Supplementary Information (SI) for Digital Discovery.

Supporting Information: Setting New Benchmarks in AI-driven Infrared Structure Elucidation

Marvin Alberts[1,2,3], Federico Zipoli[1,3], and Teodoro Laino[1,3]

[1]*IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland*
[2]*University of Zurich, Department of Chemistry, Winterthurerstrasse 190, 8057 Zurich, Switzerland*
[3]*NCCR Catalysis, Switzerland*

1 Detailed Results

In-depth results for all models trained as part of this manuscript are provided in the table below. Results include Top–1, –5, and –10 performances both on simulated data (after pretraining) and experimental data (after finetuning). Results for all finetuned models were verified with 5-fold cross-validation.

	Layer Norm.	Pos. Enc.	Gated Linear Unit	Patch Size	Pretraining Augmentation	Finetuning Augmentation	Simulated			Experimental		
							Top–1 ↑	Top–5 ↑	Top–10 ↑	Top–1 ↑	Top–5 ↑	Top–10 ↑
Alberts et. al.	Pre	Sin-Cos	✗	N/A	Hori. + Smooth	Hori. + Smooth	45.33	72.21	78.50	44.39±5.31	66.85±3.08	69.79±2.48
AZ Patches	Post	Sin-Cos	✗	125	None	None	39.35	60.70	66.96	42.48±2.17	70.20±1.25	78.45±1.05
Wu	Post	Sin-Cos	✗	64	Noise + Smiles	Noise + Smiles	50.51	76.34	81.38	**53.56±1.13**	**74.18±0.79**	**80.36±0.70**
Ours (Patches)	Pre	Sin-Cos	✗	125	None	None	20.84	40.35	47.29	42.59±2.64	71.29±3.21	78.04±2.81
Ours (Patches)	Post	Sin-Cos	✗	125	None	None	39.86	60.68	66.52	48.36±3.14	75.47±2.11	81.58±2.08
Ours (Patches)	Post	Learned	✗	125	None	None	39.78	61.35	67.19	49.55±1.77	76.51±0.78	82.39±0.83
Ours (Patches)	Post	Learned	✓	125	None	None	42.94	63.52	69.47	**50.01±1.53**	**76.57±2.95**	**83.09±1.83**
Ours (Patches)	Post	Learned	✓	25	None	None	45.73	65.85	71.30	49.81±3.49	74.54±2.64	81.26±1.71
Ours (Patches)	Post	Learned	✓	50	None	None	44.48	64.63	70.89	51.03±2.82	77.51±2.88	82.35±2.83
Ours (Patches)	Post	Learned	✓	75	None	None	44.23	65.27	70.68	**52.25±2.71**	**77.67±3.56**	83.00±2.14
Ours (Patches)	Post	Learned	✓	100	None	None	43.49	64.36	69.72	51.72±3.08	77.03±3.54	82.62±2.19
Ours (Patches)	Post	Learned	✓	125	None	None	42.97	63.51	69.40	50.57±2.59	77.15±1.90	**83.57±1.67**
Ours (Patches)	Post	Learned	✓	150	None	None	41.52	63.08	68.93	48.36±3.11	75.76±2.46	82.07±2.13
Ours (Patches)	Post	Learned	✓	75	Hori.	None	43.18	63.12	68.95	50.33±2.37	76.22±1.44	83.15±1.19
Ours (Patches)	Post	Learned	✓	75	Smooth	None	42.94	62.63	67.84	48.60±1.74	75.24±0.46	81.90±0.56
Ours (Patches)	Post	Learned	✓	75	Pseudo	None	45.26	67.84	73.97	50.45±1.13	77.44±0.96	83.64±0.93
Ours (Patches)	Post	Learned	✓	75	Smiles	None	50.86	66.64	70.51	54.62±3.06	76.95±1.45	82.65±1.51
Ours (Patches)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	None	50.62	68.52	72.39	**55.58±1.75**	**79.24±1.57**	**84.19±1.78**
Ours (Patches)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Hori.	50.62	68.52	72.39	57.49±1.86	79.29±1.65	84.25±1.46
Ours (Patches)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Smooth	50.62	68.52	72.39	56.04±1.85	79.64±2.05	85.06±2.05
Ours (Patches)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Pseudo	50.62	68.52	72.39	55.10±3.00	**80.72±2.43**	**85.19±1.99**
Ours (Patches)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Smiles	50.62	68.52	72.39	**59.80±1.64**	76.55±1.89	80.99±1.33
Ours (Patches)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Hori.+ Smooth Smiles + Pseudo	50.62	68.52	72.39	**61.07±2.09**	77.59±1.41	81.94±1.16
Ours (Patches Constr. Gen.)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Hori.+ Smooth Smiles + Pseudo	50.62	68.52	72.39	**63.79±3.18**	79.29±3.12	83.95±2.54
Ours (Patches Constr. Gen. 5-35 HAC)	Post	Learned	✓	75	Hori.+ Smooth Smiles + Pseudo	Hori.+ Smooth Smiles + Pseudo	50.62	68.52	72.39	59.41±1.96	73.52±1.18	77.59±0.90

