Supplementary Information

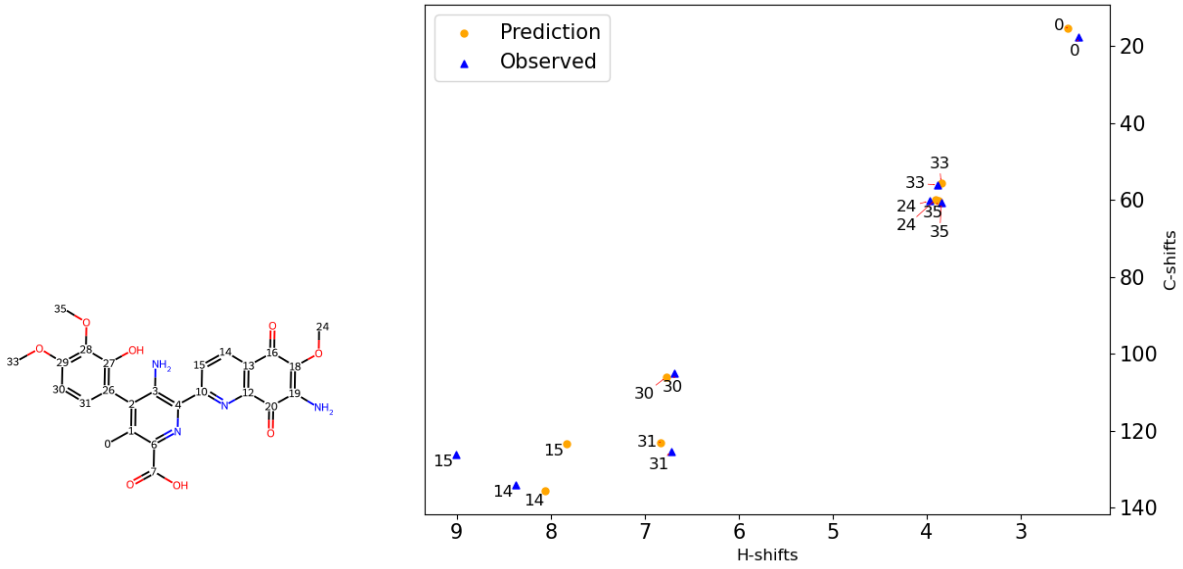
TransPeakNet for Solvent-Aware 2D NMR Prediction via Multi-Task Pre-Training and Unsupervised Learning

Yunrui Li,[‡a] Hao Xu, [‡b] Ambrish Kumar,[c] Duosheng Wang,[d]
Christian Heiss,[c] Parastoo Azadi,[c] and Pengyu Hong*[a]

February 4, 2025

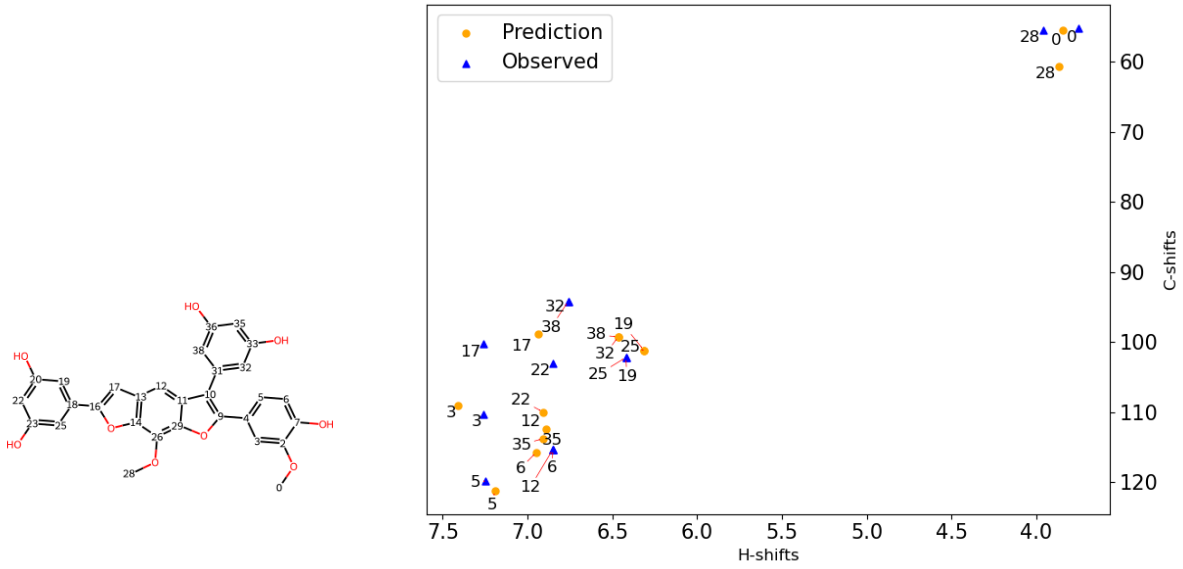
[a] *Department of Computer Science, Brandeis University, Waltham, MA, USA*
[b] *Department of Medicine, Harvard Medical School, Boston, MA, USA*
[c] *Complex Carbohydrate Research Center, University of Georgia, Athens, Georgia, USA*
[d] *Department of Chemistry, Boston College, Chestnut Hill, MA, USA*

