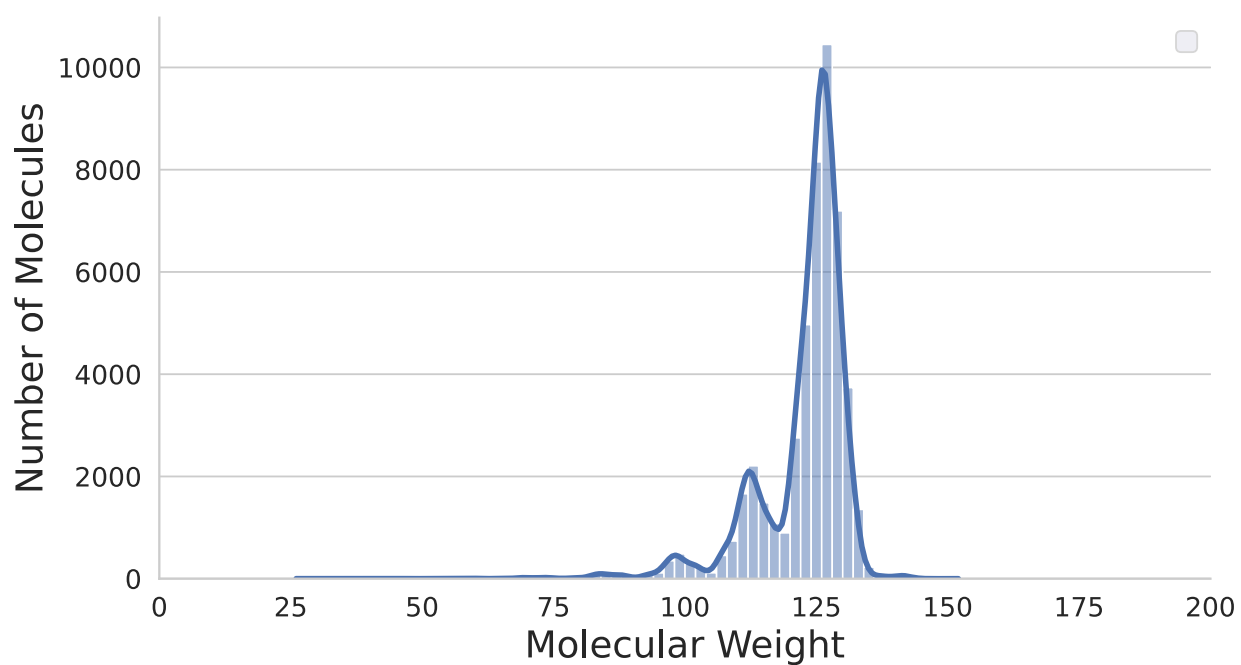


Supplementary Information for DeepSPInN

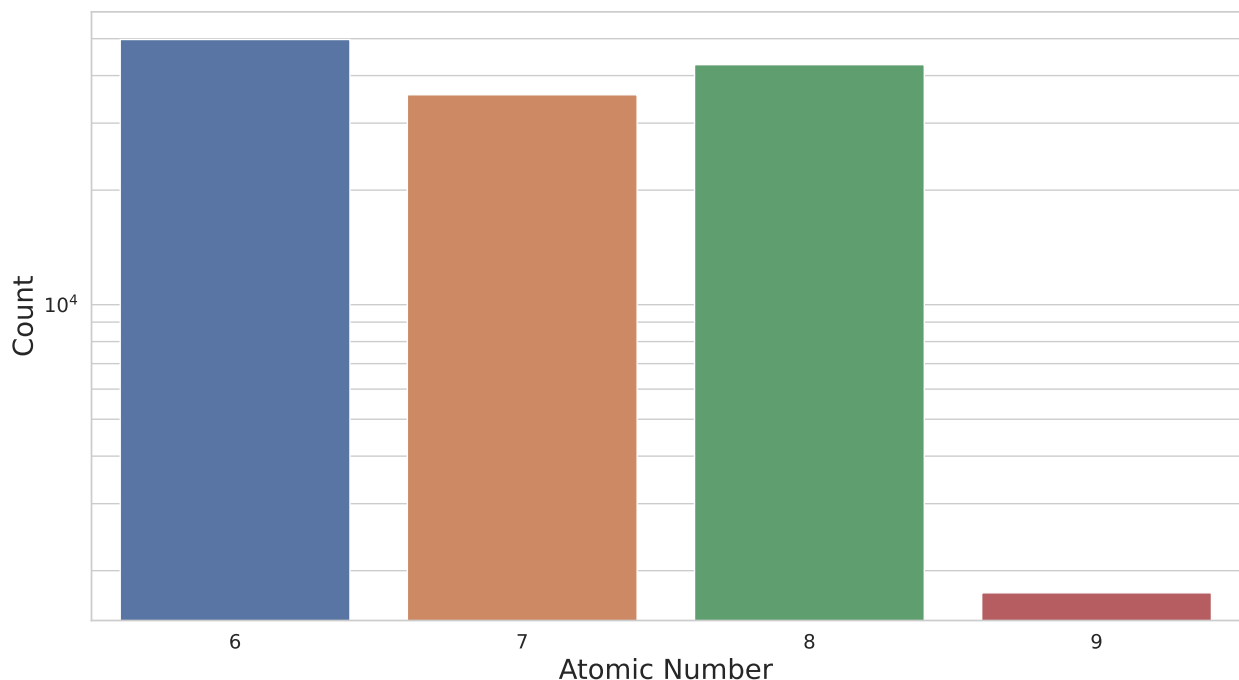
E-mail:

Statistics for the dataset

To understand the distribution of the molecules in the dataset, Figures 1 and 2 show the distribution of the molecular