2 Training Metrics: Patch size 25 vs 75

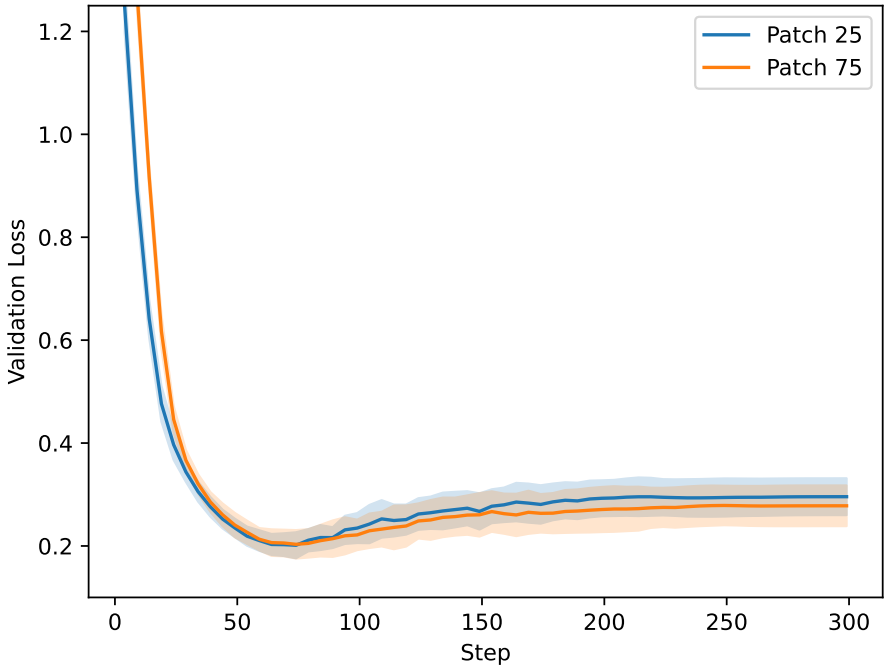


Figure 1: Validation loss of training a model with patch size 25 and 75. The average validation loss across five cross validations is plotted.