1 Additional annotation examples for medium and large molecules



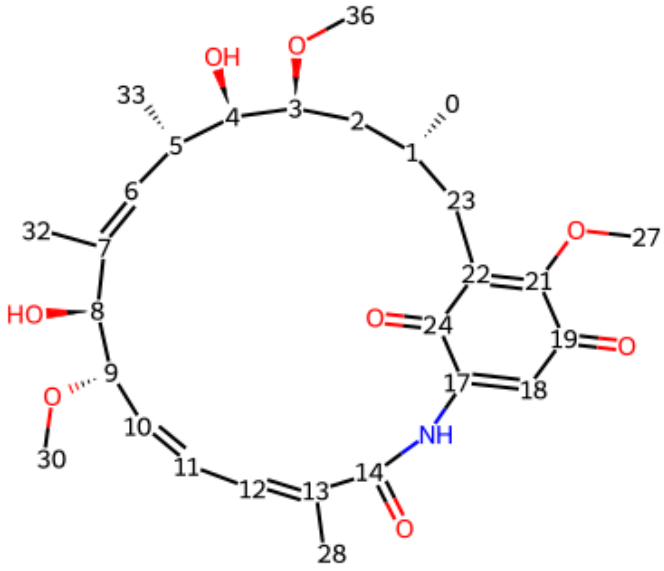
(a) Molecule (b) Prediction and Annotation

