TRANSFER LEARNING BASED ON ATOMIC FEATURE EXTRACTION FOR THE PREDICTION OF EXPERIMENTAL ^{13}C CHEMICAL SHIFTS

Žarko Ivković
Institut de Química Teòrica i Computacional (IQTC), Faculty of Chemistry
Univesity of Barcelona , zivkoviv7@alumnes.ub.edu
Department of Chemistry
KU Leuven

Jesús Jover
Institut de Química Teòrica i Computacional (IQTC)
Faculty of Chemistry
Univesity of Barcelona

Jeremy Harvey
Department of Chemistry
KU Leuven

ABSTRACT

This study indicates that atomic features derived from a message passing neural network (MPNN) forcefield are robust descriptors for atomic properties. A dense network utilizing these descriptors to predict ^{13}C shifts achieves a mean absolute error (MAE) of 1.68 ppm. When these features are used as node labels in a simple graph neural network (GNN), the model attains a better MAE of 1.34 ppm. On the other hand, embeddings from a self-supervised pre-trained 3D aware transformer are not sufficiently descriptive for a feedforward model but show reasonable accuracy within the GNN framework, achieving an MAE of 1.51 ppm. Under low-data conditions, all transfer-learned models show a significant improvement in predictive accuracy compared to existing literature models, regardless of the sampling strategy used to select from the pool of unlabeled examples. We demonstrated that extracting atomic features from models trained on large and diverse datasets is an effective transfer learning strategy for predicting NMR chemical shifts, achieving results on par with existing literature models. This method provides several benefits, such as reduced training times, simpler models with fewer trainable parameters, and strong performance in low-data scenarios, without the need for costly ab initio data of the target property.

1 INTRODUCTION

1.1 NMR CHEMICAL SHIFTS

NMR chemical shifts are valuable in the structure elucidation of organic compounds within classical and computer-assisted frameworks.(Stothers, 2012; Sternberg et al., 2004; Bagno & Saielli, 2007; Wu et al., 2023; Huang et al., 2021) Carbon chemical shifts have been used to elucidate reaction products(Michels et al., 2012), metabolites(DiBello et al., 2023), and natural products, including in the revision of the structures.(Rychnovsky, 2006; Sánchez-Martínez et al., 2023; Tantillo, 2013) Furthermore, chemical shifts carry information about the local chemical environments of atoms and have been used as descriptors for predicting chemical reactivityGordon et al. (2019); Guan et al. (2021) and in QSAR/QSPR modelsVerma & Hansch (2011). Prediction of carbon chemical shifts from the molecular structure has been extensively studied and many methods have been developed, ranging from ab initio to fully data-driven methods.(Jonas et al., 2022; Cortés et al., 2023)

Predicting carbon NMR shifts from molecular structures from the first principles is computationally intensive. First, the geometry is optimized, followed by calculating the electronic structure. In addition to errors from the electronic structure calculations, treatment of solvation, conformational

flexibility, and rovibronic effects introduce further errors.(Lodewyk et al., 2012) Considering all these factors comprehensively is computationally impractical at any level of theory that ensures reasonable accuracy. For example, even a basic DFT calculation of chemical shifts on an inexpensive geometry is too resource-intensive for large-scale rapid structure elucidation. The chosen functional, basis set, and solvation model influences the precision of DFT predictions for NMR shifts.(Benassi, 2017; Cimino et al., 2004) Although different results in the literature are reported on different sets for the same computational protocols, the best-reported protocol achieves a root mean square error (RMSE) of 3.68 ppm when compared to experimental shifts.(Benassi, 2017) This is insufficient for typical applications, as an initial investigation has shown that an accuracy of 1.1-1.2 ppm of MAE is necessary for correctly identifying 99% of molecules in the metabolomic database. (Yesiltepe et al., 2022)

The errors of DFT-predicted shifts have a systematic component that can be corrected using available experimental data. Lodewyk et al. (2012) developed a linear scaling protocol for different combinations of levels of theory, solvents, and solvation models, and their findings were compiled in the CHESHIRE repository.(Cheshire et al.) This became the standard for chemical shift prediction using DFT. Gao et al. (2020) went beyond linear interpolation and constructed a deep neural network that takes molecular structure and descriptors derived from calculated DFT shielding constants as input to predict experimental chemical shifts. Their method demonstrated superior performance, achieving an RMSE of 2.10 ppm, which is a significant notable improvement over the 4.77 ppm RMSE the authors report from linear regression on the same small test set.

