Enhancing Peak Assignment in ^{13}C NMR Spectroscopy – A Novel Approach Using Multimodal Alignment

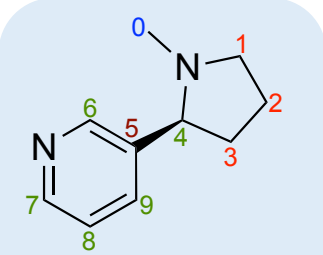
Dr. Hao Xu†, Zhengyang Zhou†, Dr. Pengyu Hong*

Department of Computer Science,
Brandeis University

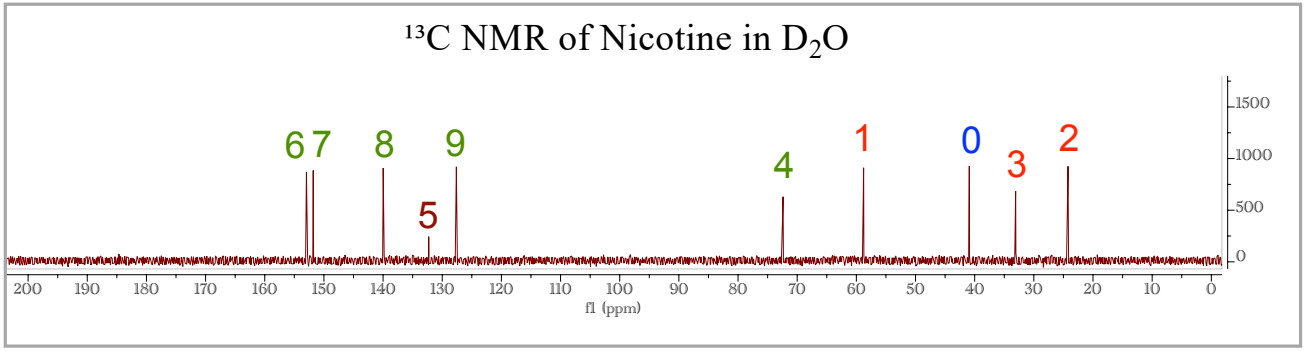
† Equal contribution

Introduction to ^{13}C NMR Spectroscopy

- Nuclear Magnetic Resonance (NMR) spectroscopy is a technique used to determine the structure of molecules by providing the environment (resonance frequency) of focused nuclei within a molecule, which is termed as chemical shifts and expressed in ppm (parts per million).
- ^{13}C NMR spectroscopy focuses on the ^{13}C isotope. ^{13}C chemical shifts, typically range from 0 to 200 pm, tell the positions of carbon atoms.