Figure S 1: Additional Annotation Example 1

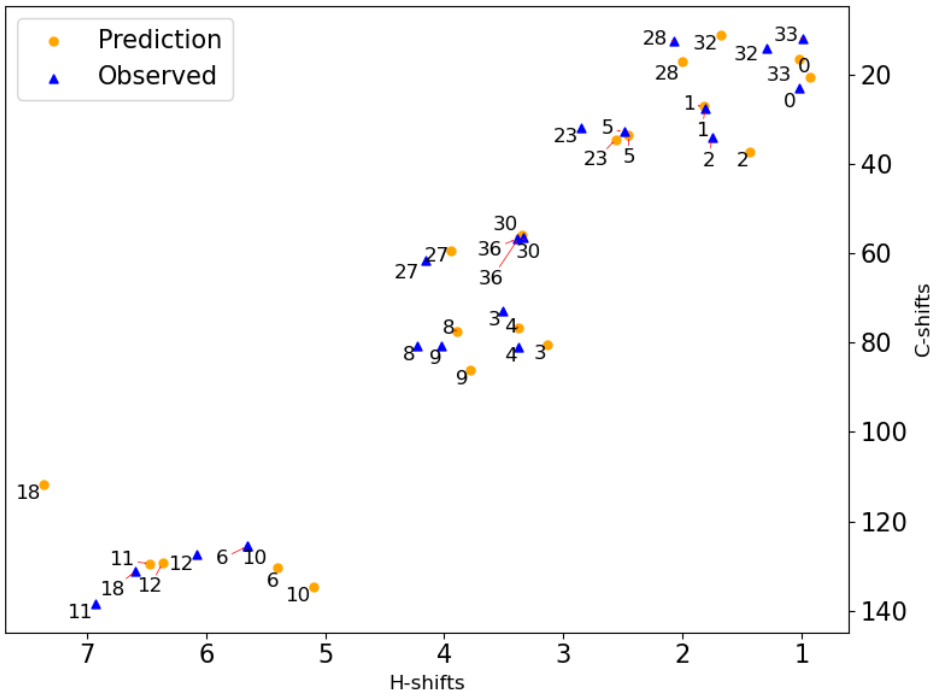


(a) Molecule (b) Prediction and Annotation

Figure S 2: Additional Annotation Example 2

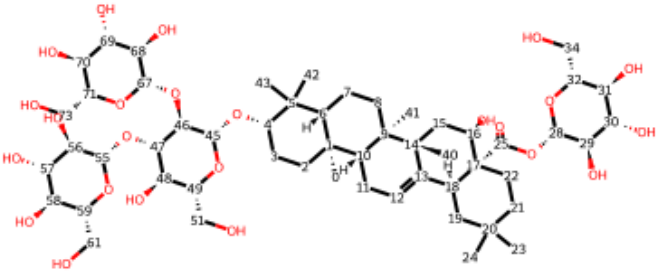