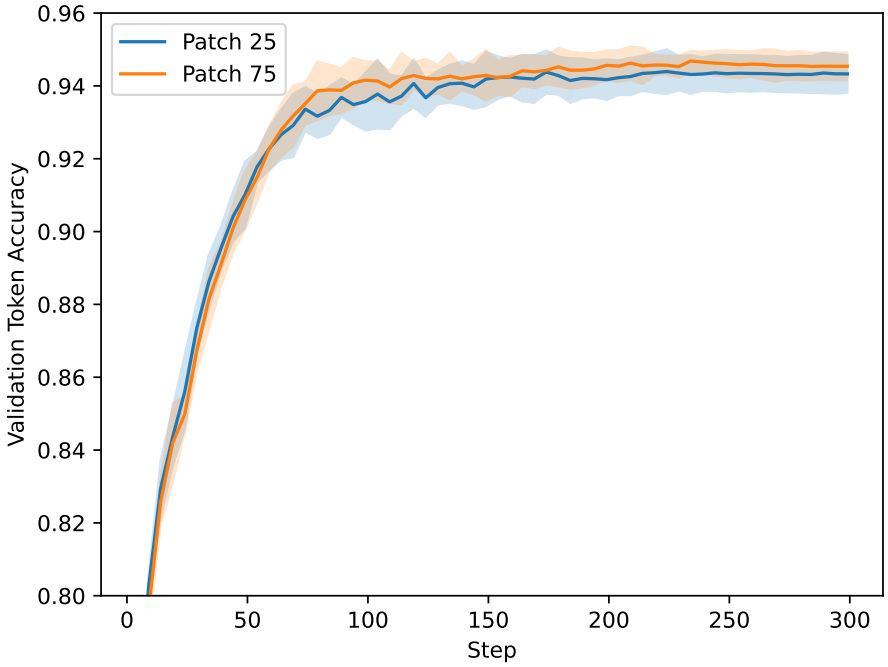


Figure 2: Token Accuracy of a model with trained patch size 25 and 75. The average token accuracy across five cross validations is plotted.

3 Transfer Function

3.1 Training

We trained a model to represent the transfer function from simulated to experimental IR spectra on a total of 2,000 paired simulated experimental spectra. We split the data into a 80/20 train and test split and train for a total of 50 epochs. Adam is used as optimizer with β_1=0.9 and β_2=0.999. In addition to the simulated IR spectrum, we added a 2048-dimensional Morgan fingerprint[1,2] of the molecule as well as an occurrence vector of the functional groups as defined by Jung *et al.*[3] We ablate the number of layers (4 and 6), the dimension of the bottleneck layer (128, 256 and 512), the learning rate (1e-3 and 1e-4), different final activation functions (sigmoid, absolute smoothing and exponential) and various loss functions (SID and MAE). Definitions for the final activation and loss functions are given below.

3.2 Final Activation Functions

Sigmoid

We used the standard sigmoid function as defined as: $Sigmoid(x) = \frac{1}{1+exp(-x)}$.

Absolute Smoothing

Here we first take the absolute of the output of the model, followed by applying a learned convolutional kernel of size five. Finally we take the absolute again. Defined as: $Abs\_Smoothing(x) = |Conv1D(|x|)|$.

Exponential

Standard exponential. Defined as $exp(x) = e^x$.

3.3 Loss Functions

Mean Absolute Error (MAE)

MAE defined as:

$$MAE(y_{pred}, y_{target}) = \frac{\sum_{i=1}^{n} |y_{pred} - y_{target}|}{n}$$

Spectral Information Divergence (SID)[4]

SID first introduced by McGill *et al.*[4].

$$SID(y_{pred}, y_{target}) = \sum_{i=1}^{n} y_{pred,i}\ ln\left(\frac{y_{pred,i}}{y_{target,i}}\right) + y_{target,i}\ ln\left(\frac{y_{target,i}}{y_{pred,i}}\right)$$

Table 2: Hyperparameter tuning for the transfer function. Parameters tuned include the loss and activation functions, learning rate, number of layers and the size of the bottleneck layer. The models were evaluated using MSE, MAE and SID. In increasing order of the sum of all three metrics.

