Supplementary Information

Transfer learning based on atomic feature extraction for the prediction of experimental ^{13}C chemical shifts

Žarko Ivković[1,2], Jesús Jover[1], and Jeremy Harvey[2]

[1]Institut de Química Teòrica i Computacional (IQTC), Faculty of Chemistry, University of Barcelona, Spain

[2]Department of Chemistry, KU Leuven, Belgium

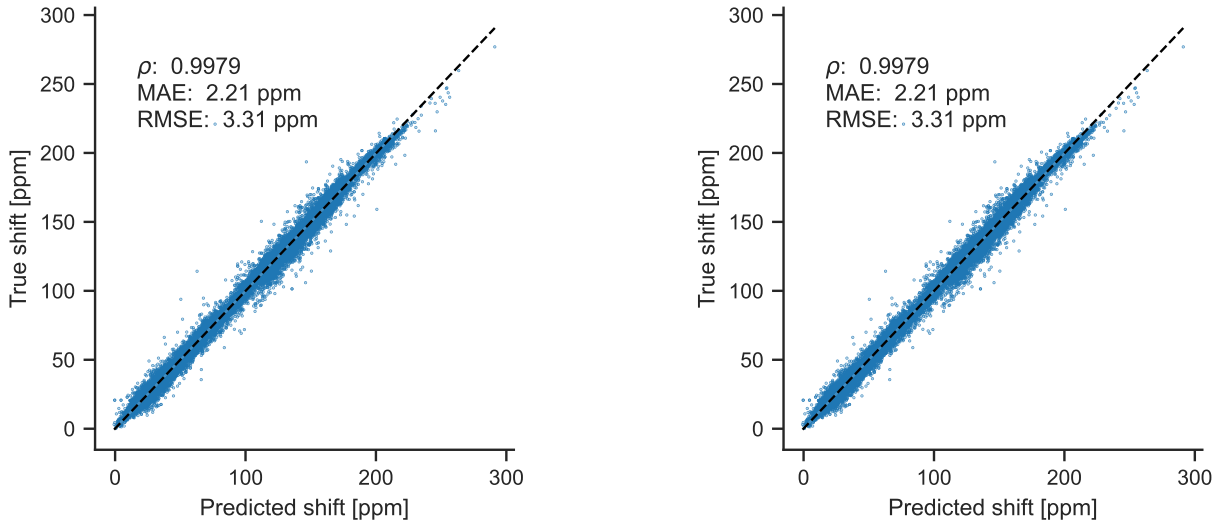
June 20, 2024

Contents

1 Atom order distortion between Exp5K and DFT8K

Certain carbon atoms exhibit DFT shifts within the 0-10 ppm range. A detailed inspection of the *exp_ dft_ outlier* file reveals that the atom labels were altered, with these atoms being hydrogen atoms in the DFT8K dataset but carbon atoms in the NMR8K dataset. This label distortion likely occurred due to the addition of explicit hydrogen atoms in functional groups, where they were previously implicit, using the *AddHs* function of rdkit during data transformation.

Atoms in any chemical structure can be ordered uniquely, as in the SMILES canonization process. Consequently, there is also a unique mapping of two different atom labelings of the same chemical structure. Rdkit stores the mapping from any atom labeling to canonical labeling in the *_smilesAtomOutputOrder* property of the canonicalized molecule. If we denote the mapping of the experimental structure as f and the DFT structure as g, then the mapping from experimental atom labels to DFT atom labels is $g \circ f$. Using this approach, we can correctly deduce MAE and RMSE of DFT predicted shift compared to experiments.



(a) Forcefield geometries

(b) DFT geometries

Figure 1: True against scaled DFT predicted chemical shifts of Exp5K dataset. More accurate DFT geometries don't result in better shift prediction

2 Architectures

Models architecture are shown in Figure 2 and 3. Dropout is applied after each layer.