weights and the counts of molecules that have the heavy atoms C, N, O, and F.



Supplementary Figure 1: Distribution of the molecular weights in the dataset



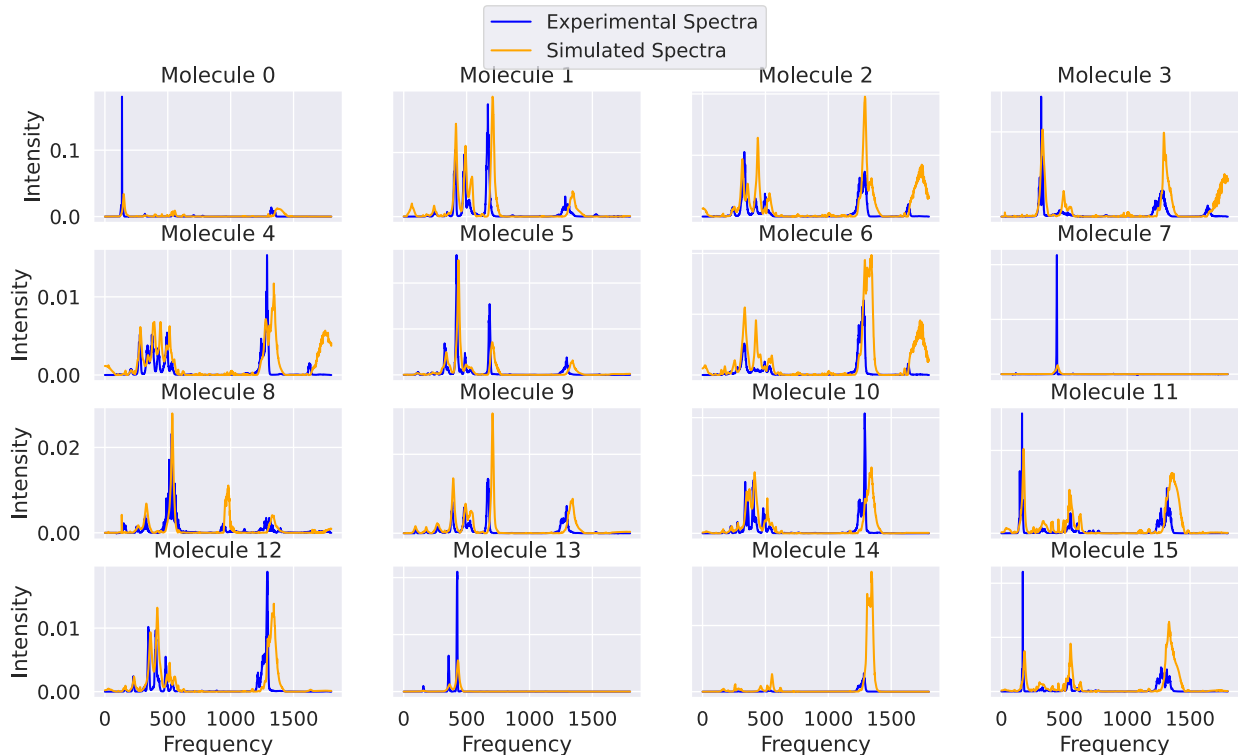
Supplementary Figure 2: Number of molecules that contain each element