Loss function	Activation function	Learning Rate	n–Layers	Bottleneck Layer dim.	MSE	MAE	SID
SID	Sigmoid	0.001	4	258	0.005091	0.02972	0.03489
MAE	Abs. Smoothing	0.001	4	258	0.005072	**0.02921**	0.03561
MAE	Abs. Smoothing	0.001	4	512	**0.004959**	0.03072	**0.03440**
SID	Exponential	0.001	4	258	0.005688	0.03095	0.03650
SID	Sigmoid	0.001	4	512	0.005156	0.03261	0.03639
MAE	Abs. Smoothing	0.001	6	258	0.006153	0.03111	0.03997
SID	Exponential	0.001	4	512	0.006043	0.03287	0.03884
MAE	Abs. Smoothing	0.001	6	512	0.006462	0.03133	0.04167
SID	Sigmoid	0.0001	4	512	0.006618	0.03358	0.04098
SID	Abs. Smoothing	0.001	6	512	0.006613	0.03354	0.04111
SID	Abs. Smoothing	0.001	4	512	0.006530	0.03333	0.04231
SID	Sigmoid	0.0001	4	258	0.006717	0.03398	0.04156
MAE	Abs. Smoothing	0.0001	4	512	0.005933	0.03414	0.04403
SID	Sigmoid	0.001	6	258	0.006863	0.03390	0.04400
SID	Exponential	0.001	6	512	0.007226	0.03386	0.04371
SID	Exponential	0.0001	4	512	0.007073	0.03484	0.04396
SID	Sigmoid	0.001	6	512	0.006919	0.03517	0.04394
SID	Abs. Smoothing	0.001	6	258	0.007129	0.03515	0.04426
MAE	Sigmoid	0.0001	4	258	0.007252	0.03339	0.04630
SID	Abs. Smoothing	0.0001	4	512	0.006640	0.03536	0.04529
MAE	Abs. Smoothing	0.0001	6	258	0.007457	0.03355	0.04651
MAE	Sigmoid	0.0001	4	512	0.007119	0.03460	0.04599
SID	Exponential	0.0001	4	258	0.007415	0.03588	0.04611
SID	Exponential	0.001	6	258	0.007022	0.03720	0.04521
SID	Abs. Smoothing	0.001	4	258	0.007178	0.03647	0.04791
MAE	Exponential	0.0001	4	512	0.007963	0.03468	0.05107
MAE	Exponential	0.0001	4	258	0.008008	0.03488	0.05148
SID	Sigmoid	0.0001	6	512	0.008078	0.03638	0.04993
SID	Exponential	0.0001	6	512	0.008004	0.03701	0.05006
SID	Sigmoid	0.0001	6	258	0.008241	0.03635	0.05092
MAE	Abs. Smoothing	0.0001	6	512	0.008168	0.03608	0.05291
SID	Abs. Smoothing	0.0001	6	512	0.008023	0.03821	0.05180
SID	Exponential	0.0001	6	258	0.008388	0.03772	0.05193
SID	Abs. Smoothing	0.0001	6	258	0.007995	0.03886	0.05258
MAE	Abs. Smoothing	0.0001	4	258	0.007473	0.03814	0.05796
MAE	Sigmoid	0.0001	6	258	0.009770	0.03806	0.06579
SID	Abs. Smoothing	0.0001	4	258	0.008790	0.04519	0.07062
MAE	Sigmoid	0.0001	6	512	0.01071	0.03985	0.08027
MAE	Exponential	0.0001	6	258	0.01085	0.04148	0.08268
MAE	Sigmoid	0.001	6	512	0.02235	0.06158	0.5496
MAE	Sigmoid	0.001	6	258	0.02149	0.06072	0.5575
MAE	Sigmoid	0.001	4	258	0.02185	0.06108	0.5573
MAE	Exponential	0.0001	6	512	0.01887	0.05805	0.5672
MAE	Sigmoid	0.001	4	512	0.02011	0.05934	0.5671
MAE	Exponential	0.001	4	512	0.01916	0.05839	0.5741
MAE	Exponential	0.001	4	258	0.01916	0.05839	0.5741
MAE	Exponential	0.001	6	258	0.01916	0.05839	0.5741
MAE	Exponential	0.001	6	512	0.01916	0.05839	0.5741

4 Performance analysis compared to baselines

In the following sections, we provide a detailed performance analysis of our original model[5], the model by Wu *et al.*[6] and the model presented in this work. We analyse the performance of each model with regard to the heavy atom count, functional groups and the Tanimoto similarity between the predicted and ground truth molecules. The analyses are carried out on the performance of the models on the 6-13 heavy atom subset of the NIST database. Results across the five cross-validation folds are averaged.