(a) Molecule

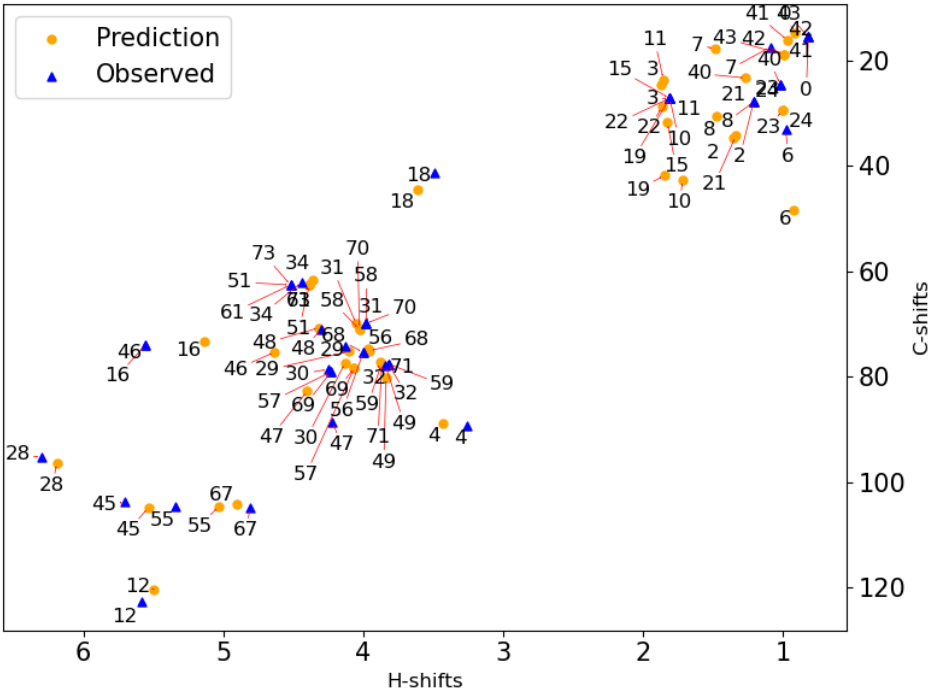


(b) Prediction and Annotation

Figure S 3: Additional Annotation Example 3



(a) Molecule



(b) Prediction and Annotation

