

Leveraging Infrared Spectroscopy for Automated Structure Elucidation

Supporting Information

Marvin Alberts^{1,2*}, Teodoro Laino^{1,2}, and Alain C. Vaucher^{1,2}

¹*IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*

²*National Center for Competence in Research-Catalysis (NCCR-Catalysis), Zurich, Switzerland*

^{*}*Correspondence: marvin.alberts@ibm.com*

1 Similarity between simulated and experimental spectra

In the following we assess the similarity between the experimental and simulated spectra. We utilise two similarity metrics: The Spectral information similarity as introduced by McGill et. al. and the cosine similarity. We compare the simulated spectra against a total of 721 experimental ones in three metrics: First we compare the simulated spectra directly to it’s corresponding experimental one. Second we assess the similarity between the simulated spectrum and the experimental spectrum of the molecule with the closest Tanimoto similarity to demonstrate that similar molecules can be differentiated. Last we calculate the mean of the similarity between the simulated and all experimental spectra.

Table 1: The performance of models trained on different sections of the IR spectrum.

Metric	Sim. vs Exp.	Sim. vs Tanimoto	Sim. vs all Others	Tanimoto Similarity
SIS $\uparrow$	0.114 ± 0.091	0.111 ± 0.082	0.086 ± 0.036	0.776 ± 0.122
Cosine Similarity $\downarrow$	0.265 ± 0.135	0.272 ± 0.135	0.468 ± 0.140	0.776 ± 0.122

2 Window selection

The experiments in the main text show that removing the section from $2000\text{-}2800\text{cm}^{-1}$ and $3300\text{-}4000\text{cm}^{-1}$ improves performance as relatively few peaks are present in these section. This raises the question if removing further section of the spectrum further increases performance. To this end we train three further models removing more sections of the IR spectrum. The results are shown in Table 2. All experiments include the chemical formula and each spectrum is represented with 400 tokens.

Table 2: The performance of models trained on different sections of the IR spectrum.

	Window	Top-1%	Top-5%	Top-10%
Window	400-3982 cm^{-1} (Full Spectrum)	26.21	51.38	59.63
	400-2000 and 2800-3300 cm^{-1}	30.22	55.87	63.88
	800-2000 and 2800-3300 cm^{-1}	30.44	55.64	63.55
	400-1800 and 2800-3300 cm^{-1}	32.70	58.13	66.14
	800-1800 and 2800-3300 cm^{-1}	31.68	56.01	63.59

Table 2 supports the hypothesis that removing further section of the IR spectrum increases performance. Especially removing the section from $1800\text{-}2000\text{cm}^{-1}$ yields a performance improvement of almost two percent. However, when training the same models on augmented data the difference in performance becomes negligible and the model trained on a window from 400-2000 and $2800\text{-}3300\text{cm}^{-1}$ performs the best (see 3). The data was augmented using horizontal shifting as explained above.

Table 3: Performance of models trained on different windows with augmentation.

	Window	Top-1%	Top-5%	Top-10%
Window	400-3982cm ⁻¹ (Full Spectrum)	31.21	58.13	64.86
	400-2000 and 2800-3300cm ⁻¹	36.49	62.63	69.96
	800-2000 and 2800-3300cm ⁻¹	36.42	62.58	69.72
	400-1800 and 2800-3300cm ⁻¹	35.87	61.98	69.18
	800-1800 and 2800-3300cm ⁻¹	35.74	61.85	69.03

These seemingly contradictory results can be explained by through the effect of augmentation. Data augmentation increases the diversity of the data which can promote the model to learn more complex pattern present in the data. In this case through data augmentation the model is able to extract more information from the peaks that are present in the region of 1800-2000cm⁻¹.

3 Scaffolds diversity in the NIST Dataset

The 3,453 molecules in the NIST dataset contain a total of 257 unique scaffolds. Figure 1 shows the number of molecules contained in the top 100 scaffolds. However, there is a heavy imbalance towards molecules containing a benzene scaffold and linear molecules not containing any rings. These two scaffolds account for ca. 40 and 30% of all molecules respectively. In addition, 165 scaffolds are only represented by one molecule.