The Exp5K dataset, developed as part of the CASCADE project,(Guan et al., 2021) is the largest dataset that compares empirically scaled DFT chemical shifts with experimental shifts. The authors excluded structures where DFT significantly disagreed with experimental results to avoid introducing noise from potential misassignments in the experimental data. This exclusion inevitably removes challenging examples where the disagreement arises from DFT's inability to accurately predict shifts due to molecular complexity. Additionally, the atom ordering was altered when comparing DFT with experimental shifts, leading to the unjustified exclusion of some examples from the dataset. After correcting the atom order, the calculated shifts deviate from the experiments with an MAE of 2.21 ppm and an RMSE of 3.31 ppm. (Appendix A) This should be considered the most realistic measure of the accuracy of DFT-calculated shifts corrected with linear scaling. These correction methods, along with others reported in the literature,(Sarotti & Pellegrinet, 2009; Xin et al., 2017) enhance the accuracy of predictions but do not reduce their computational cost.

On the other hand, data-driven methods are significantly faster by several orders of magnitude. The efficiency of machine learning in predicting carbon chemical shifts arises from the avoidance of expensive geometry optimizations or electronic structure computations. Nevertheless, the top models in the literature explicitly include geometrical data of the lowest energy conformers in their predictions.(Williams & Jonas, 2023; Guan et al., 2021; Han et al., 2022; Kwon et al., 2020) The compromise is achieved by utilizing inexpensive forcefield geometries instead of costly DFT-optimized geometries.

The accuracy of predictions in data-driven models is influenced by the quality and quantity of the training data.(Budach et al., 2022; Fan & Shi, 2022) By using experimental data for training, common errors in ab initio methods can be avoided. The most extensive open NMR shift database with fully assigned spectra is nmrshiftdb2.(Kuhn & Schlörer, 2015; Kuhn et al., 2024) User-contributed databases like this often face issues such as missing solvent and temperature details, peak misassignments, measurement noise, and incorrect structure identification. A model's performance is limited not only by the quantity but also by the quality of data. Thus, models that perform well in low-data scenarios are necessary when data is scarce and when prioritizing high-quality data over quantity.

1.2 TRANSFER LEARNING

Transfer learning involves using a model trained on one task as a foundation for training on another task, known as a downstream task. (Farahani et al., 2021) Generally, pre-training is performed on a similar task with a much larger dataset, followed by training on a smaller dataset for the specific task

of interest. Feature extraction and fine-tuning are two main implementations of transfer learning.[1] The choice of method depends on task similarity, the size and architecture of the pre-trained model, and the amount of available data. Feature extraction is commonly used in computer vision,(Kumar & Bhatia, 2014; Puls et al., 2023) while fine-tuning is widely used in language models.(Brown et al., 2020; Weng, 2024)

One of the major challenges for machine learning in chemistry is the scarcity of training data.(van Tilborg et al., 2024; Espley et al., 2023) Acquiring experimental and high-quality ab initio data is costly, and more affordable ab initio data often comes with substantial errors. Complex models, which are generally necessary to represent intricate chemical phenomena, demand a large amount of data for training. Integrating chemical and physical knowledge and intuition into the model architecture is one strategy to lessen the required training data.(Karniadakis et al., 2021) Transfer learning provides an alternative method to enhance models and can be used alongside other techniques to address issues related to limited data for chemical problems.

Most previous studies employ transfer learning for chemical models by initially training models on data generated from ab initio methods and then fine-tuning them on experimental data.(Guan et al., 2021; Han & Choi, 2021; Vermeire & Green, 2021) This quasi-transfer approach is effective if a significantly larger amount of ab initio data compared to the available experimental data can be produced. However, certain experimental properties like the smell, catalytic activity, and reaction yield are difficult or impossible to model using ab initio methods, while calculating others such as NMR properties, free energies, and absorption spectra can be prohibitively costly. In such cases, pre-training must be conducted on less relevant tasks where it is feasible to generate large-scale datasets.