Figure S 4: Additional Annotation Example 4

2 Partially correct annotation examples

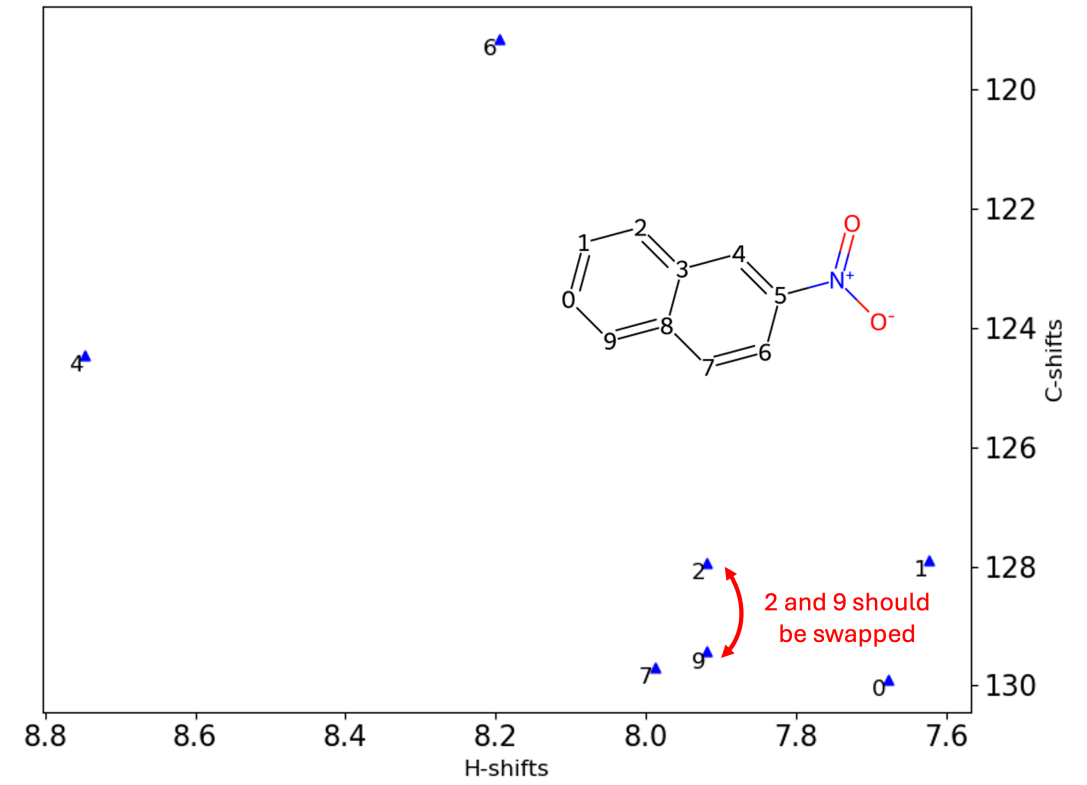


Figure S 5: Partially Correct Annotation: Example 1