4.1 Heavy Atom Count

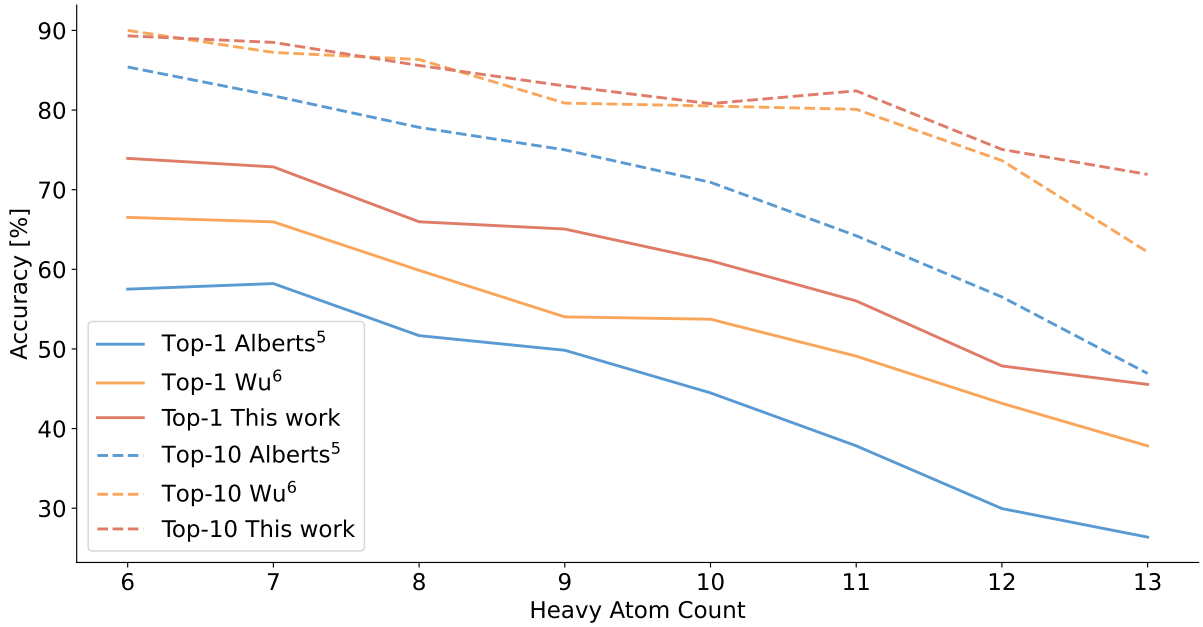
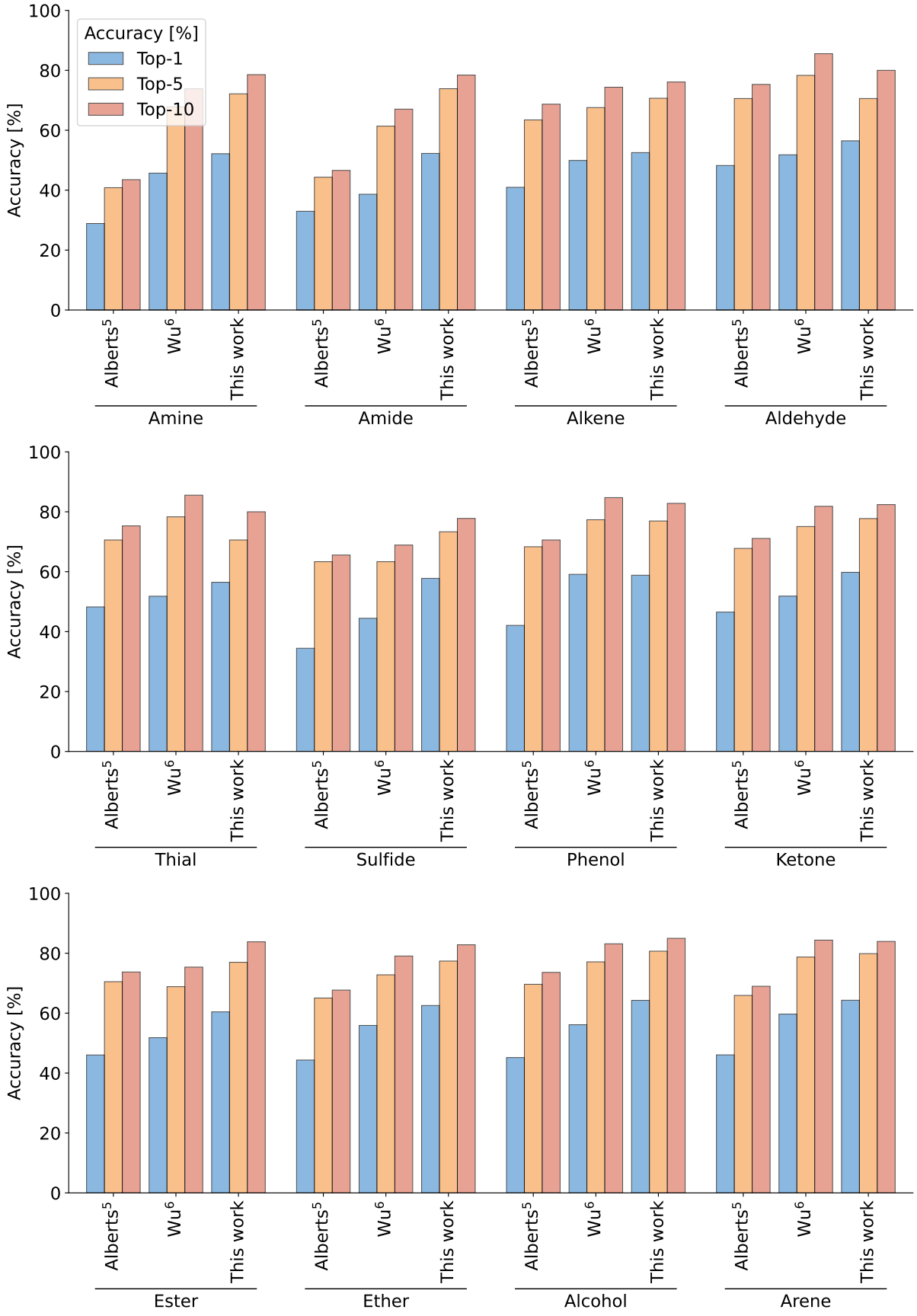