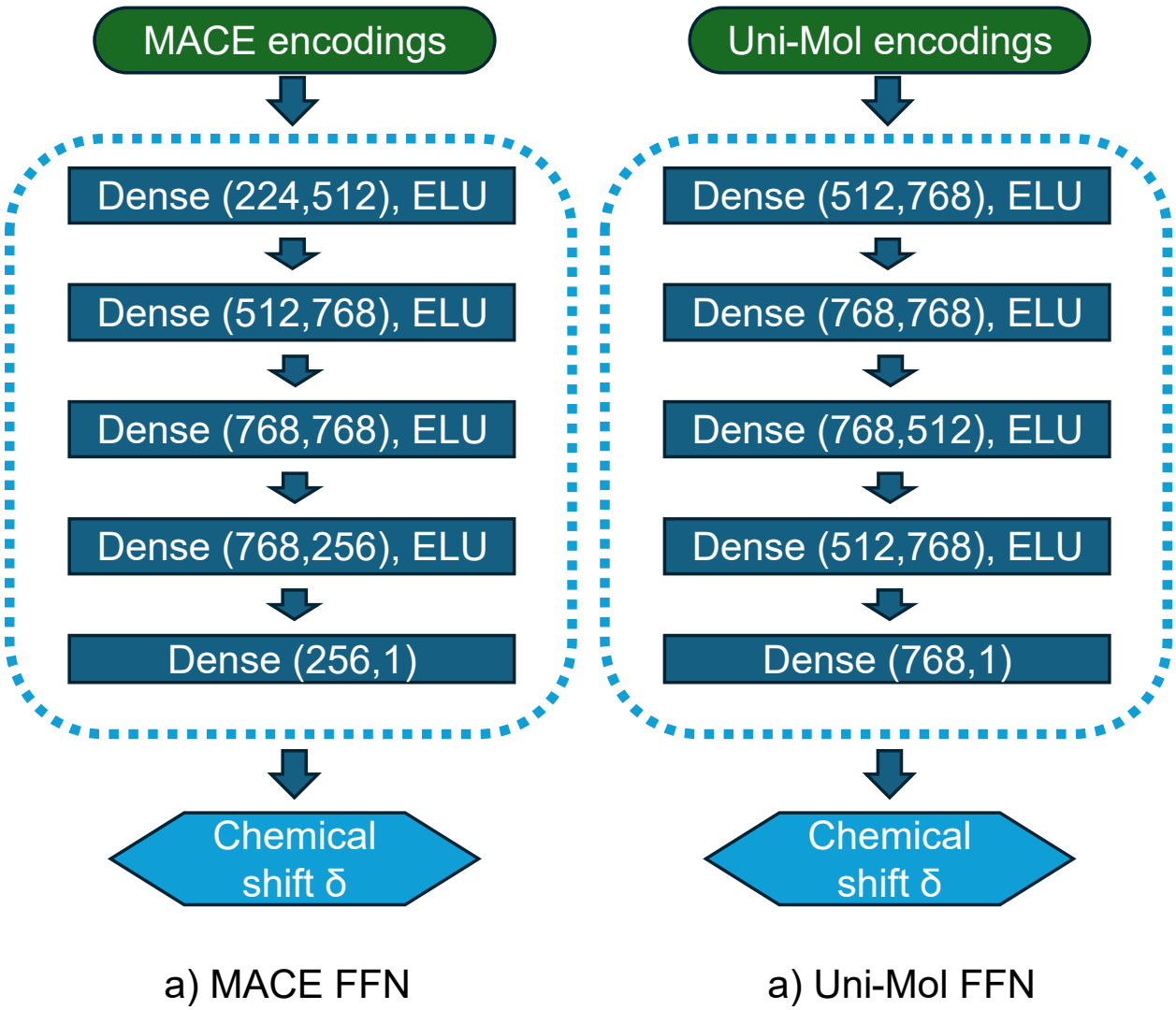


Figure 2: FFN models. Dropout layers are not shown. ELU = Exponential Linear Unit

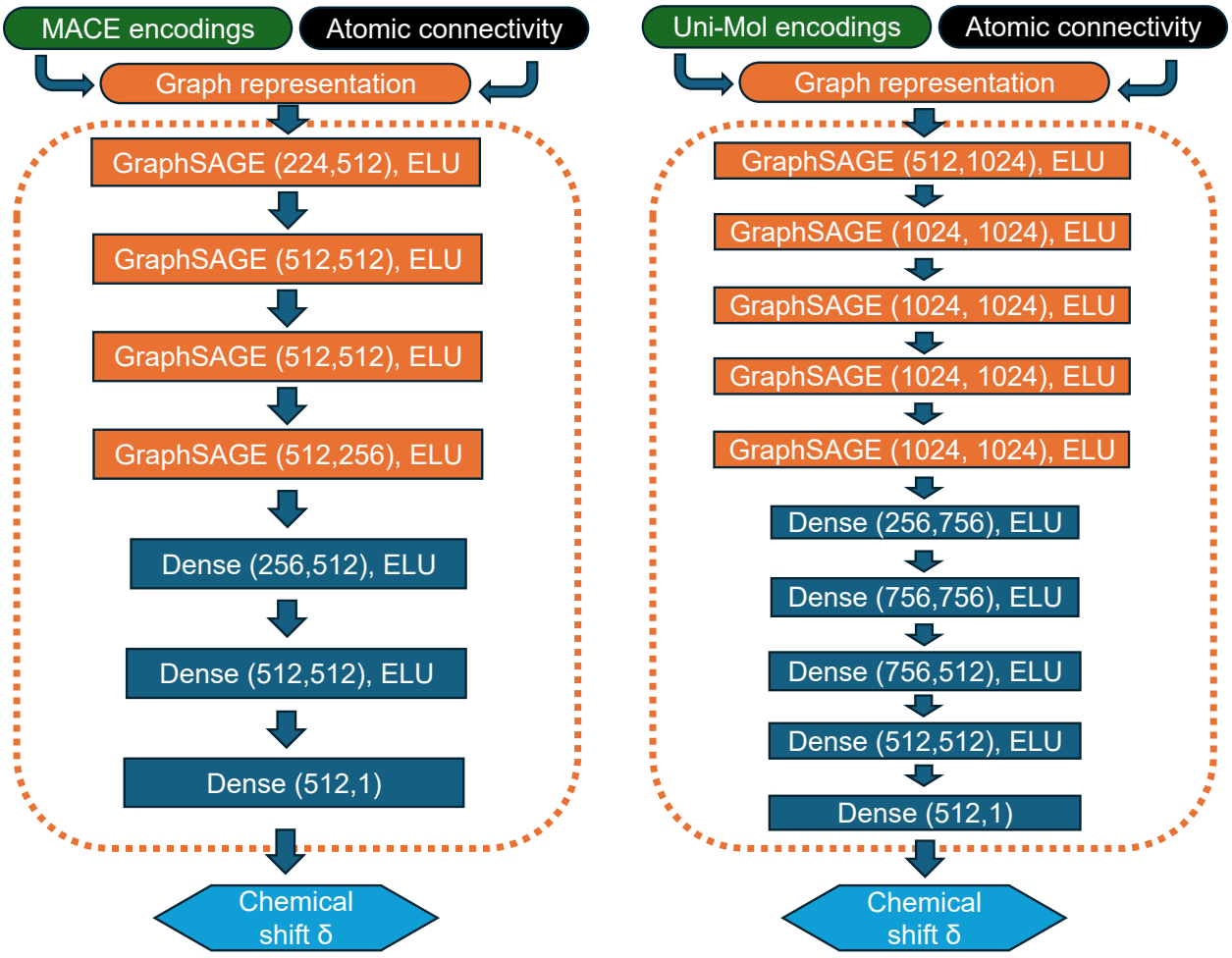


Figure 3: GNN models. Dropout layers are not shown. ELU = Exponential Linear Unit