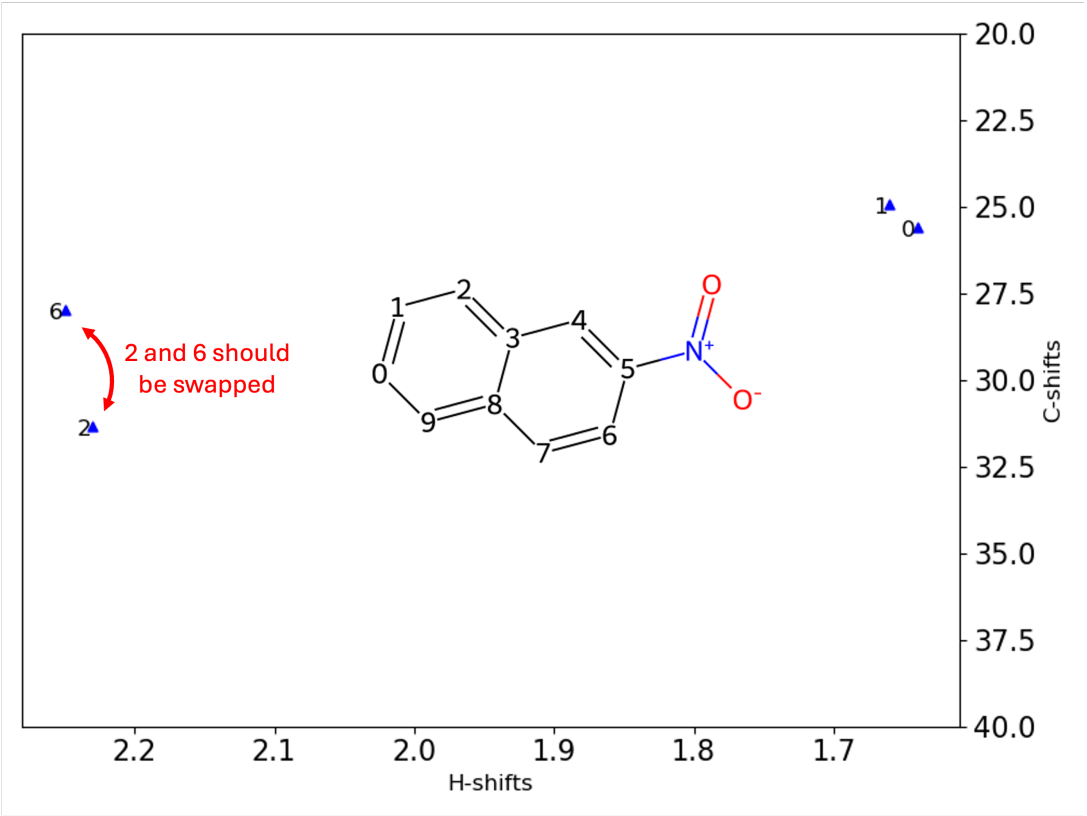


Figure S 6: Partially Correct Annotation: Example 2

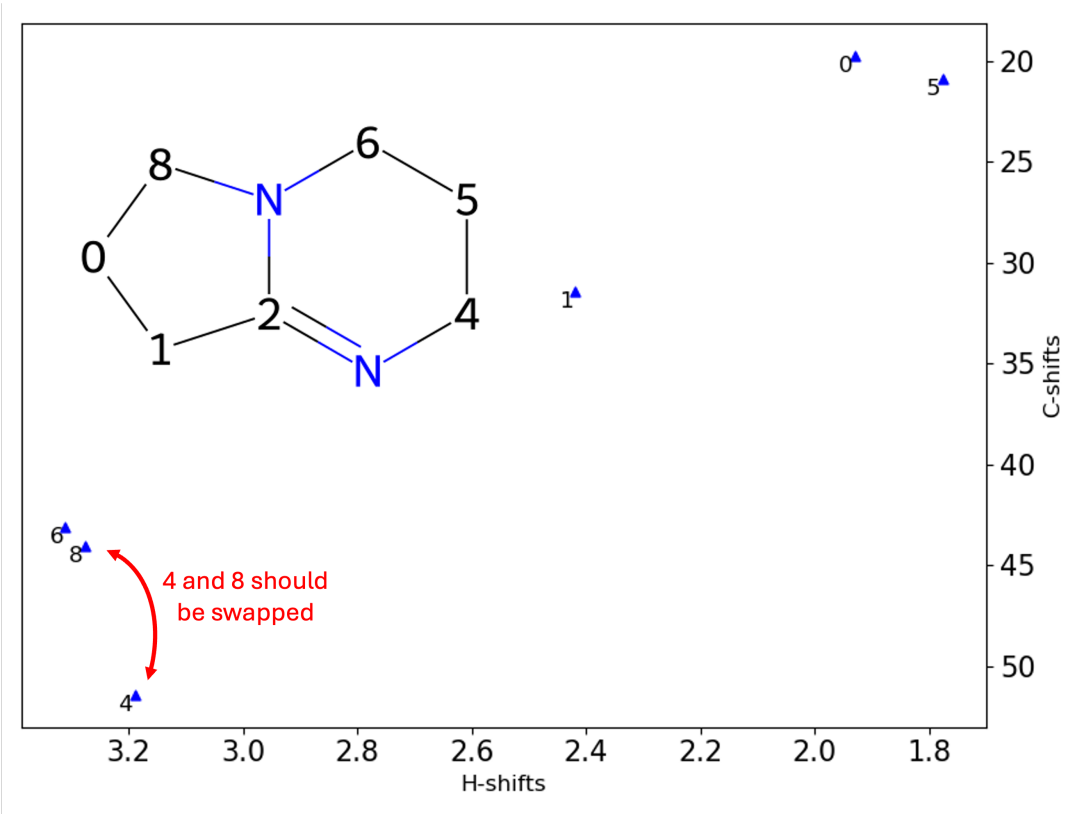


Figure S 7: Partially Correct Annotation: Example 3

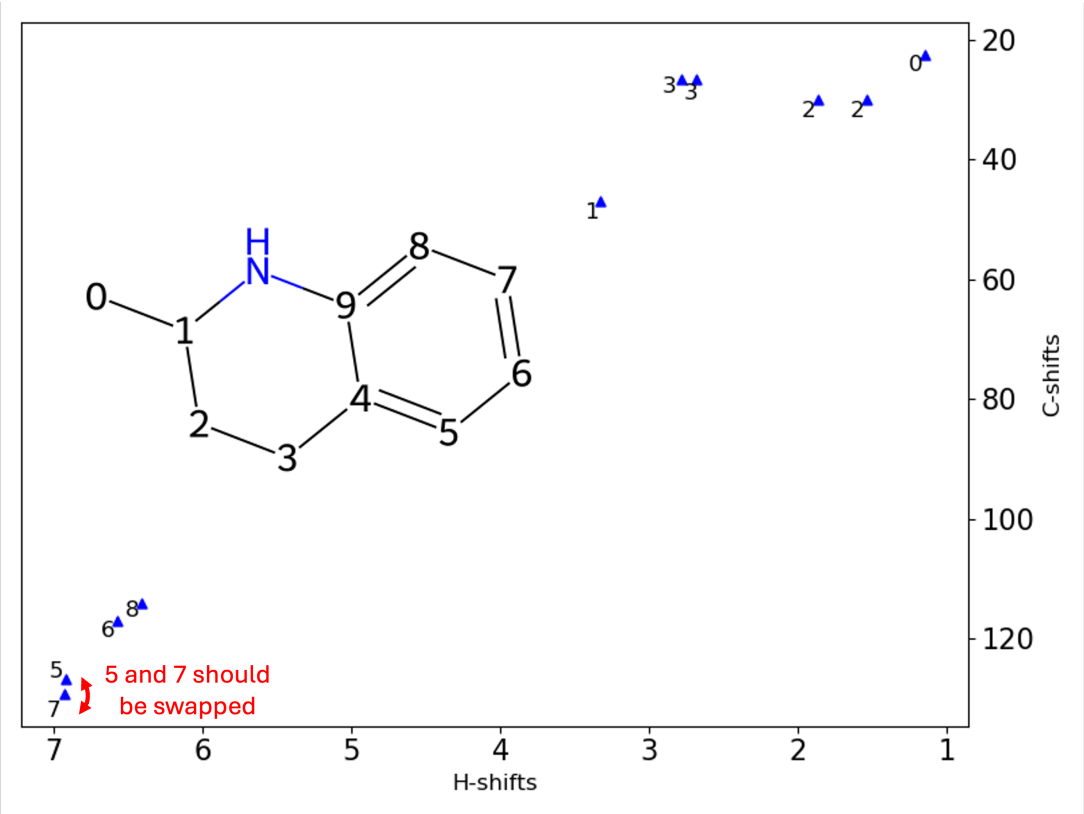