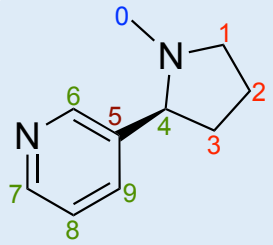


Nicotine



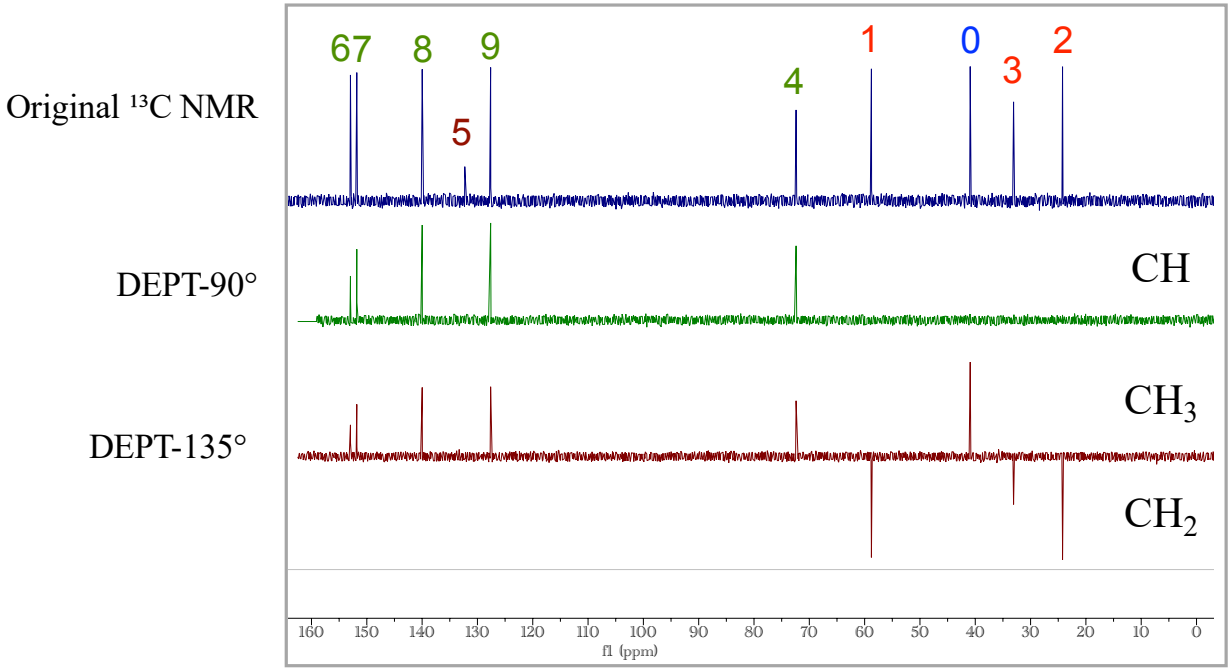
Introduction to ^{13}C NMR Spectroscopy

- DEPT (Distortionless Enhancement by Polarization Transfer) aids in determining the multiplicities of carbon atoms (CH, CH_2, and CH_3)