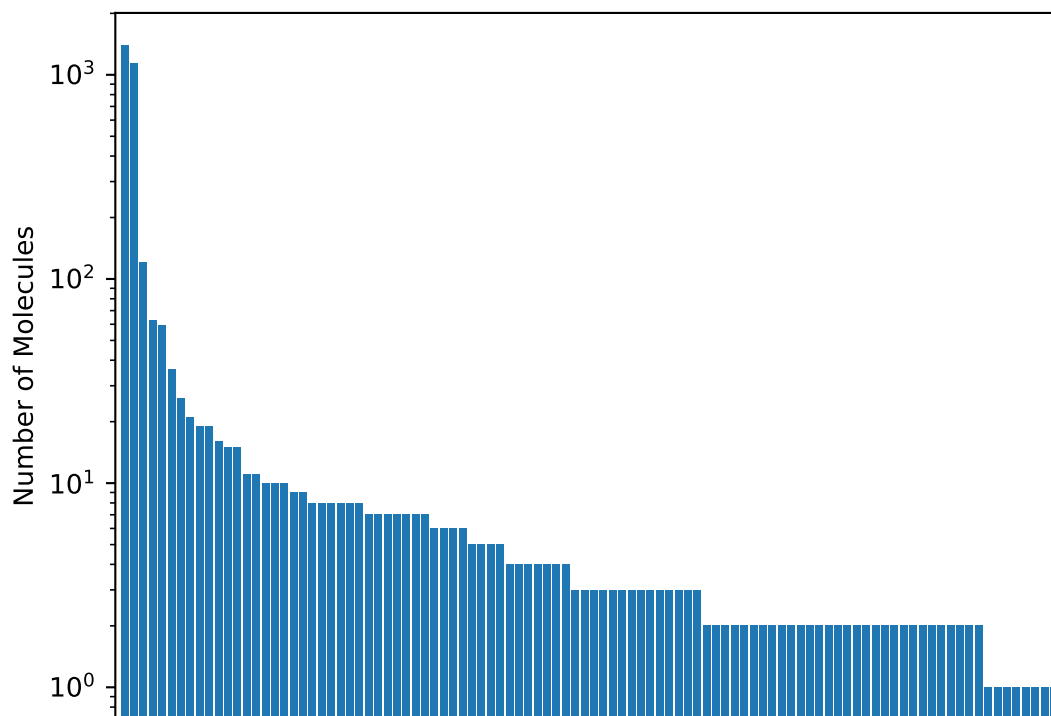


Figure 1: Bar chart of the number of molecules containing each scaffold.

4 Similarity between the Training and Test set

We assess the similarity of randomly sampled training and test sets from the NIST database by investigating the Tanimoto similarity between the two sets. Figure 2 shows the percentage of molecules in the test set having at least one molecule above a given Tanimoto similarity. The plot indicates a high similarity between the training and test set with more than 50% of molecules in the test set having a molecule in the training with a Tanimoto similarity above 0.7.