Figure 3: Top–1 and Top–10 performance against heavy atom count for the three models.

Shown in Figure 3 is the performance of all three models against the heavy atom count. Across all models, a decrease in performance with the heavy atom count is observed, stemming from the higher complexity of larger molecules and more noisy spectra. With regards to the Top–1 accuracy, our model outperforms both other models. The model by Wu *et al.*[6] and ours are comparable, in terms of Top–10 accuracy for molecules with six to ten heavy atoms. For larger molecules our model starts to outperform the one by Wu *et al.*[6]. Interestingly, the Top–1 performance of the model presented in this work is comparable to the Top–10 performance of our previous work at a heavy atom count of 13, demonstrating the robustness of our revised architecture and training pipeline.

4.2 Functional Group Performance


Accuracy [%]
Top-1
Top-5
Top-10
Accuracy [%]
0
20
40
60
80
100
Alberts[5]
Wu[6]
This work
Amine
Amide
Alkene
Aldehyde
Thial
Sulfide
Phenol
Ketone
Ester
Ether
Alcohol
Arene

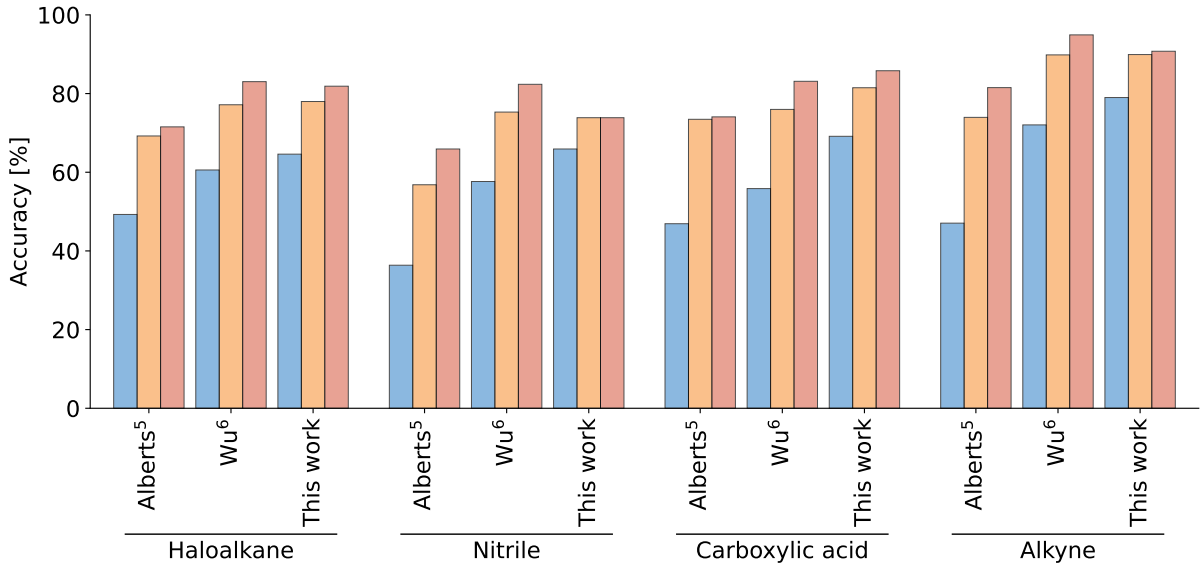


Figure 4: Performance of the models on subsets of molecules containing a specific functional group. The Top–1, Top–5 and Top–10 accuracies are plotted.