Figure S 8: Partially Correct Annotation: Example 4

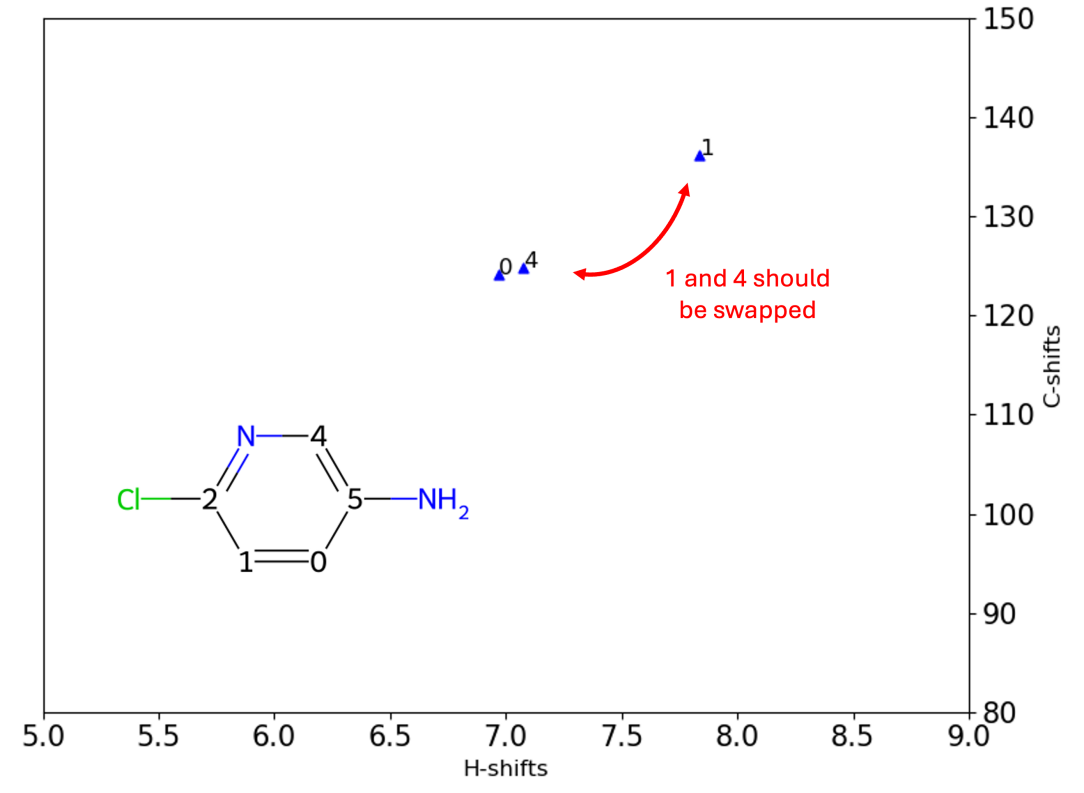
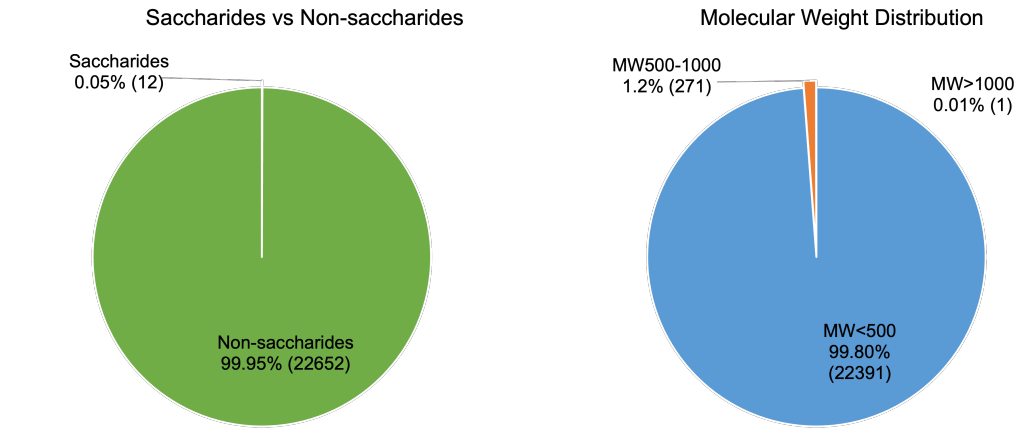
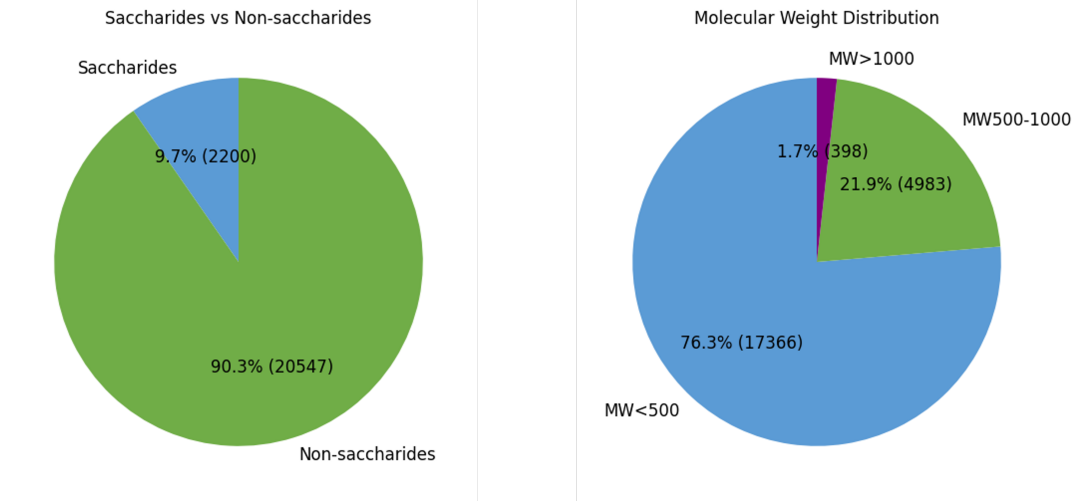