2 Related work

In the notable CASCADE study,(Guan et al., 2021) graph neural networks (GNN) were employed to predict experimental chemical shifts. The ExpNN-ff model takes 3D structures optimized using MMFF forcefield as the way to incorporate geometrical information while maintaining relatively low computational cost. The authors implemented an interesting double-transfer learning training. First, the model was trained on DFT-optimized geometries and scaled DFT shifts. Second, the model was retrained on DFT-optimized geometries and experimental shifts, keeping the interaction layers frozen. Finally, the model was retrained again on forcefield geometries and experimental shifts, keeping the readout layers frozen. It is unclear what advantage this approach has over doing single-step transfer learning, updating all layers in the model simultaneously. Still, the ExpNN-ff model with an MAE of 1.43 ppm on a 500 hold-out test set performs better than the DFT with empirical scaling which has an MAE of 2.21 ppm on the whole training dataset of around 5000 compounds.

To avoid the costly DFT calculations for large molecules during the generation of the pre-training dataset, Han & Choi (2021) pretrained a GNN using the QM9 dataset of DFT shielding constants. They subsequently fine-tuned the model using an experimental chemical shifts database that includes larger molecules and atoms such as P, Cl, and S, which are absent in the QM9 dataset. The authors evaluated the model in low data scenarios, achieving an MAE of approximately 2.3 ppm with 2112 training examples. Nonetheless, the authors pre-trained on ab initio NMR data on a dataset comparable to the size of the experimental dataset used to fine-tune the model, similar to the approach used in CASCADE.

The first example of adopting true transfer learning for predicting chemical shifts was done in a recent work by El-Samman et al. (2024). The authors extracted atomic embeddings from the last interaction layer from the SchNet model trained to predict molecular energies on the QM9 dataset. The authors tested linear and feedforward network models for different chemical tasks, including predicting carbon chemical shifts calculated by HOSE codes. However, the dataset for the chemical shifts consisted of only 200 examples of shifts predicted by the HOSE code, so the performance relative to the literature models trained from scratch could not be assessed.

[1] In the literature, the term fine-tuning is not well-defined; it can refer to the second phase of training in general or to training models with weights initialized from other models. Here, we refer to the latter and simply call the second phase of training 'training,' as opposed to the 'pre-training' in the first phase.

To tackle low-data scenarios without resorting to transfer learning, Rull et al. (2023) modified a GNN architecture to enhance its efficiency in such conditions. While the modified architecture performed better in low-data scenarios than a similar GNN model, it significantly underperformed in high-data scenarios. This underscores the importance of considering the volume of training data when evaluating model performance and designing model architectures.

3 APPROACH

In an ideal situation, pre-training is performed on a highly similar task for which either more data is available or it is significantly cheaper to generate. However, such tasks are rarely available for any downstream chemical task, necessitating some form of compromise. Many of the latest pre-trained chemical models employ self-supervised pre-training tasks on huge unlabeled datasets of 2D chemical structures.(Ahmad et al., 2022; Ross et al., 2022; Xia et al., 2022; Rong et al., 2020) Conversely, there are numerous instances of quasi-transfer learning, involving pre-training on datasets of ab initio calculated properties of the size comparable to the available experimental datasets.(Guan et al., 2021; Han & Choi, 2021) We propose the atomic feature extraction from the models pre-trained for different chemical tasks on larger datasets, and we evaluate it by predicting experimental ^{13}C chemical shifts.

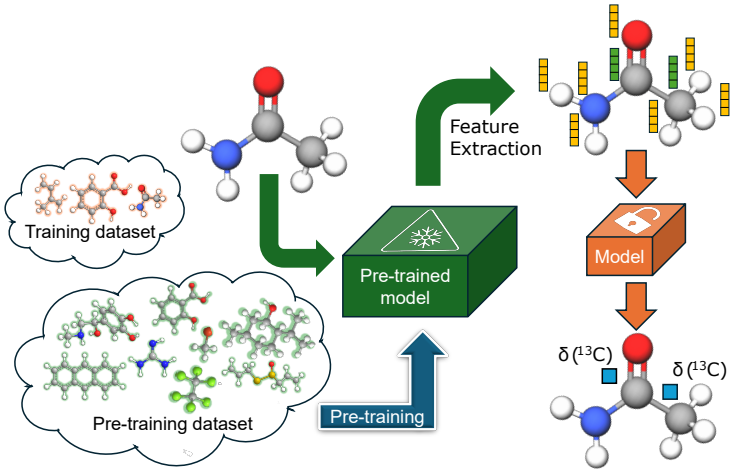


Figure 1: Transfer Learning based on atomic feature extraction.