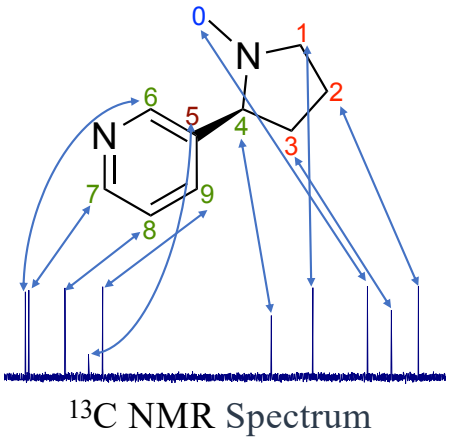
Nicotine

CH_3: 1° Carbon
CH_2: 2° Carbon
CH: 3° Carbon
C: 4° Carbon



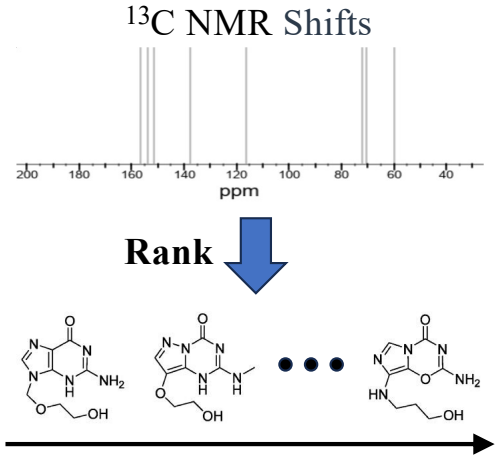
Opportunities

Peak Assignment

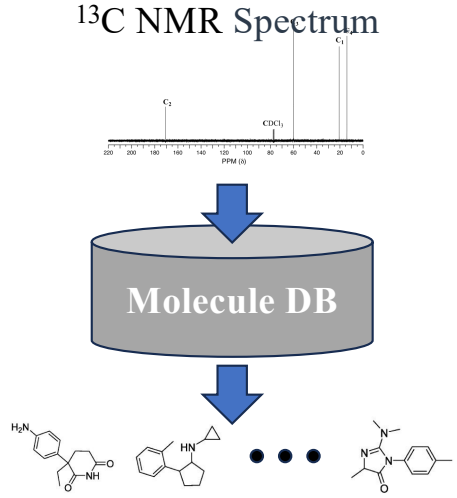


^{13}C NMR Spectrum

Candidate Ranking

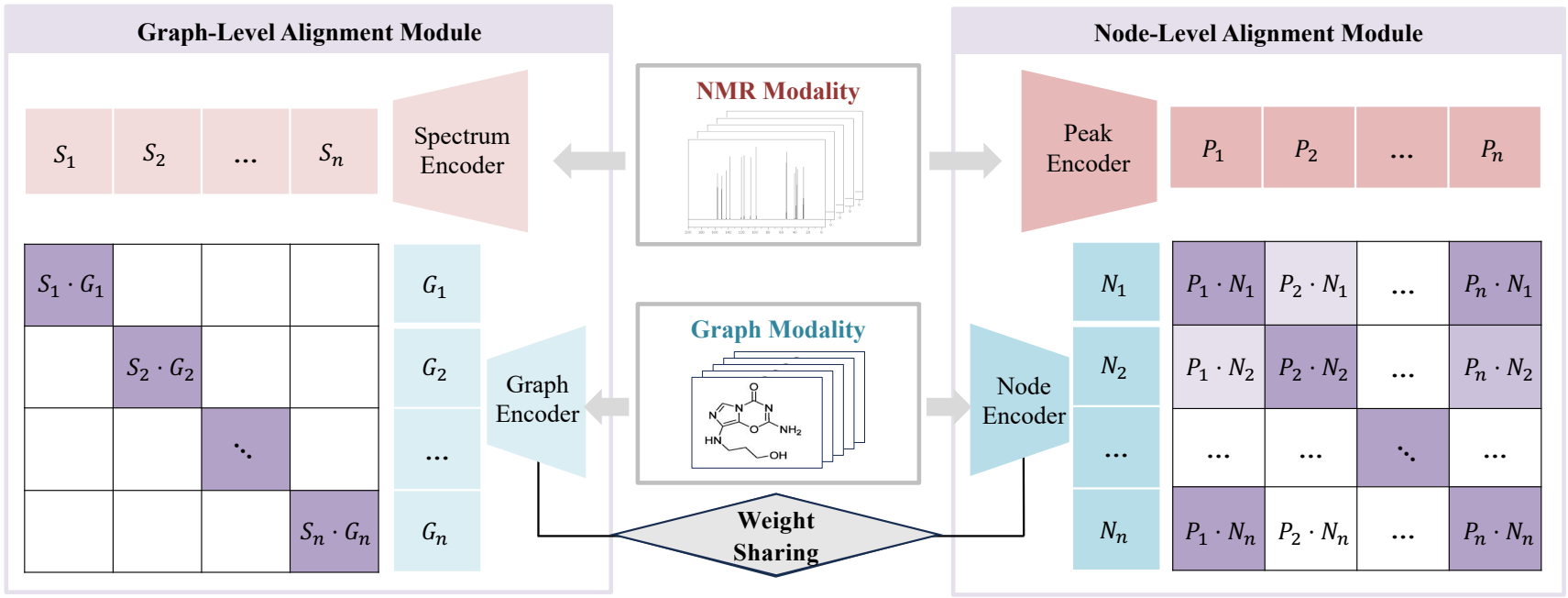


Molecule Retrieval



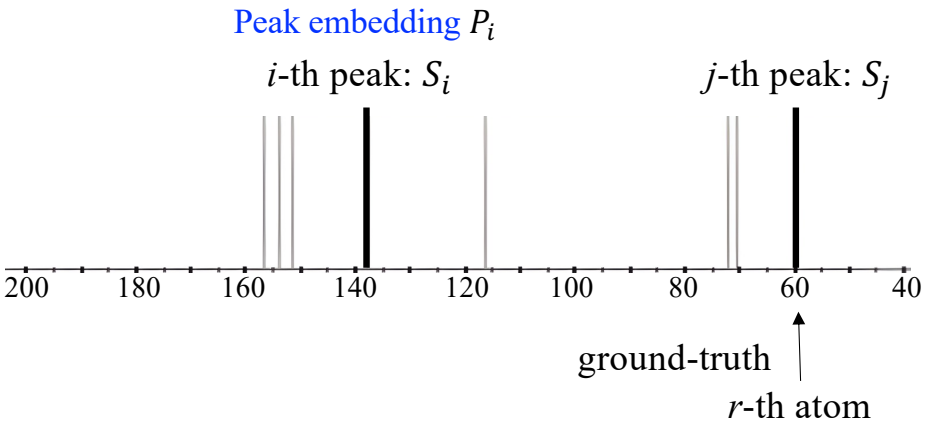
Propose: K-M^{3}AID (Knowledge-guided Multi-level Multimodal Alignment with Instance-wise Discrimination)

- Training inputs: (a) Molecular graphs: nodes represent atoms; edges represent chemical bonds
  (b) ^{13}C NMR spectra: peaks positions (ppm) and peak types (CH_3, CH_2, CH, or C)
- Training outputs: (a) Probabilities of molecule-spectrum matching
  (b) Probabilities of atom–peak alignments



Knowledge-guided Instance-wise Discrimination

Use domain knowledge to design soft contrastive learning
- More flexible than using binary negative/positive pairs
- Reduce potential bias in pre-defining positive/negative pairs



The similarity between the i-th peak (P_i) and the r-th node (N_m) should be "consistent" with the similarity between the i-th and j-th peaks

$$softmax(P_i \cdot N_r) \propto softmax\left(\frac{\tau_2}{|S_i - S_j| + \tau_1}\right)$$

where τ_1 and τ_2 are hyperparameters

Experiment Setting

- **Dataset:** ^{13}C NMR spectra of ~20k molecules in nmrshiftdb2*.