3 Modified train and test sets

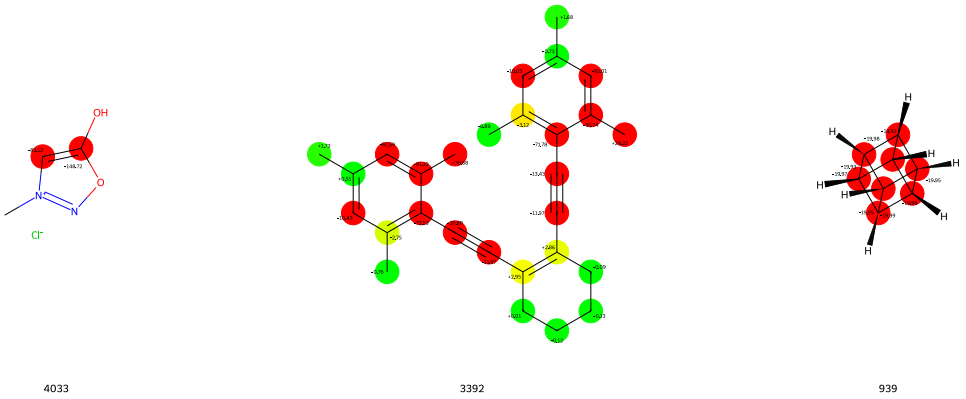


Figure 4: Molecules removed from the test set. The first molecule has the wrong graph connectivity(not shown), and the second molecule has the erroneous geometry with two methyl groups overlapping. The third molecule is removed only from the low-data regime test set due to erroneous behavior in MACE models when trained in low-data regimes, likely due to inaccurate geometry obtained from the forcefield.

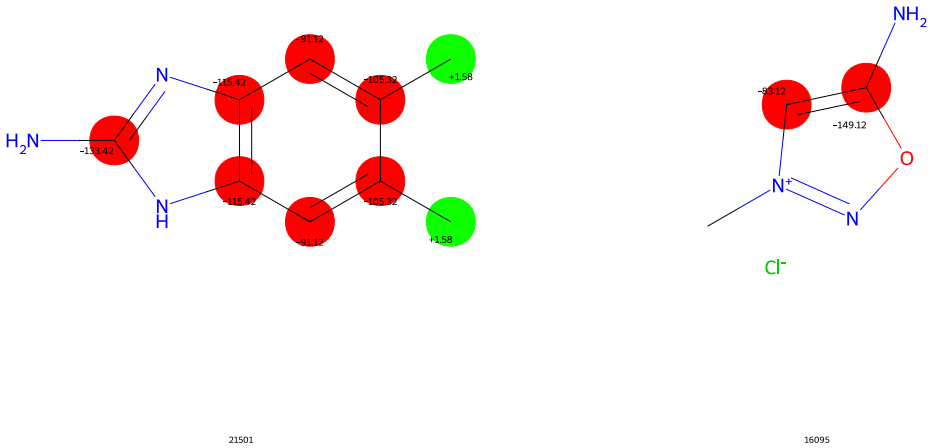


Figure 5: Molecules removed from the training set before sampling for low-data regimes. The first molecule has the wrong graph connectivity and lacks one hydrogen atom in the structure(not shown), while the second molecule has the wrong graph connectivity (not shown).

4 Training and hyperparameters

The hyperparameters for all models are listed in Table 1. A custom learning rate decay of 4% was applied every 15 epochs. The cost function was the mean absolute error. The AdamW optimizer, with a weight decay of 0.01, was used to minimize the cost function. A validation set comprising 10% of the training data was utilized. The train/validation split was done on molecule level. Once the optimal hyperparameters, including the number of epochs, were identified, the models were retrained from scratch using the entire training set.

Table 1: Hyperparameters and number of trainable parameters

Model	Initial LR	Batch size	Dropout rate	Epochs
MACE FFN	6e-4	96	0.10	1000
Uni-Mol FFN	8e-6	96	0.15	900
MACE GNN	1e-3	48	0.1	1000
Uni-Mol GNN	6e-4	64	0.15	900

5 Dataset summary

Table 2: Data Description of training and test data

Dataset	**N° spectra**	**N° atoms**		**N° C atoms**		**N° labeled at.**		**N° heavy at.**	
		average	range	average	range	average	range	average	range
train	21509	26.96	3-64	10.68	1-34	9.93	1-34	14.20	1-44
test	5386	26.74	5-64	10.62	1-33	9.88	1-32	14.14	2-38