3.1 CHOICE OF PRE-TRAINING TASK AND MODEL

The downstream task in this study is to predict the chemical shifts of carbon atoms. Predicting other atomic properties influenced by the chemical environment of the atom is the most relevant task. However, no other atomic properties have as extensive experimental data as chemical shifts. Fortunately, many models designed for predicting molecular properties incorporate atomic representations within their architectures.(Heid et al., 2024; Gilmer et al., 2017) Moreover, the pre-trained model must consider geometrical information since chemical shifts are influenced by molecular conformation. Therefore, most pre-trained models based on 2D molecular structures are not suitable candidates. This leads us to neural network forcefields, whose architectures are designed to sum atomic energy contributions.[1] We selected the MACE-OFF23 transferable organic forcefield(Kovács et al., 2023; Batatia et al., 2022), which is state-of-the-art for predicting DFT molecular energies, open-source, and trained on a reasonably large dataset. Since we are not concerned with inference time, we chose the large variant of the forcefield. The other model we tested is Uni-Mol(Zhou et al., 2023), a 3D-aware self-supervised pre-trained transformer known for its performance in downstream

[1]This architecture design is not mandatory. The only requirement for architecture is the presence of atomic embeddings within the model

molecular property prediction tasks. Although self-supervised pre-training is less directly related to atomic property prediction, it is done on an even larger dataset. The model includes atomic representation in its architecture, and integrates geometrical information in its embeddings, making it appropriate for this transfer learning approach. Finally, both pre-trained models are publicly available including their weights, making them ready to use without pre-training them again.

3.2 Feature extraction

We extract atomic embeddings from the first of two interaction layers in the large variant of the MACE-OFF23 forcefield. This approach contrasts with the method of El-Samman et al. (2024), where embeddings are extracted from the final interaction layer of the SchNet model. We suggest the extraction from initial layers works better when the similarity of tasks is low.(Schütt et al., 2017) We retain only the invariant portion of the embedding to ensure rotational and translational invariance, resulting in a 244-dimensional vector atomic embedding. Given that Uni-Mol is intended as a backbone pre-trained model for various downstream tasks, we directly extract the atomic representation from the output of the backbone, yielding a 512-dimensional vector per atom, invariant to translation and rotation. Both models use atomic coordinates and identities as inputs, akin to the input used by typical ab initio codes, and produce atomic embeddings for each atom as outputs.

3.3 Models architecture

We evaluated two distinct types of downstream models: a feedforward network (FFN) and a graph neural network (GNN). For the feedforward network, we assume that the pre-trained model has captured all necessary information regarding the chemical environment of each carbon atom. We use the embeddings of carbon atoms as input and train the network to predict chemical shifts. Additionally, we tested the GNN based on the GraphSAGE(Hamilton et al., 2017) architecture, which facilitates the exchange of information between different atomic environment embeddings. This leads to a more robust model as it can learn more relevant embeddings for NMR shifts. Unlike the other methods where fully connected graphs with a cutoff distance or graphs with implicitly represented hydrogens have been used, we used a chemical graph where all atoms are explicitly included. Consequently, GNN models require atomic connectivity as input, whereas FFN models only need atomic coordinates. Finally, after the message passing layers, the atomic embeddings of carbon atoms are fed into a readout feedforward network to predict chemical shifts. Both methodologies are illustrated in Figure 2. Finally, the ensembles of two models of the same type and different pre-training tasks, implemented as the average of the prediction of each model are also tested.

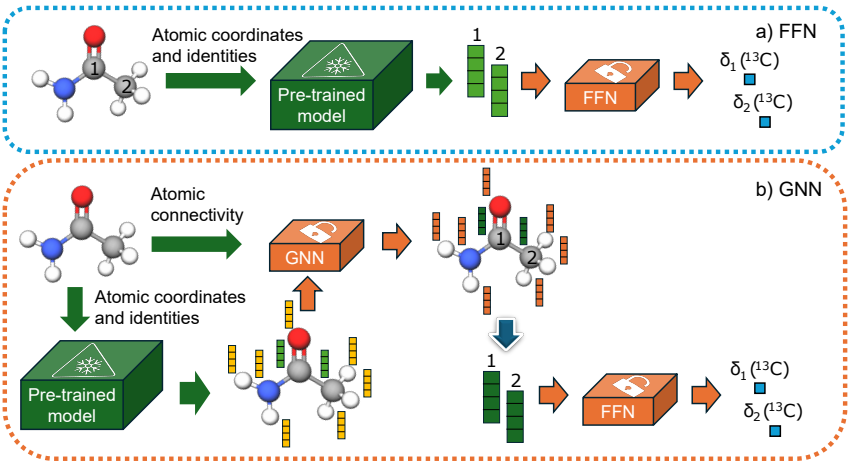


Figure 2: a) FNN model b) GNN model.
Only orange models are trained, while the green models' weights are frozen.