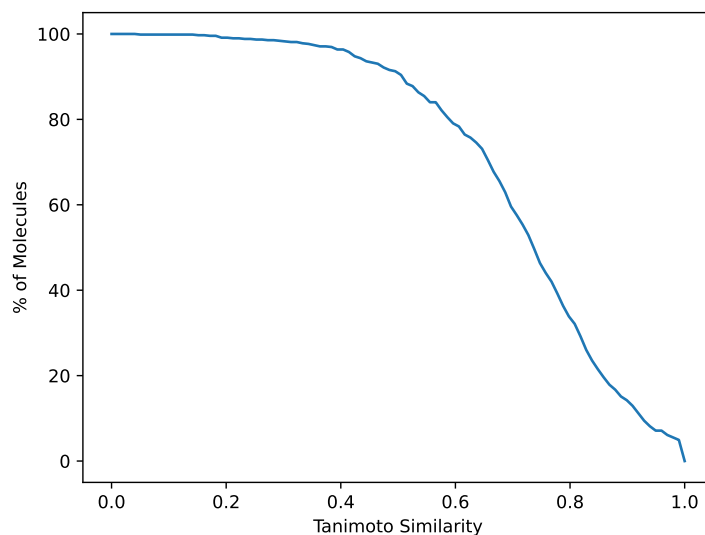


Figure 2: The percentage of molecules in the test set with at least one similar molecule in the training at a given Tanimoto similarity is plotted against the Tanimoto similarity.

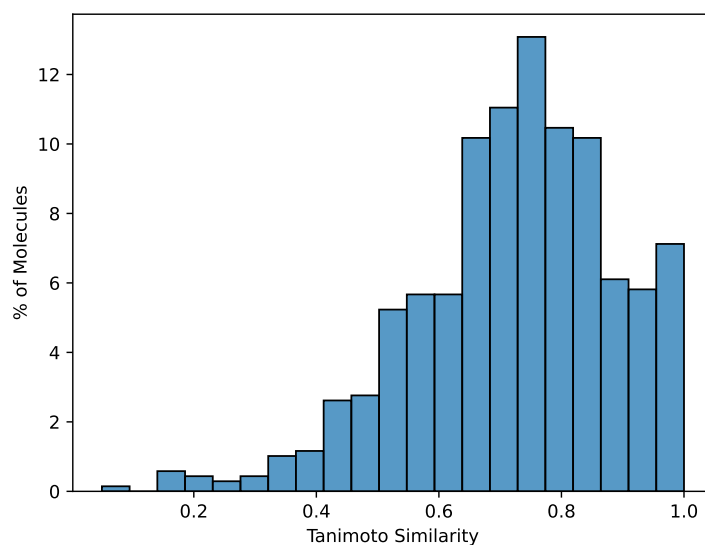


Figure 3: Histogram of the maximum similarity of molecules in the test set with molecules in the training set.

5 Heavy atom count distribution

The heavy atom distribution of the NIST IR dataset set is shown in Figure 4. Conversely, the distribution of the simulated data test set is shown in 5. As the molecules were randomly sampled from PubChem [1], this distribution reflects the distribution of the PubChem database. The number of molecules increases exponentially with the heavy atom count, demonstrating the exponentially larger chemical space with an increase in the heavy atom count.

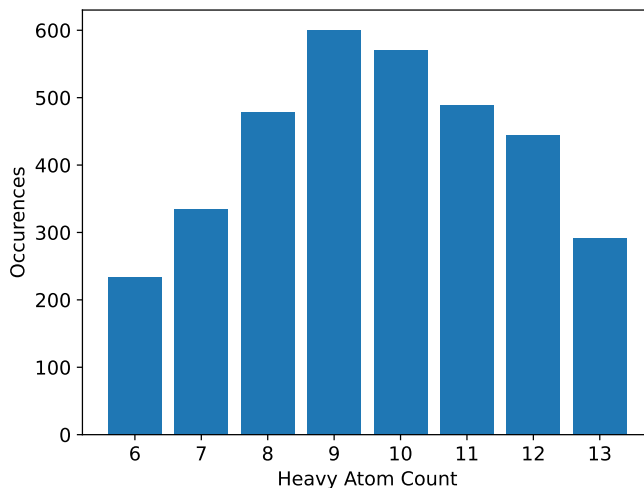


Figure 4: Heavy atom distribution of the NIST IR dataset.