Congruence of simulated and experimental Infrared spectra

Of the 40 molecules whose infrared spectra we were able to download from the NIST Quantitative Infrared Database, we had the simulated spectra of 15 molecules. The average SIS of the experimental and simulate infrared spectra is 0.2241. Although predictions are expected to have SIS in the range 0.40–0.70 to even be considered as loosely predictive, the simulated spectra should not be considered as replacements for experimental spectra.

Threshold reward for MCTS

MCTS will continue finding child nodes to the search tree until it reaches a terminal state. For a state to be considered a terminal, at least one of these following termination criteria have to be met:



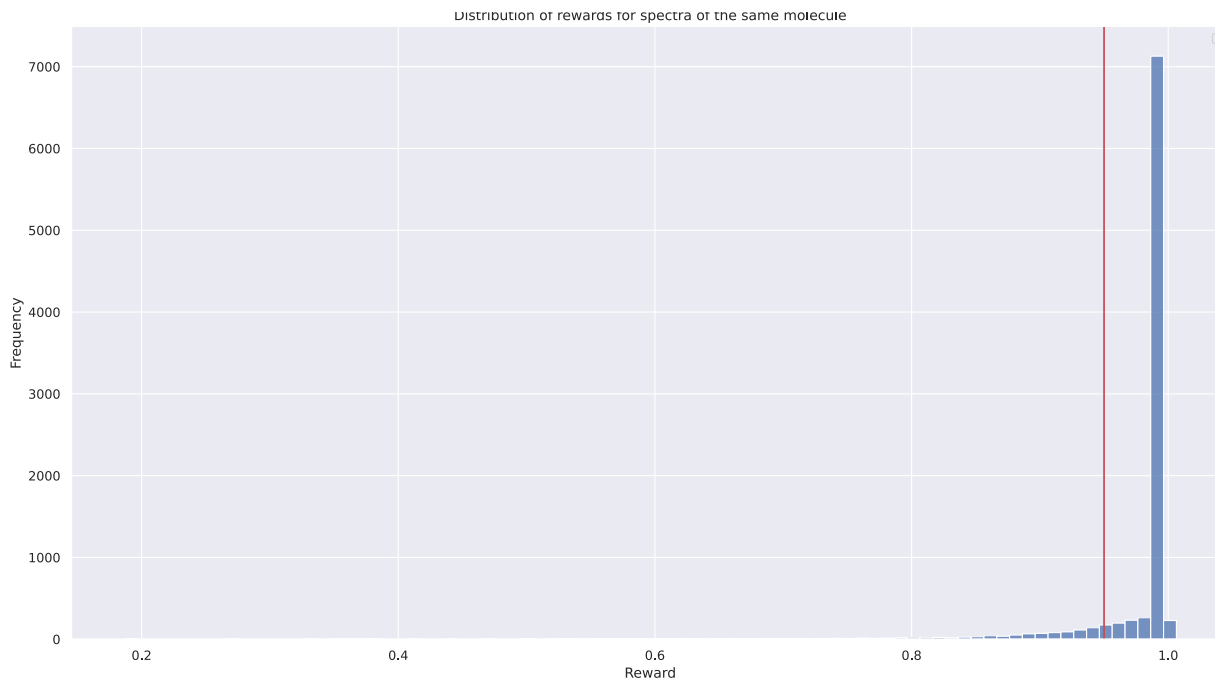
Supplementary Figure 3: Experimental and Simulated IR spectra