3.4 LOW-DATA REGIMES

To evaluate model performance with fewer training examples, we selected varying quantities of samples from the original dataset, treating it as a pool of unlabeled examples. Although this dataset is smaller than the typical molecular datasets of unlabeled molecules, it is sufficiently large to compare different sampling methods. We examined three sampling strategies: random sampling, MaxMin(Ashton et al., 2002) sampling based on the Tanimoto distance(Bajusz et al., 2015) between Morgan fingerprints(Rogers & Hahn, 2010), and MaxMin sampling based on the undirected Hausdorff distance(Birsan & Tiba, 2006) between sets of transferred embeddings of all carbon atoms in two molecules. The directed Hausdorff distance between two sets of vectors A and B is defined as:

$$h(A, B) = \max_{a \in A} \min_{b \in B} d(a, b)$$

where $d(a, b)$ is any distance metric between two vectors. However, the directed Hausdorff distance is not symmetric, so we use the undirected Hausdorff distance, employing the Euclidean distance as the distance metric d:

$$H(A, B) = \max\big(h(A, B), h(B, A)\big)$$
$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b||^2$$

In our scenario, sets of vectors represent sets of transferred embeddings of carbon atoms. While we could have used embeddings of all atoms, the carbon atom embeddings also convey information about their neighboring atoms. Since our primary interest lies in the differences in carbon atom environments between two molecules, we used only the embeddings of carbon atoms, which also reduces the computational cost, a crucial factor when sampling large pools of examples.

4 RESULTS

The mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation coefficient (ρ) for all models are presented in Table 2. The results are based on a modified test set, where we excluded a couple of broken examples from the original test set. (Appendix C). The ensemble of two independently trained GNN models performs the best, with the lowest MAE and RMSE. MACE models outperform their Uni-Mol equivalents significantly, indicating that the forcefield is an excellent option for the pre-training task. Even though the Uni-Mol GNN has a lower MAE than the MACE FFN model, its RMSE is higher, highlighting the necessity to report at least both MAE and RMSE when reporting the model's performance. Regarding parameter efficiency, MACE GNN is by far the best model.

Table 1: Performance on a test set and number of trainable parameters

Model	MAE [ppm]	RMSE [ppm]	ρ	N° params
MACE FFN	1.68	2.74	0.9986	1.3×10^6
Uni-Mol FFN	2.07	3.40	0.9978	1.8×10^6
Ensemble MACE & Uni-Mol FFN	1.65	2.68	0.9986	3.1×10^6
MACE GNN	1.34	2.38	0.9989	1.9×10^6
Uni-Mol GNN	1.51	2.81	0.9985	9.3×10^6
Ensemble MACE & Uni-Mol GNN	**1.28**	**2.37**	**0.9989**	1.0×10^7

A comparison with relevant literature models that take forcefield geometries as input is shown in Figure 3. The ensemble of two GNNs and MACE GNN performs equally well as the best-reported literature models. Comparison with models trained using the same train/test split is more reliable, and the FullSSPrUCe model is trained on the larger portion of the nmrshiftdb2 database, which explains its slightly better performance. In any case, since all reported models are solvent agnostic, it is clear that the accuracy has reached its limit because it is not unusual for ^{13}C shifts to differ by more than 1 ppm in different solvents.

The distinct advantages of our models are their simpler architectures(Appendix B) and fewer trainable parameters, which result in significantly reduced training time. We do not consider the parameters of pre-trained models because the entire training dataset can be encoded by pre-trained models

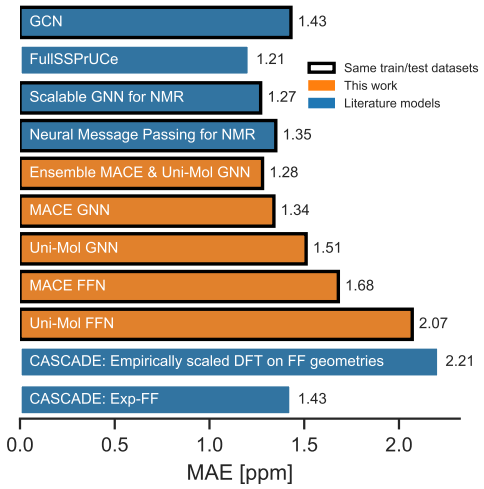


Figure 3: Comparison with the literature models.(Jonas & Kuhn, 2019; Williams & Jonas, 2023; Han et al., 2022; Kwon et al., 2020; Guan et al., 2021)

before training, making the training time independent of the number of parameters of the pre-trained model. However, the complexity of pre-trained models affects inference speed. Fortunately, the bottleneck in inference is conformer generation, so our models are faster to train and equally fast for inference.