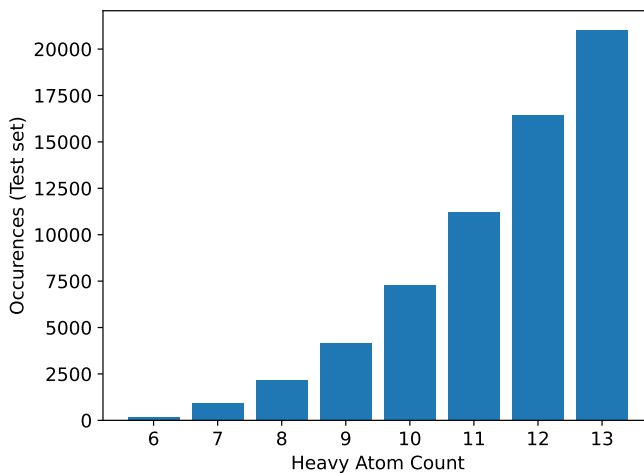


Figure 5: Heavy atom distribution of the simulated test set.

6 Functional group definitions

Functional groups are defined in SMARTS as shown in Table 4. Using these SMARTS and RDKit the presence of a certain function group is determined by invoking `<RDKit molecule>.GetSubstrucMatches(<RDKit molecule from SMARTS pattern>)`

Table 4: Functional group definitions used.

	Definition
Alcohol	[OX2H] [CX4; !\$ (C ([OX2H]) [O,S,#7,#15])]
Carboxylic Acid	[CX3] (=O) [OX2H1]
Ester	[#6] [CX3] (=O) [OX2H0] [#6]
Ether	[OD2] ([#6]) [#6]
Aldehyde	[CX3H1] (=O) [#6]
Ketone	[#6] [CX3] (=O) [#6]
Alkene	[CX3] = [CX3]
Alkyne	[\$ ([CX2] #C)]
Benzene	c1ccccc1
Primary Amine	[NX3; H2; !\$ (NC = [!#6]) ; !\$ (NC# [!#6])] [#6]
Secondary Amine	[NH1, nH1])
Tertiary Amine	[NH0, nH0])
Amide	[NX3] [CX3] (= [OX1]) [#6]
Cyano	[NX1] # [CX2]
Fluorine	[#6] [F]
Chlorine	[#6] [Cl]
Iodine	[#6] [I]
Bromine	[#6] [Br]
Sulfonamide	[#16X4] ([NX3]) (= [OX1]) (= [OX1]) [#6]
Sulfone	[#16X4] (= [OX1]) (= [OX1]) ([#6]) [#6]
Sulfide	[#16X2H0]

[†] Adapted from [2]

7 Results: Model accuracy depending on specific functional groups

In Table 5 the accuracy of the model depending on the presence of specific functional groups in the target molecule is shown. “Count” represents the number of molecules with this functional group in the test set. Additionally, the average heavy atom count (“Avg. HAC” in the table) is calculated to rule out bias.

Table 5: The model’s ability to predict the correct molecular structure based on if a specific functional group is present in the target molecule.

	Count	Avg. HAC	Top-1%	Top-5%	Top-10%
Tertiary Amine	507	10.23	21.89	35.70	39.05
Secondary Amine	167	10.05	26.95	43.11	45.51
Amide	88	10.10	32.95	44.32	46.59
Sulfide	90	9.47	35.56	64.44	65.56
Cyano	88	9.83	36.36	60.23	65.91
Alkene	591	8.99	41.12	63.96	68.70
Chlorine	465	9.69	41.29	63.66	65.81
Primary Amine	217	9.63	44.70	61.75	63.13
Ester	278	10.37	46.04	70.86	73.74
Benzene	1376	10.78	46.22	66.86	68.97
Ketone	301	10.23	46.51	69.10	71.10
Alcohol	563	9.36	46.71	70.16	75.13
Carboxylic Acid	162	10.23	46.91	73.46	74.07
Ether	341	10.36	46.92	67.16	69.50
Alkyne	119	8.92	47.06	76.47	81.51
Aldehyde	81	9.75	50.62	72.84	76.54
Fluorine	139	9.97	54.68	71.22	71.94
Bromine	242	9.26	58.68	78.51	80.58
Iodine	37	8.97	67.57	78.38	78.38

8 Results: Functional group accuracy

In Table 6 we present the models ability to predict functional groups based on the IR spectrum and the chemical formula. To calculate the scores we compare the functional groups present in the target molecule to those present in the top-1 prediction. The precision, recall and F1-score were calculated.