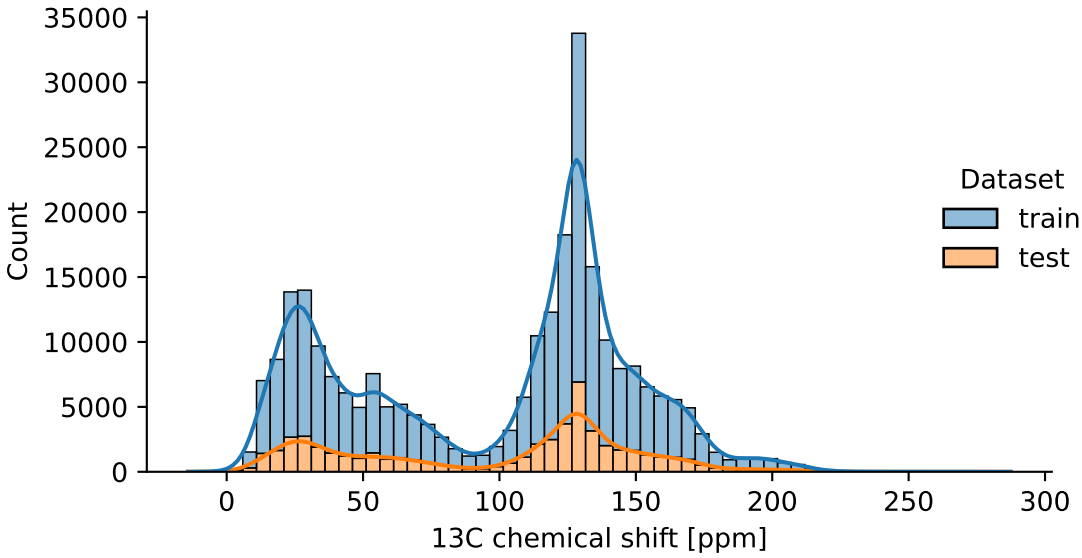


Figure 6: Distribution of chemical shifts

6 Performance of the models

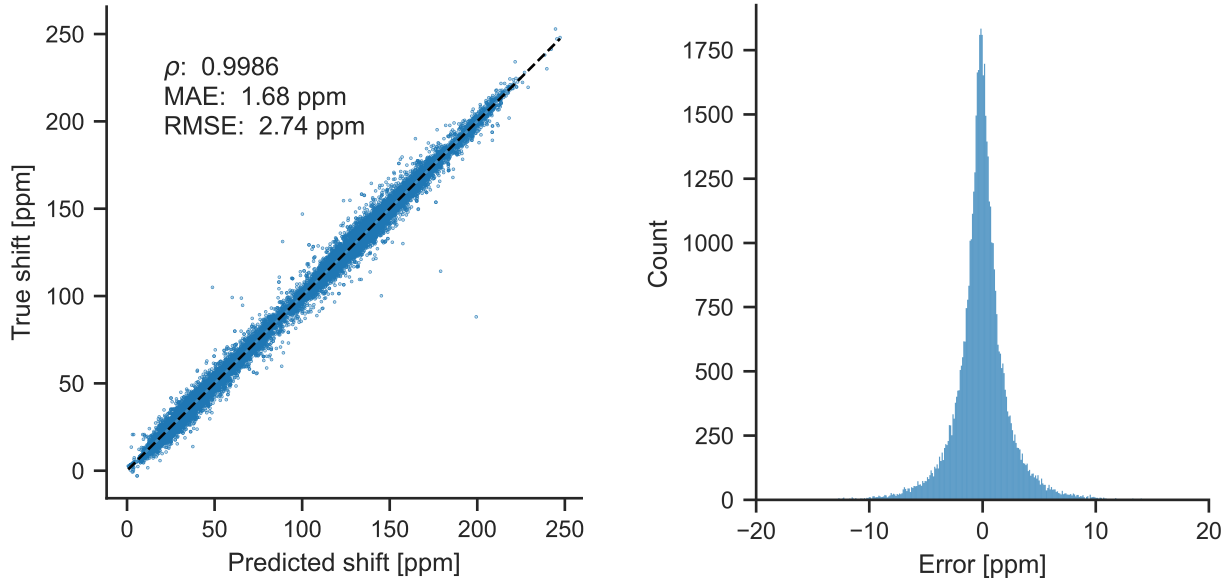


Figure 7: MACE FFN

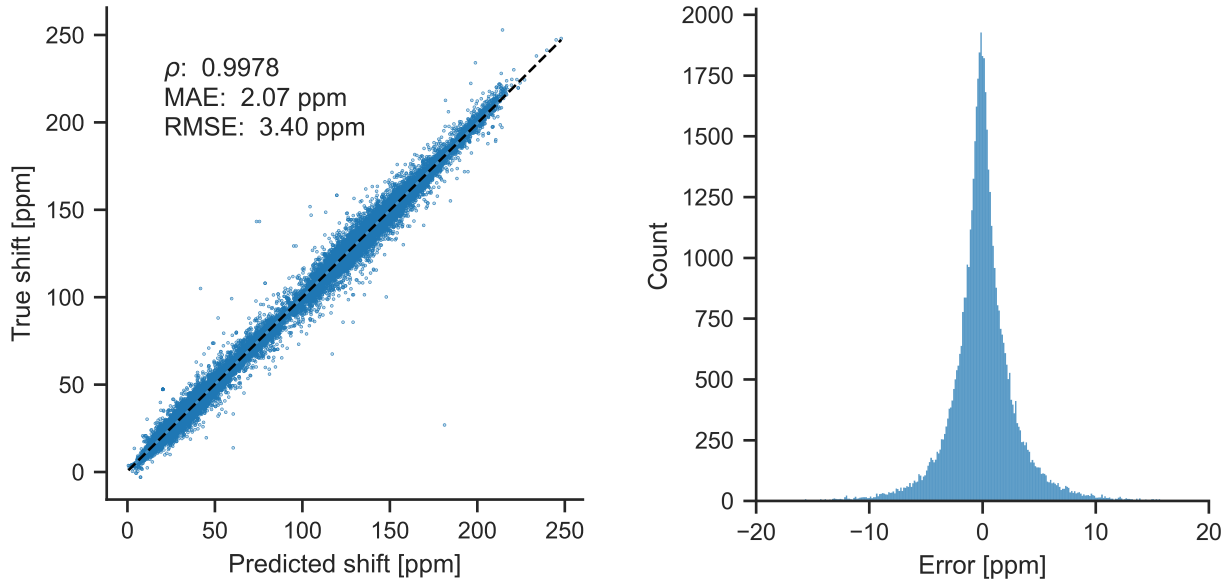


Figure 7: Uni-Mol FFN

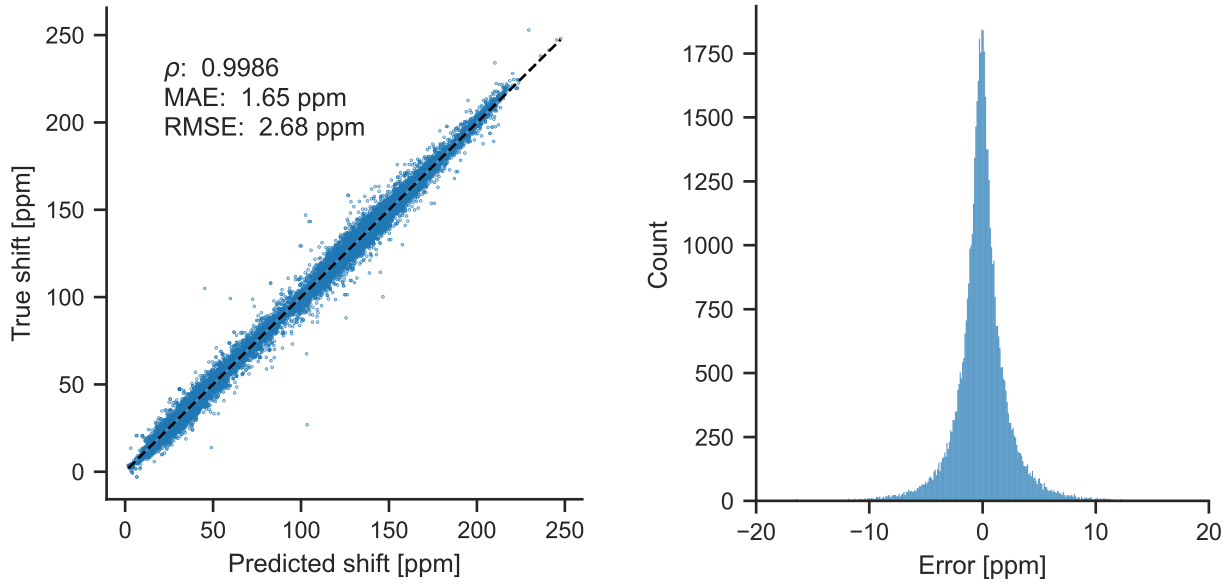


Figure 8: Ensemble MACE & Uni-Mol FFN