1. There are no more valid actions
2. The reward of this state is greater than a particular threshold
3. The NMR split values of the current state are such that the target NMR split values can never be reached

If there are no valid actions that can be taken from a state, it has to be terminal since there are no possible child nodes. If a state has a molecular graph where there are no individual nodes, further actions will just continue adding bonds within the molecule. In such a case, the environment checks if the NMR split values of a state can still allow further addition of bonds. If a state's NMR split values are such that the target NMR split values cannot be reached via the addition of more bonds, then the state is considered to be terminal.

In addition to these termination criteria, we want the framework to stop the tree search when it is confident that the target molecule has been reached. We use the reward function

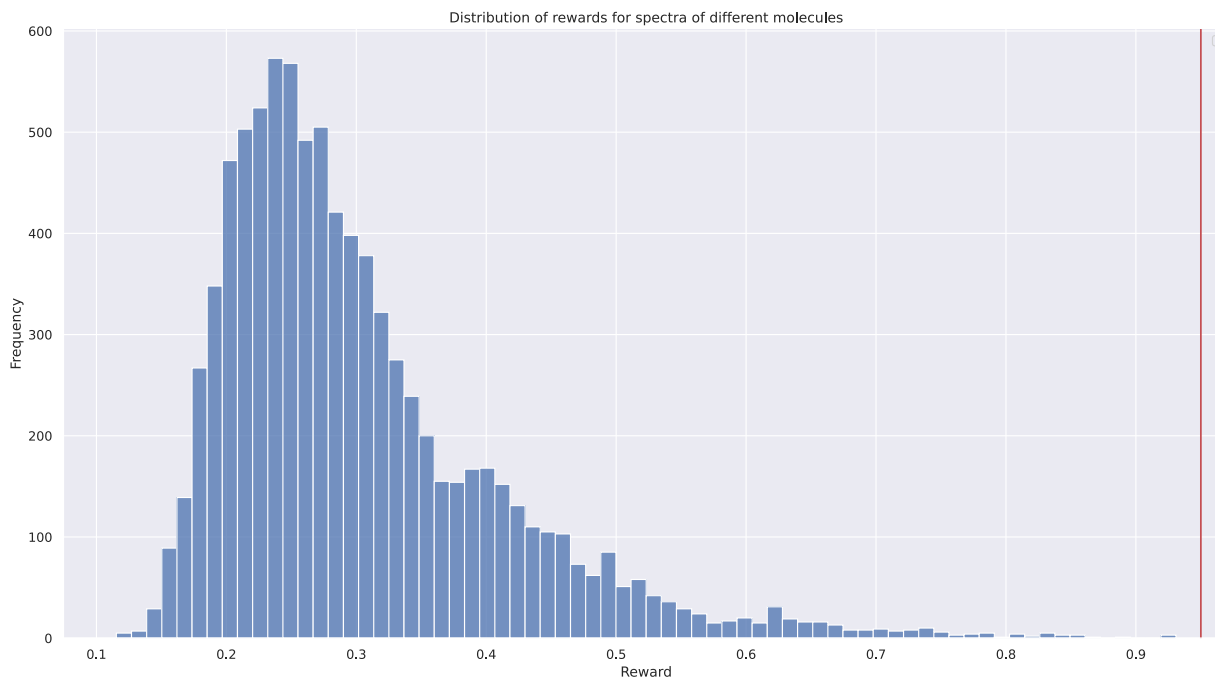
to get the reward for each state, and if the reward is greater than a particular threshold, the state is terminal. To choose this threshold value, we sampled 10,000 molecules from the test set and found the *SIS* between their IR spectrum and the predicted IR spectrum from Chemprop-IR. These rewards are plotted in Fig. 4. 88.63% of molecules have their self-rewards above 0.95. We sampled another 100 molecules and for each molecule, we found the *SIS* between the molecule’s IR spectrum and the Chemprop-IR predicted spectrum of the other 99 molecules. These rewards are plotted in Fig. 5. All these rewards are below 0.95. Choosing 0.95 as a reward threshold means that states with a reward greater than this threshold are very likely to be the target molecules themselves. This allows the framework to stop the tree search at this point.



