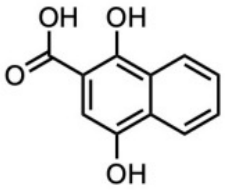
Results, qualitative

Proton-deficient compounds:

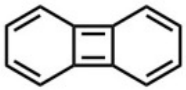
The power of multiple input data types



Aetokthonotoxin (input)



Using only ME-HSQC

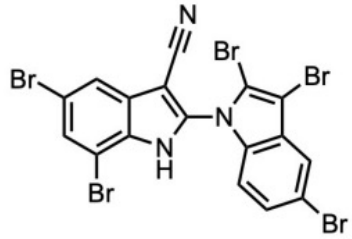


Using only ^{13}C

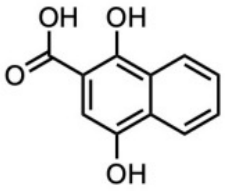
Results, qualitative

Proton-deficient compounds:

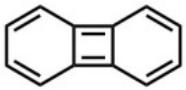
The power of multiple input data types



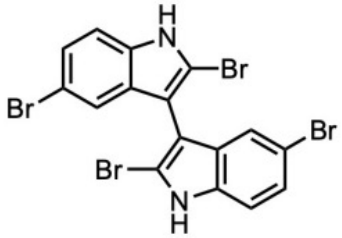
Aetokthonotoxin (input)



Using only ME-HSQC



Using only ^{13}C

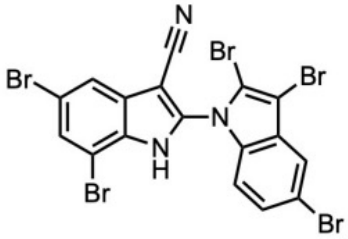


Using ME-HSQC + ^{13}C

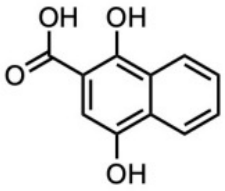
Results, qualitative

Proton-deficient compounds:

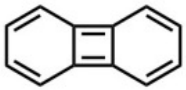
The power of multiple input data types



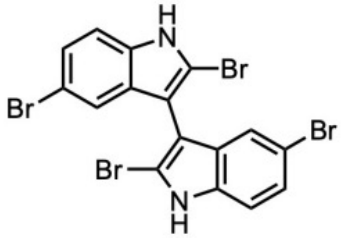
Aetokthonotoxin (input)