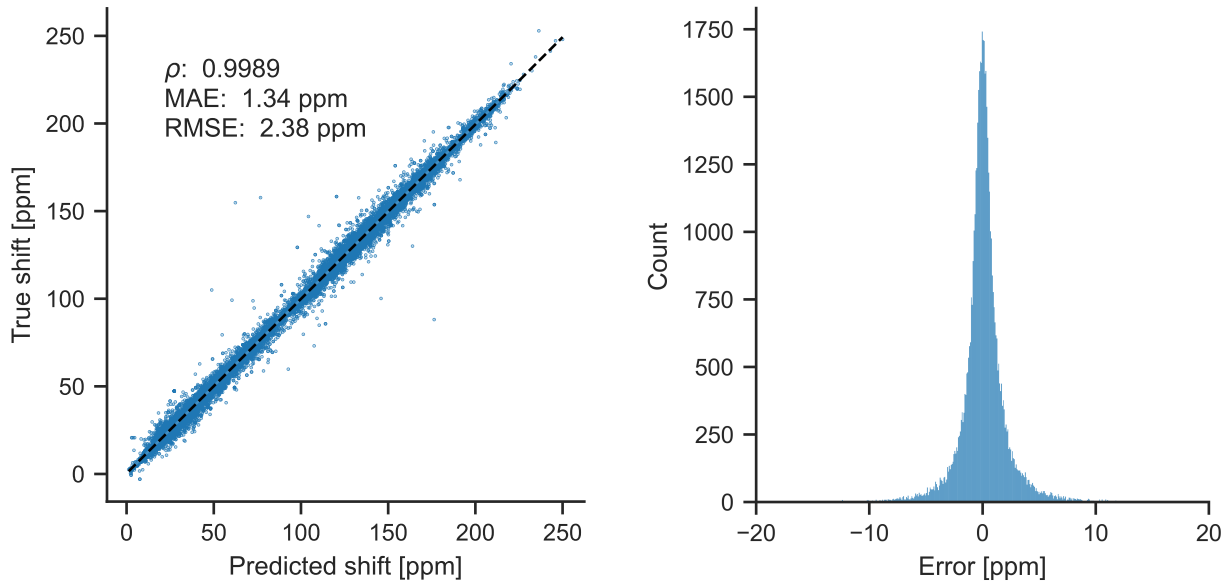


Figure 8: MACE GNN

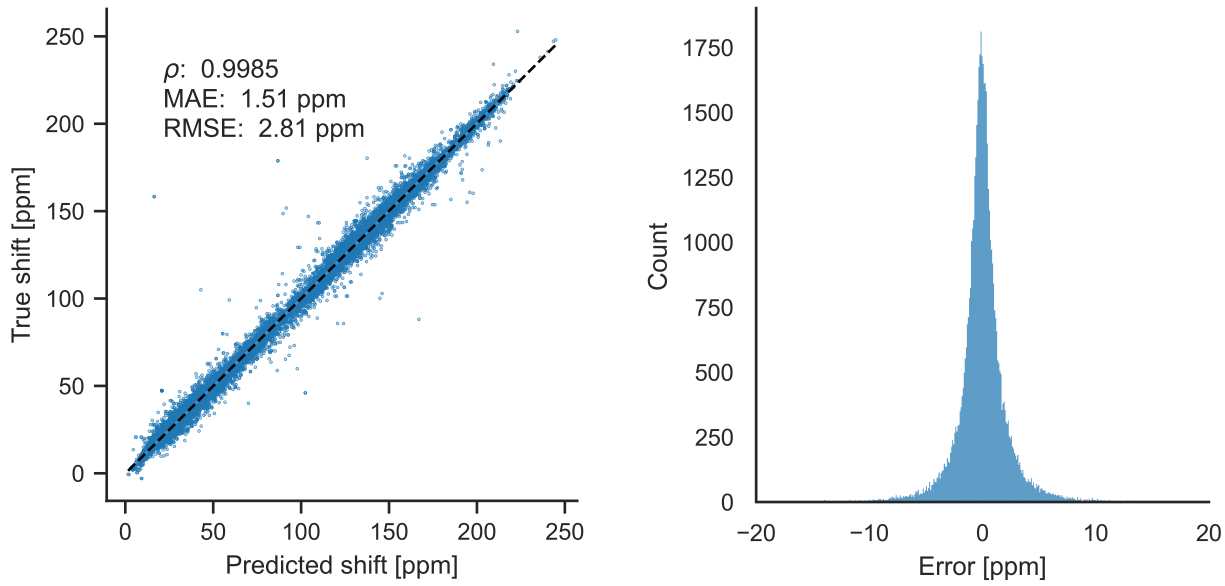


Figure 9: Uni-Mol GNN

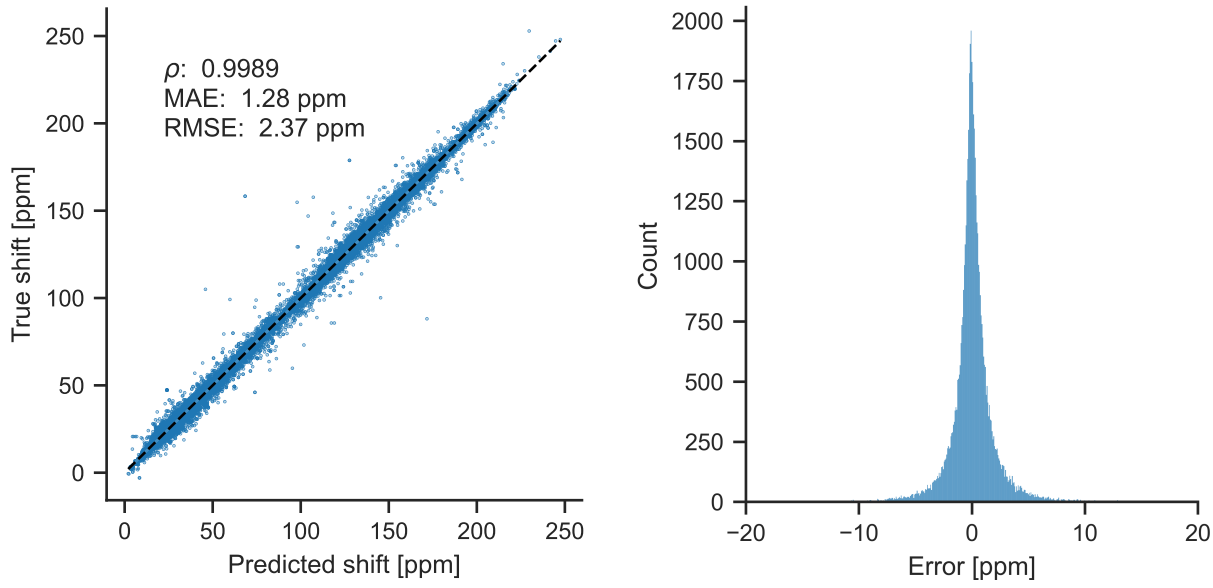


Figure 9: Ensemble MACE & Uni-Mol GNN

7 Molecules where models break down

For simplicity, we only show molecules where Ensemble MACE & Uni-Mol GNN and MACE GNN models fail. However, the code to produce similar plots for other models is available in the repository of this paper. The magnitude of errors is denoted by color: atoms with errors smaller than 1.5 ppm are green, while atoms with errors larger than 4.5 ppm are red. Errors in between are colored using linear interpolation.