Supplementary Figure 4: Distribution of rewards of the same molecules

SIS Loss

The original SIS description performs a Gaussian convolution to allow for any minor deviations in the spectral peak locations. This Gaussian convolution allows spectra with minor

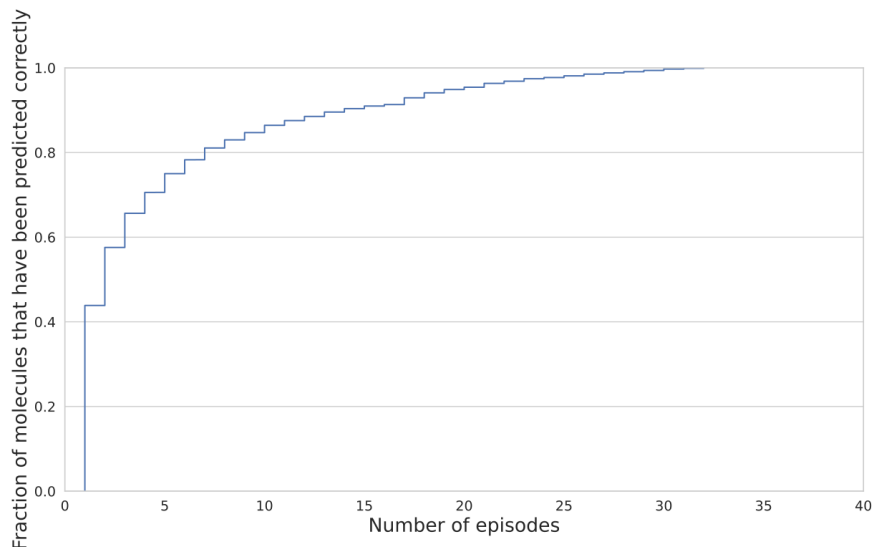


Supplementary Figure 5: Distribution of rewards between different molecules

differences in the peak locations and intensities to have a relatively high SIS score. We do not perform this Gaussian convolution while calculating SIS since the dataset used in this work already has broadened stick spectra and Chemprop-IR also predicts broadened IR spectra.

Training on molecules with ≤ 7 heavy atoms and testing on molecules with 8 or 9 heavy atoms

For the generalization study where the framework was trained on molecules with up to 7 heavy atoms, and tested on molecules with 8 or 9 heavy atoms, the test-train split of the dataset was heavily unbalanced. There are a total of 2,099 molecules with ≤ 7 heavy atoms, with all these molecules making the training set. The remaining 47,650 molecules have either 8 or 9 heavy atoms and these molecules make up the testing set. Due to computational constraints of testing the framework on the entire set, the framework was tested on a random subset of 5000 molecules that contain either 8 or 9 heavy atoms. Out of these 5000 molecules, 771 molecules have 8 heavy atoms and 4,229 have 9 heavy atoms.