4.1 Low-data regimes

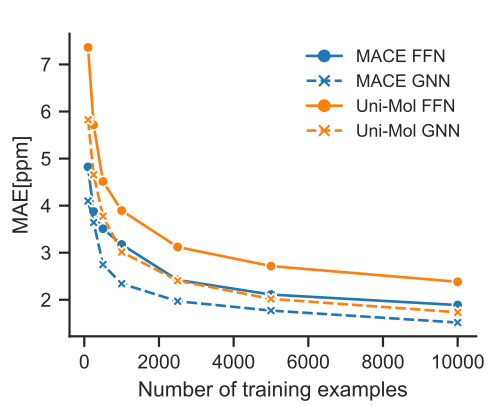
To simulate low-data regimes, we sampled data points from the training dataset, maintaining the same model architectures(Appendix B) as used in the full data scenario to emphasize the effectiveness of transfer learning. Nonetheless, the performance can be enhanced by optimizing hyperparameters for low-data regimes, especially by reducing model complexity and the dropout rate. Furthermore, an additional molecule was excluded from the test set because MACE-based models gave erroneous predictions for that molecule.(Appendix C)

Figure 6a illustrates that the performance of all models is improved with an increased number of training examples. Notably, the MACE FFN model outperforms the Uni-Mol GNN model in extremely low-data scenarios, whereas the reverse is true in high-data scenarios. The varying complexities of the models can explain this difference, as smaller models need less training data. Figure 6b compares models in this paper with a model that performs similarly on the full dataset, a model specifically designed for low-data scenarios, and a classical HOSE Code model.(Rull et al., 2023; Bremser, 1978) Transfer learning significantly boosts accuracy in low-data scenarios compared to models trained from scratch. Furthermore, there is no trade-off between performance in high-data and low-data scenarios, unlike in the 2019 model. (Rull et al., 2023)

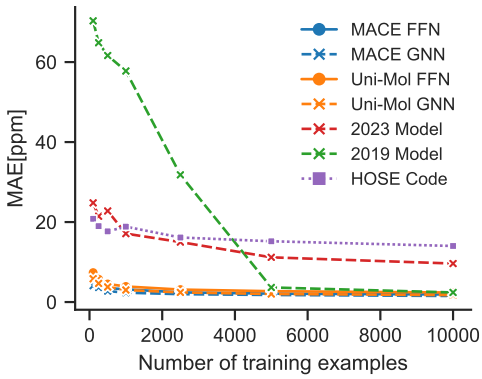
Conclusion

We introduced atomic feature extraction as a transfer learning method applicable to both atomic and molecular-level prediction tasks. Unlike previous quasi-transfer methods, this approach does not require generating ab initio data for the target property. Moreover, the only information needed are atomic coordinates and atomic connectivity.

We evaluated this method on the prediction of experimental ^{13}C chemical shifts, a well-studied atomic property prediction task. Our method performs on par with the best models trained from scratch and surpasses them in low-data scenarios. When using this transfer learning approach, we demonstrated that the details of the sampling strategy used to select from the pool of unlabeled



(a) This work.



(b) Comparison with literature models.(Rull et al., 2023; Bremser, 1978)

Figure 4: Low-data regimes simulated using random sampling.