- **Contrast Learning Baselines:**
  - Strong Positive (**SP**): Pair (*i*-th atom, *j*-th peak) is positive **iff** the chemical shift of the *i*-th atom == the ppm of the *j*-th peak
  - Weak Positive (**WP**): Pair (*i*-th atom, *j*-th peak) is positive **iff** the difference between the chemical shift of the *i*-th atom and the ppm of the *j*-th peak ≤ a user-defined threshold (***th***).

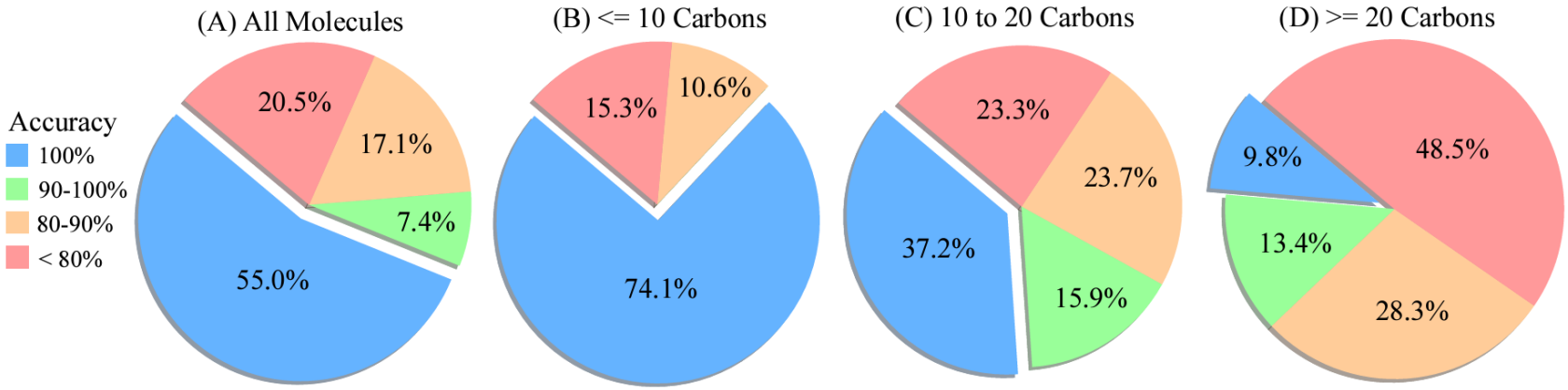
* nmrshiftdb2 (https://nmrshiftdb.nmr.uni-koeln.de/)

Experiments – Peak Assignment

Five-fold Cross-Validation Results

Alignment	SP	WP (*th* = 1)	WP (*th* = 5)	WP (*th* = 10)	K-M^{3}AID
Graph-Level	93.5 ± 0.6	91.3 ± 0.8	90.3 ± 0.6	88.4 ± 1.4	**95.5 ± 0.4**
Node-Level	89.3 ± 0.4	83.7 ± 0.6	79.8 ± 0.5	66.1 ± 2.5	**90.3 ± 0.1**