Supplementary Figure 6: Cumulative plot of the fraction of correctly predicted molecules and the number of episodes that were taken to find the right molecule

Choosing the number of episodes hyperparameter

Due to the way we parallelized DeepSPInN and with the resources that we used, it takes the same amount of time to run the number of episodes that is the next closest multiple of 8. For example, it takes the same amount of time to run 22 episodes and 24 episodes. This is why we chose 32 episodes even though running for 28 episodes has a negligible decrease in the fraction of correctly predicted molecules when compared to 32.

Top N metrics for various functional groups/structural motifs

To identify if DeepSPInN has any affinity to predict the molecular structures that contain specific functional groups better than others, we analyzed the *Top N* metrics of the molecules that contain different functional groups. DeepSPInN performs well for molecules with ketones, with a *Top 1 (%)* accuracy of 96.24%, and performs the worst for molecules with amines with a *Top 1 (%)* accuracy of 86.99%. Table 1 shows the *Top N (%)* metrics of other

functional groups as well as the number of molecules in the test set that contained these functional groups.

Table 1: *Top N* metrics for molecules that have specific functional groups

Functional Group (Number of molecules)	Alcohols (2987)	Aldehydes (1214)	Amines (1983)	Ester (437)	Ketone (1064)	Phenol (556)
Top 1 (%)	92.37	93.49	86.99	93.82	96.24	90.83
Top 3 (%)	93.34	94.48	87.59	94.51	97.27	91.01
Top 5 (%)	93.40	94.65	87.64	94.51	97.37	91.01
Top 10 (%)	93.44	94.65	87.70	94.51	97.37	91.01

Using proton-coupled ^{13}C NMR spectra

In an experiment where we assume that we also have the ^{13}C NMR split values, we use these values to help prune the search tree by identifying molecules that are invalid according to the input NMR spectrum and nullify the rewards for these molecules. This discourages the tree search from exploring these molecules. The NMR split values for the NMR shift of each carbon atom are equivalent to the number of hydrogen atoms attached to it. Each carbon atom can be either a singlet (S, quaternary), doublet (D, tertiary), triplet (T, secondary), or a quartet (Q, primary) atom with each of these denoted by S, D, T, and Q respectively. Since each valid action is defined as the addition of a bond between two atoms in the MDP reformulation, each valid action can only convert the carbon atoms from $Q \rightarrow T \rightarrow D \rightarrow S$. If the target Q-splits are more than the Q-splits at one such state, this state is not valuable since no valid action from this state would be able to increase the count of Q-splits. If the Q-splits match, checking the T and D-splits subsequently in the same way further identify more states that get zero-rewards.

Tables 2, 3, and 4 present the same results from the main paper, but with the ^{13}C NMR split values being used. The dataset split is different from the main paper, with a 80-20 split being used for the train and test sets.

Table 2: *Top N* metrics for varying n_{mcts} values

n_{mcts}	IR+NMR			
	100	200	400	800
Top 1 (%)	82.311	90.773	94.934	95.980
Top 3 (%)	82.874	91.597	95.839	96.925
Top 5 (%)	82.994	91.778	96.060	97.106
Top 10 (%)	83.015	91.798	96.120	97.166
Top 40 (%)	83.015	91.798	96.120	97.206

Table 3: Performance of IR-and-NMR-trained, IR-trained, and NMR-trained models for $n_{\text{mcts}} = 400$

	IR and NMR	Only IR	Only NMR
Top 1 (%)	94.934	74.075	40.462
Top 3 (%)	95.839	74.396	61.849
Top 5 (%)	96.060	74.396	68.201
Top 10 (%)	96.120	74.396	73.105
Top 40 (%)	96.120	74.423	74.472

Testing DeepSPInN checkpoints trained on simulated spectra to elucidate experimental spectra

The simulated Infrared and ^{13}C NMR spectra used to train and test DeepSPInN reflect the complexity of experimental spectra but can not serve as replacement for the experimental spectra. Since DeepSPInN is able to work with simulated spectra, it can analogously learn to work with experimental spectra.