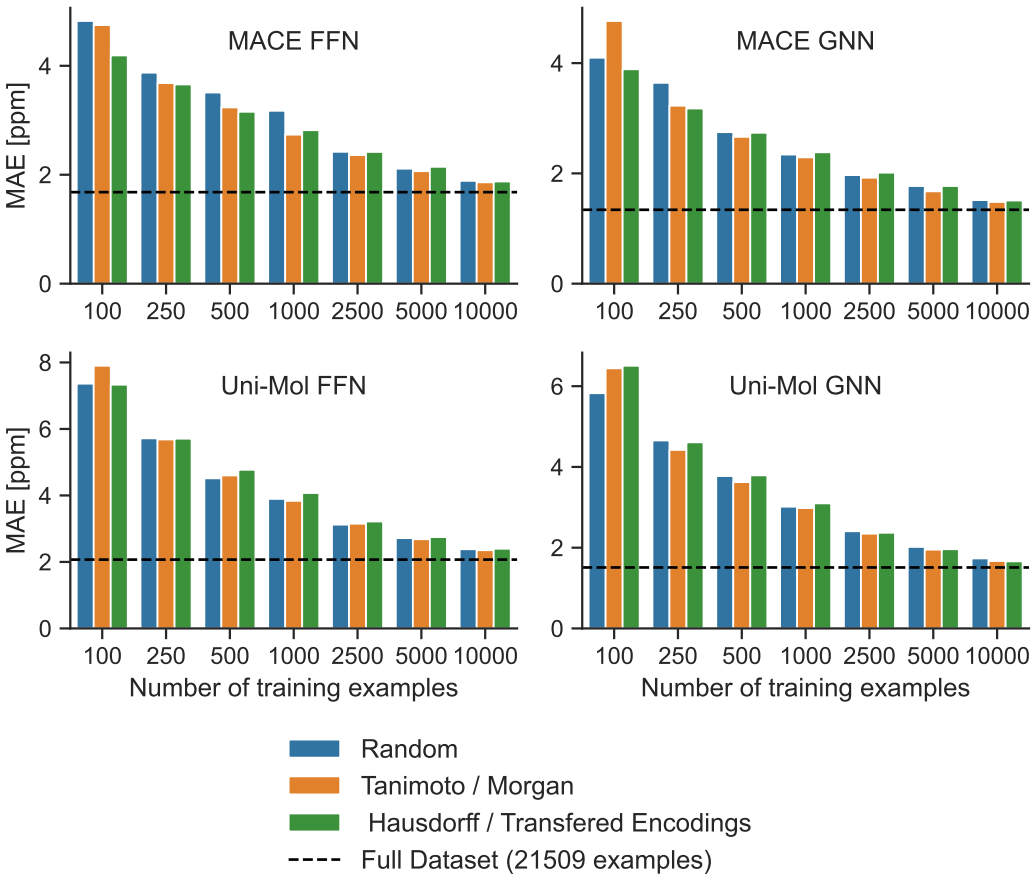


Figure 5: The effect of three different sampling strategies.

examples don't matter. Lastly, we identified the MPNN forcefield as a superior candidate for pre-trained models for transfer learning compared to self-supervised pre-trained models.

The proven efficacy in low-data scenarios reveals new potential uses for this transfer learning approach in chemical problems with limited experimental data and in tasks where plenty of data exists

but predictions are limited by data quality. For chemical shifts, employing more precise geometries and data with recorded solvents and peaks assigned through multiple spectra will enhance the accuracy of data-driven models. This enhancement is feasible only if models can be trained on less data, which can be achieved through the transfer learning method described here.

METHODS

DATA

The dataset utilized in this work is taken from Kwon et al. (2020) and is derived from the original dataset published by Jonas & Kuhn (2019) It includes a predefined train/test split. This dataset comprises molecules with experimental spectra from nmrshiftdb2, which contain elements H, C, O, N, P, S, and F, and have no more than 64 atoms. The molecular geometries are obtained as the lowest energy conformers found in EDTKG conformer search(Riniker & Landrum, 2015) followed by MMFF minimization(Halgren, 1996). Molecules that failed rdkit sanitization, likely due to version discrepancies, were excluded. A detailed summary of the resulting dataset is available in the supplementary information.(Appendix E)

MODELS

FFN models consist of simple fully connected layers with exponential linear unit (ELU) activation functions.(Clevert et al., 2015) The final layer is linear without any activation function. GNN models employ GraphSAGE message passing layers with ELU activation function, followed by a readout feedforward network of the same type as FFN models. Dropout was applied after each layer in all models.(Srivastava et al., 2014) The models were trained using L1 loss (mean absolute error) as the cost function and the AdamW optimizer with a weight decay of 0.01.(Loshchilov & Hutter, 2017) Hyperparameters were optimized through automated hyperparameter tuning and manual adjustments. Additional training and model architecture details can be found in the SI.(Appendix B and D)