Validation results of K-M^{3}AID, stratified by the number of carbons

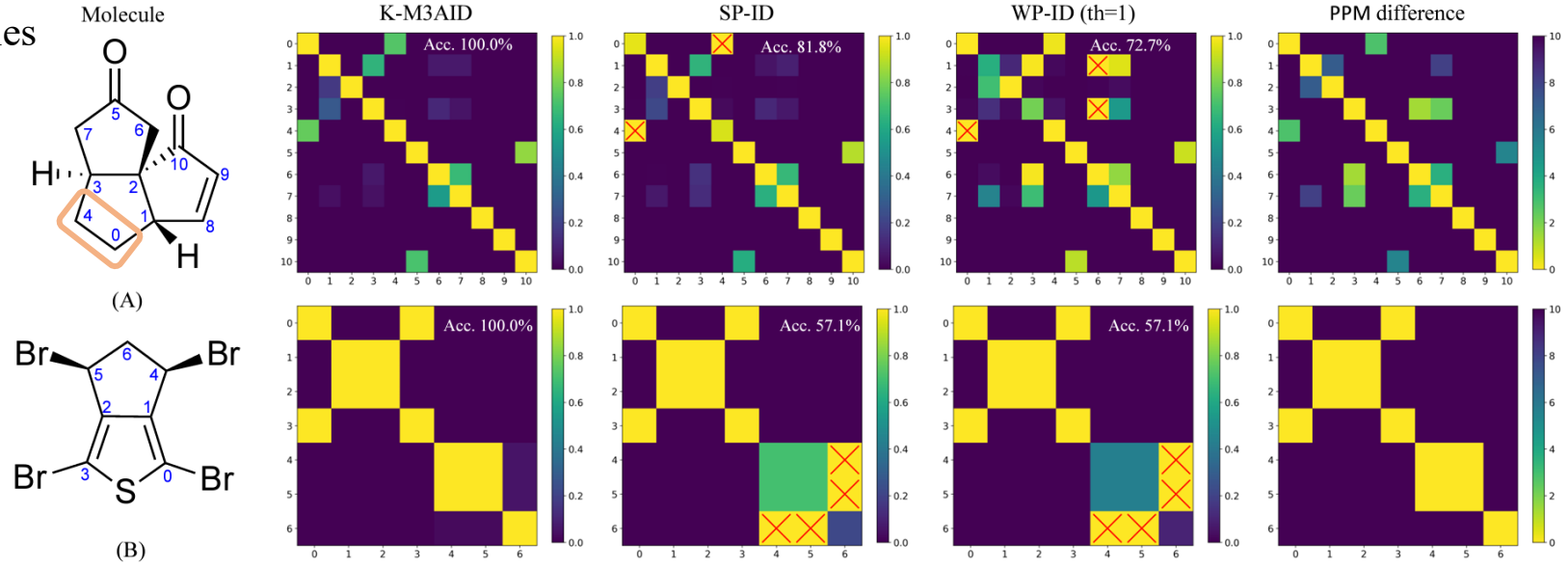


Experiments – Peak Assignment

K-M^{3}AID excels in two challenging scenarios:

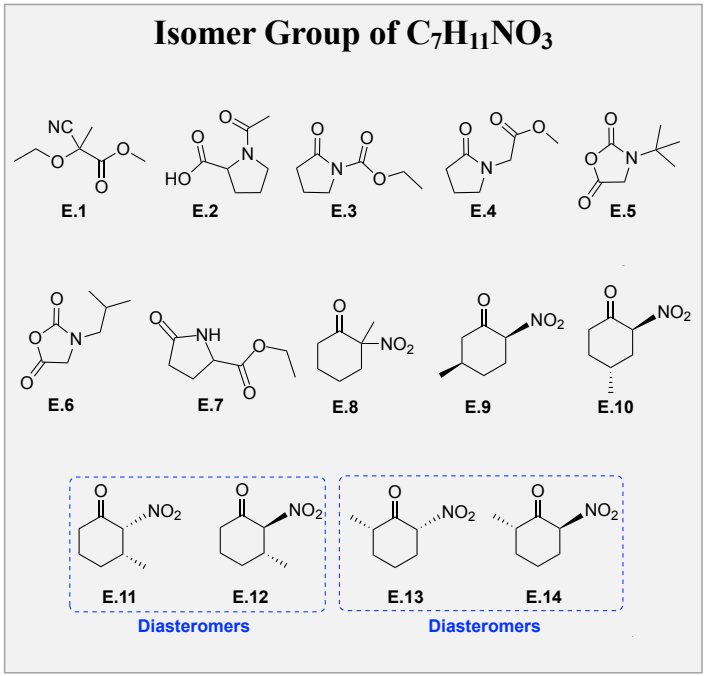
- Local contexts of carbons exhibit a high degree of similarity, such as C_0 and C_4 in the first example below, both are all secondary aliphatic carbons, next to tertiary carbons, and on the same ring.
- Carbons exhibit symmetry within the same molecule (second example below).

Examples



Experiments – Isomer Recognition

Formula	#Isomers	SP	WP (*th* = 1)	K-M^{3}AID
C_4H_6O	15	86.7	86.7	**100**
C_9H_9N	15	86.7	80.0	**100**
$C_7H_{11}NO_3$	14	78.6	85.7	**100**
$C_6H_{13}NO$	23	91.3	91.3	**100**
$C_8H_7NO_4$	13	92.3	84.6	**100**
$C_{15}H_{24}O$	16	93.8	93.8	**100**
$C_{11}H_{14}$	10	90.0	80.0	**100**
$C_7H_{15}NO$	14	85.7	85.7	**100**
$C_{10}H_{16}O_2$	26	92.3	84.6	**100**
$C_8H_{15}N$	11	81.8	90.9	**100**



Summary

- Developed K-M^{3}AID – a knowledge-guided cross-modal contrastive learning approach, leveraging domain-specific continuous features with natural order.
- Demonstrated the effectiveness of K-M^{3}AID in three tasks
  - ^{13}C NMR peak assignment
  - Molecular retrieval using ^{13}C NMR
  - Isomer recognition using ^{13}C NMR.

Acknowledge: NSF DMR 1933525
NSF DRL 2314156
NSF OAC 1920147

Thank You!