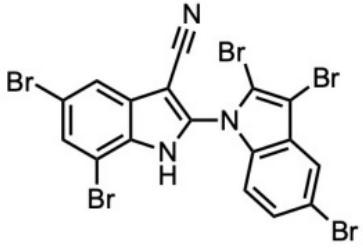
Using only ME-HSQC



Using only ^{13}C



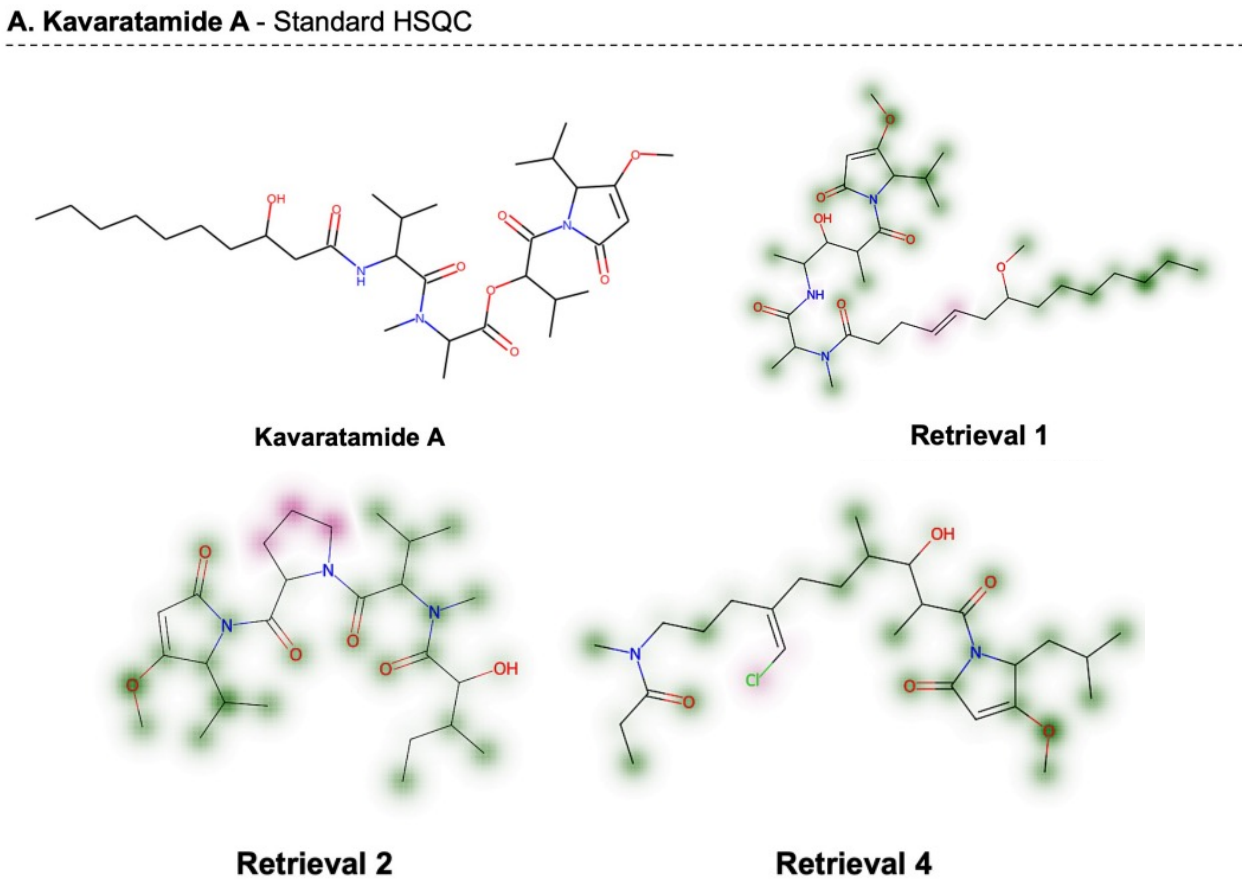
Using ME-HSQC + ^{13}C



Using ME-HSQC + ^{13}C + MW

A perfect match – the molecule has been "dereplicated"

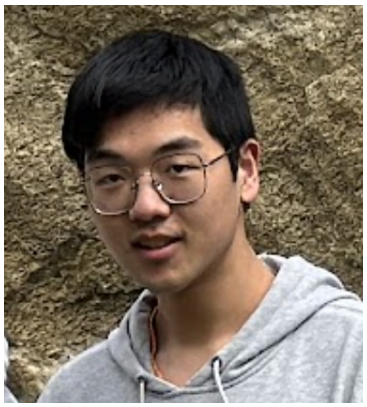
The advantage of unique bits in our entropy-optimized Morgan Fingerprints
We can mark what matches and what doesn't



Conclusions

- SPECTRE is a major advance over our previous work
- It automagically combines data from multiple sources to obtain the best result, given the amount of data provided
- It uses an advanced form of Morgan Fingerprint – we call "entropy-optimized Morgan Fingerprints"
- These allow highlighting of matching substructures – providing more information than just – hey, this is similar!
- It beats our previous state of the art model by a lot!

Thanks!!
To these very smart folks!


Wangdong Xu


Byeol Ryu


Henry Mao


Hyunwoo Kim


James Zhao


Chen Zhang


Anthony Tong


Yiran Xu


Ming Wang


Bill Gerwick

Funders: NIH GM118815, GM107550 Gordon and Betty Moore Foundation

And thank YOU for listening!

Questions?