Figure S 9: Partially Correct Annotation: Example 5

3 Data distribution comparison between NMRshiftDB and HSQC data



(a) 1D Data



(b) HSQC Data

Figure S 10: Data Distribution Comparison

4 Uncertainty Measurement

Our model incorporates dropout in the prediction module, allowing us to run it multiple times to calculate the mean and standard deviation of the predictions as a measure of uncertainty. However, it is worth noting that dropout is typically disabled during inference because its primary purpose is to prevent overfitting during training by randomly deactivating a subset of neurons in each forward pass. This encourages the model to learn more robust and generalized patterns by reducing the dependence on specific neurons. The statistics reported in our manuscript were generated under the standard inference mode (i.e., without dropout).

Nevertheless, dropout can be enabled to estimate uncertainty by averaging multiple predictions for each atom as the final output and using their standard deviation as the uncertainty score. The dropout rate is set to 0.2 and the model is run 10 times. An example is provided below:

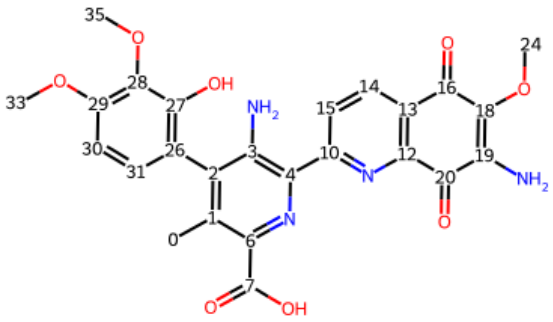


Figure S 11: Example Molecule for Uncertainty Measurement

Carbon Index	^{13}C **Mean Shift (std)**	^{1}H **Mean Shift (std)**	^{13}C **Shift (Ground Truth)**	^{1}H **Shift (Ground Truth)**	^{13}C **Shift (Dropout Disabled)**	^{1}H **Shift (Dropout Disabled)**
0	15.65 (3.87)	2.47 (0.10)	17.5	2.38	15.24	2.49
14	140.90 (6.28)	7.91 (0.45)	134.1	8.37	135.71	8.06
15	122.37 (5.29)	7.79 (0.46)	126.3	9.01	123.53	7.84
24	59.88 (0.91)	3.91 (0.04)	60.3	3.97	59.98	3.90
30	105.29 (2.06)	6.64 (0.28)	105.2	6.68	106.13	6.77
31	123.88 (5.25)	6.73 (0.30)	125.4	6.72	123.06	6.83
33	55.61 (0.23)	3.83 (0.02)	56.1	3.88	55.58	3.84
35	60.22 (0.77)	3.87 (0.04)	60.7	3.84	60.14	3.86

Table S 1: Comparison of predictions with and without dropout enabled, alongside ground truth shifts for ^{13}C and ^{1}H.