To investigate where our model improves compared to the two baselines we analyse the performance of our model, the one by Wu *et al.*[6] and our original model[5] on subsets of molecules containing 16 different functional groups. Only functional groups are considered with more than 80 occurrences across the experimental dataset. The results are shown in Figure 4 in order of increasing Top–1 performance. Our model outperforms the Top–1 accuracy of the model by Wu *et al.*[6] across all functional groups and the same is the case for the Top–5 and Top–10 accuracy for 11 out of 16 functional groups.

4.3 Tanimoto Similarity

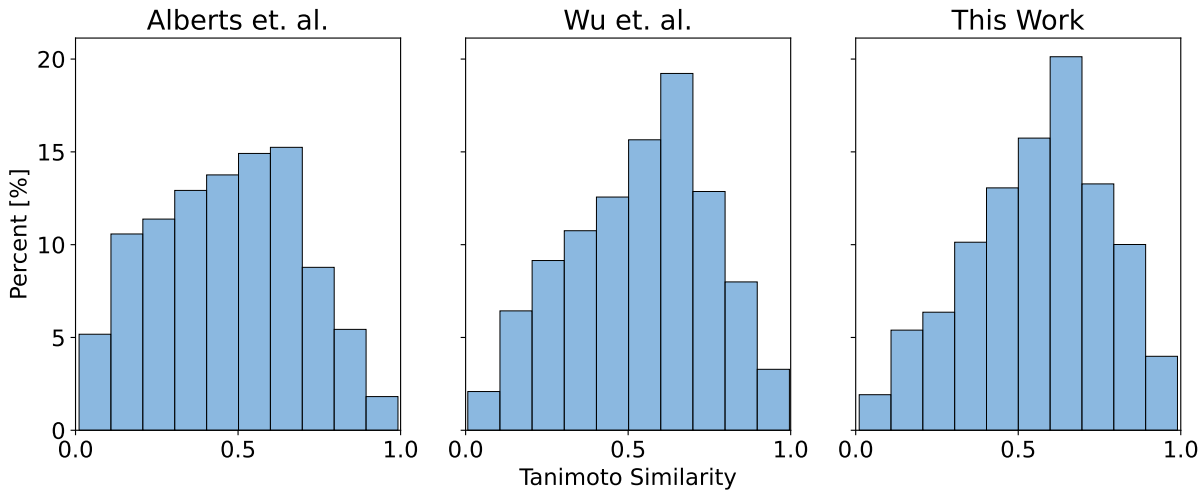


Figure 5: Tanimoto similarity between the Top–5 predicted molecules and the ground truth for the three models. Only samples are considered in which the ground truth is not present in the Top–5 predictions.

Figure 5 show the Tanimoto similarity between the Top–5 predicted molecules and the ground truth. To analyse the extent to which the predicted molecules are close to the ground truth in the case of model failure, only cases were considered in which the ground truth is not present in the Top–5 predictions. The distribution is similar for the model by Wu *et al.*[6] and the model presented in this work, with both significantly outperforming our earlier work[5]. The median Tanimoto similarity lies at 0.476 for our earlier work, 0.560 for the model by Wu *et al.*[6] and 0.583 for the model presented in this work, demonstrating that even if the model does not predict the correct molecule among the Top–5 predictions the generated molecules are still very similar to the ground truth.

References

[1] H. L. Morgan, *Journal of Chemical Documentation*, 1965, **5**, 107–113.

[2] D. Rogers and M. Hahn, *Journal of Chemical Information and Modeling*, 2010, **50**, 742–754.

[3] G. Jung, S. G. Jung and J. M. Cole, *Chemical Science*, 2023, **14**, 3600–3609.

[4] C. McGill, M. Forsuelo, Y. Guan and W. H. Green, *Journal of Chemical Information and Modeling*, 2021, **61**, 2594–2609.

[5] M. Alberts, T. Laino and A. C. Vaucher, *Communications Chemistry*, 2024, **7**, 1–11.

[6] W. Wu, A. Leonardis, J. Jiao, J. Jiang and L. Chen, *The Journal of Physical Chemistry*, 2025, **129**, 2077–2085.