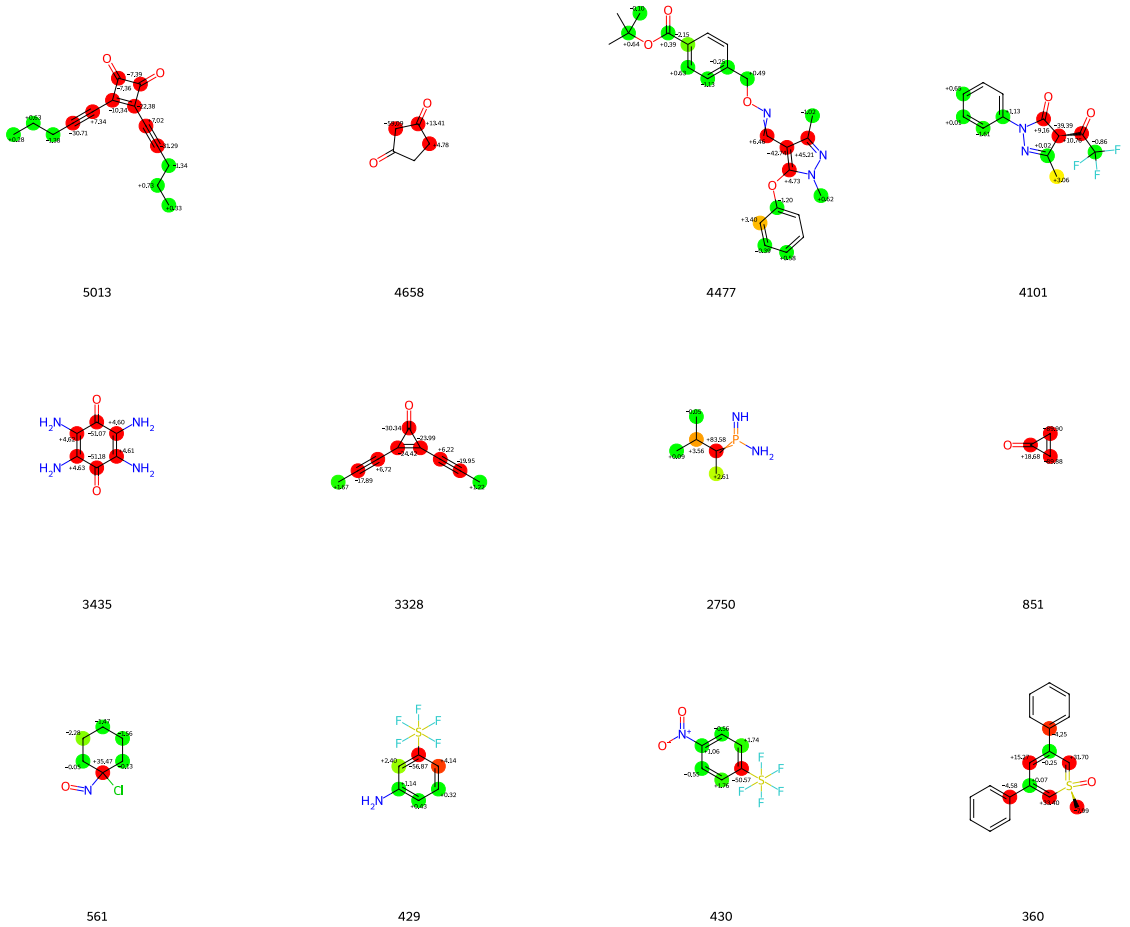


Figure 10: Molecules where Ensemble MACE & Uni-Mol GNN predicts at least one shift with absolute error larger than 30 ppm.

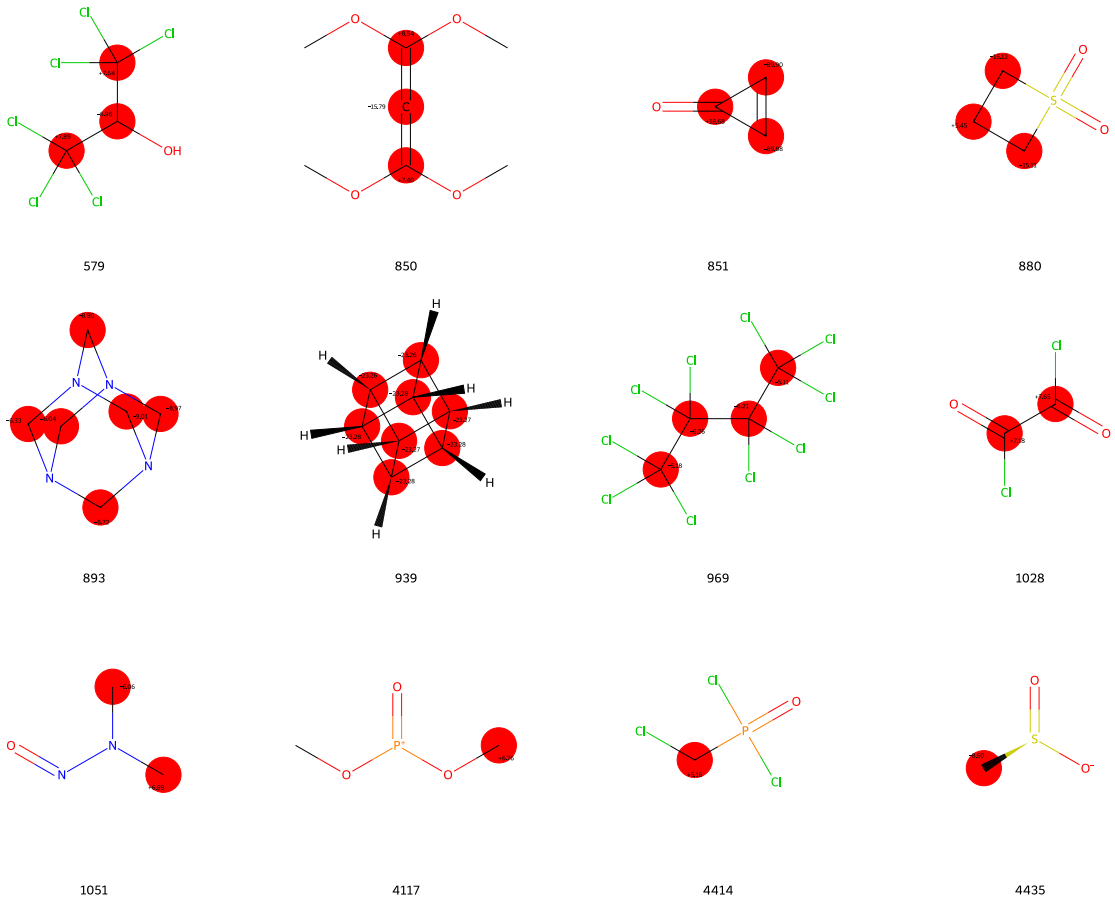


Figure 11: Molecules where Ensemble MACE & Uni-Mol GNN predicts all shifts with the absolute error larger than 5 ppm.