With this manuscript focusing on the development of DeepSPInN as a proof of concept, the current DeepSPInN model checkpoints can not (should not) be used for testing on experimental data since it was only trained on simulated data. DeepSPInN is able to learn the complexity of spectra, as seen by its performance on simulated spectra, and would perform well on unseen experimental spectra when it is also trained on experimental spectra. Future works would be able to construct databases of experimental Infrared and ^{13}C NMR spectra


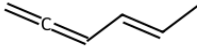
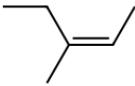
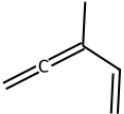
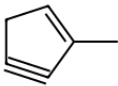
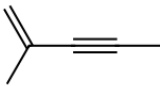
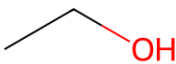
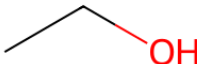
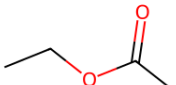
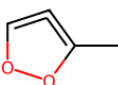
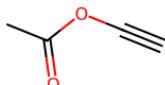
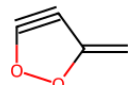
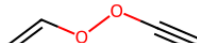
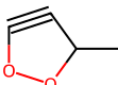
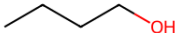
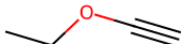


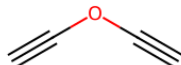
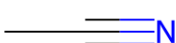
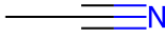

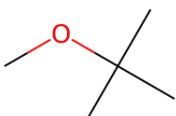
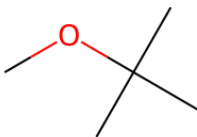
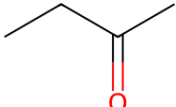
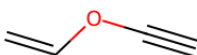

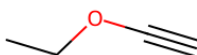
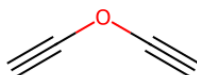
Table 4: Training on molecules with ≤ 7 atoms and testing on molecules with ≥ 8 atoms for $n_{\text{mcts}} = 400$

	≥ 8 atom molecules	8-atom molecules	9-atom molecules
Top 1 (%)	80.971	96.024	77.948
Top 3 (%)	82.046	97.247	78.992
Top 5 (%)	82.148	97.247	79.115
Top 10 (%)	82.199	97.247	79.176
Top 40 (%)	82.250	97.247	79.238

to train checkpoints of DeepSPInN that would work well on new experimental data.

We demonstrate that the current DeepSPInN checkpoints do not perform well on elucidating the structures of experimental spectra by gathering the experimental Infrared and ^{13}C NMR spectra of 14 molecules from the databases NIST Quantitative Infrared Database and nmrshiftdb2. In Table 5, we show the top candidate molecules of some of these molecules as predicted by DeepSPInN. As expected, DeepSPInN did not perform well and was only able to resolve 3/14 of the molecules.

Table 5: Top candidate molecules and the rewards when given experimental IR and ^{13}C NMR spectra as input to a DeepSPInN model trained on simulated spectra

Target Molecule	Top 5 Candidate Molecules (and their final rewards)				
 Target Molecule	 Reward: 0.133	 Reward: 0.127	 Reward: 0.127	 Reward: 0.123	 Reward: 0.123
 Target Molecule	 Reward: 0.270				
 Target Molecule	 Reward: 0.289	 Reward: 0.273	 Reward: 0.233	 Reward: 0.214	 Reward: 0.192
 Target Molecule	 Reward: 0.233	 Reward: 0.231	 Reward: 0.190	 Reward: 0.107	
 Target Molecule	 Reward: 0.277	 Reward: 0.125			
 Target Molecule	 Reward: 0.242				
 Target Molecule	 Reward: 0.204	 Reward: 0.200	 Reward: 0.189	 Reward: 0.087	