Table 6: The model’s ability to predict that a functional group is present based on the IR spectrum. This table is based on the models top-1 prediction.

	Count	Precision	Recall	F1
Secondary Amine	161	0.64	0.75	0.69
Cyano	80	0.80	0.71	0.75
Amide	86	0.81	0.71	0.76
Sulfide	83	0.83	0.81	0.82
Aldehyde	78	0.82	0.86	0.84
Tertiary Amine	420	0.92	0.77	0.84
Alkyne	119	0.92	0.81	0.86
Ether	330	0.87	0.86	0.87
Ketone	294	0.84	0.93	0.88
Alkene	585	0.89	0.90	0.90
Primary Amine	203	0.87	0.93	0.90
Carboxylic Acid	155	0.90	0.91	0.91
Ester	277	0.90	0.94	0.92
Alcohol	561	0.93	0.93	0.93
Benzene	1293	0.98	0.97	0.98

9 Correlation of functional groups in the test set

Figure 6 shows the correlation between individual groups in the test set.

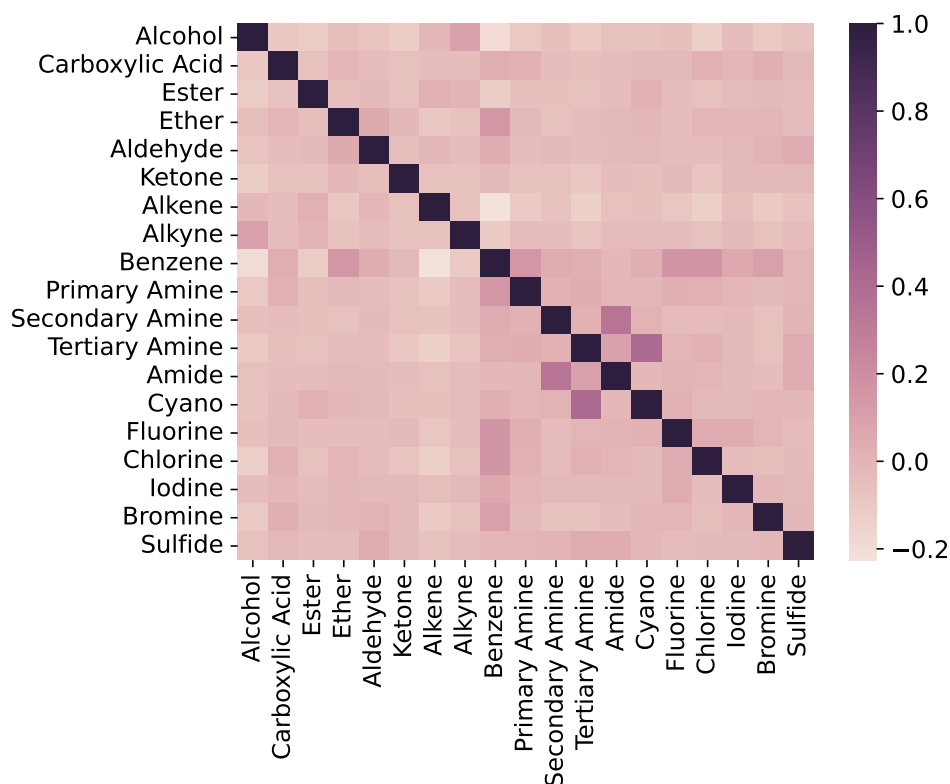


Figure 6: Correlation between functional groups of the test set.

References

- [1] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51:D1373–D1380, 2023.
- [2] Daylight: SMARTS Examples. https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html (Accessed April 20, 2023).