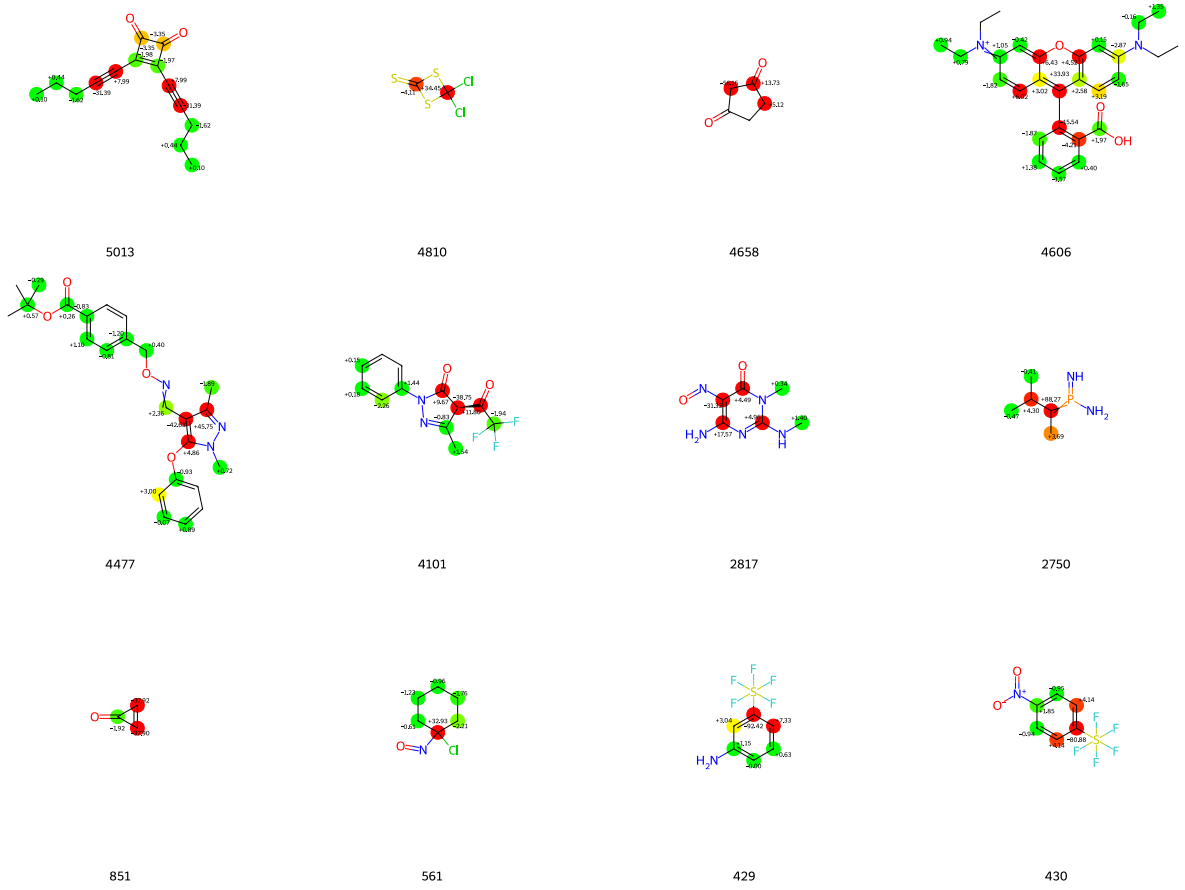


Figure 12: Molecules where MACE GNN predicts at least one shift with absolute error larger than 30 ppm.

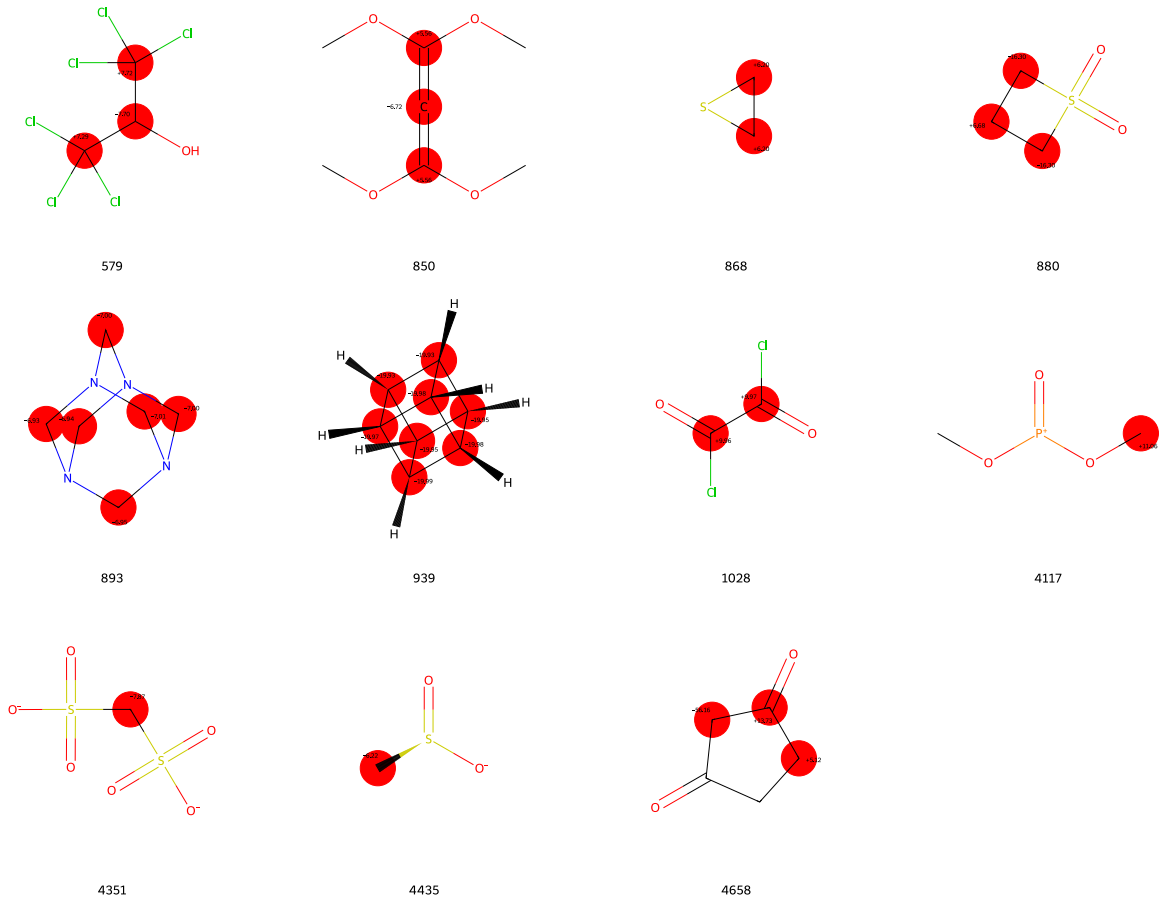


Figure 13: Molecules where MACE GNN predicts all shifts with the absolute error larger than 5 ppm.