COMPUTATIONAL DETAILS

We accessed the pre-trained models using code from the associated repositories. Rdkit(RDK; Landrum et al., 2024) (version 2023.09.5) was employed to process data, extract atomic connectivity from molecular structures, and perform MaxMin sampling. PyTorch(Paszke et al., 2019) (version 2.2.1) and PyTorch Lightning(Falcon & The PyTorch Lightning team, 2019) (version 2.2.1) were used for constructing and training FFN models, while PyTorch Geometric(Fey & Lenssen, 2019) (version 2.5.2) was used for GNN models. All models were trained on a single Nvidia L4 Tensor core GPU. MaxMin sampling and Morgan fingerprints with a radius of 3 were implemented using rdkit. The Hausdorff distance was calculated using the scipy package(Virtanen et al., 2020; Taha & Hanbury, 2015). Training for low-data examples continued until the validation loss ceased to decrease or until 800 epochs were reached. We sampled 120% of training data points for each regime, then randomly divided the data into train and validation sets. This ensured that the validation dataset size was always 20% of the training dataset size, and the train/validation split was performed as usual, making the conditions closer to a real low-data regime. Conversely, testing was conducted on the entire test set for a realistic performance evaluation. Note that this approach differs from the work we compared low-data performance to, where the test set size was proportional to the training dataset size.

CODE AND DATA AVAILABILITY

The code used in the paper is publicly available in the repository
https://github.com/zarkoivkovicc/AFE-TL-for-13C-NMR-chemical-shifts under the ASL license, including the transfer learned models' weights. Code and data related to pre-trained models can be found in the code repositories of the corresponding publications.

AUTHOR CONTRIBUTIONS

Ž.I.: conceptualization, investigation, methodology, software, visualization, writing - original draft J.J.: funding acquisition, supervision, writing - review and editing J.H.: resources, supervision, writing - review and editing.

ACKNOWLEDGMENTS

Ž.I. acknowledge the IQTC-UB Master+ grant. J.J. acknowledges Maria de Maeztu grant (code: CEX2021-001202-M).

REFERENCES

RDKit.

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa-2: Towards Chemical Foundation Models, September 2022.

Mark Ashton, John Barnard, Florence Casset, Michael Charlton, Geoffrey Downs, Dominique Gorse, John Holliday, Roger Lahana, and Peter Willett. Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions. *Quantitative Structure-Activity Relationships*, 21(6):598–604, December 2002. ISSN 0931-8771. doi: 10.1002/qsar.200290002.

Alessandro Bagno and Giacomo Saielli. Computational NMR spectroscopy: Reversing the information flow. *Theoretical Chemistry Accounts*, 117(5):603–619, May 2007. ISSN 1432-2234. doi: 10.1007/s00214-006-0196-z.

Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20, May 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0069-3.

Ilyes Batatia, Simon Batzner, Dávid Péter Kovács, Albert Musaelian, Gregor N. C. Simm, Ralf Drautz, Christoph Ortner, Boris Kozinsky, and Gábor Csányi. The Design Space of E(3)-Equivariant Atom-Centered Interatomic Potentials, 2022.

Enrico Benassi. Benchmarking of density functionals for a soft but accurate prediction and assignment of 1H and 13C NMR chemical shifts in organic and biological molecules. *Journal of Computational Chemistry*, 38(2):87–92, January 2017. ISSN 0192-8651. doi: 10.1002/jcc.24521.

T. Birsan and D. Tiba. One Hundred Years Since the Introduction of the Set Distance by Dimitrie Pompeiu. In F. Ceragioli, A. Dontchev, H. Futura, K. Marti, and L. Pandolfi (eds.), *System Modeling and Optimization*, volume 199, pp. 35–39. Kluwer Academic Publishers, Boston, 2006. ISBN 978-0-387-32774-7. doi: 10.1007/0-387-33006-2_4.

W. Bremser. Hose — a novel substructure code. *Analytica Chimica Acta*, 103(4):355–365, December 1978. ISSN 0003-2670. doi: 10.1016/S0003-2670(01)83100-7.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners, 2020.

Lukas Budach, Moritz Feuerpfeil, Nina Ihde, Andrea Nathansen, Nele Noack, Hendrik Patzlaff, Felix Naumann, and Hazar Harmouch. The Effects of Data Quality on Machine Learning Performance, 2022.

T.B. Cheshire, P.R. Ramblenm, Dean J. Tantillo, Matthew R. Siebert, and Michael W. Lodewyk. CHEmical SHift REpository with Coupling Constants Added Too. http://cheshirenmr.info/.

Paola Cimino, Luigi Gomez-Paloma, Dario Duca, Raffaele Riccio, and Giuseppe Bifulco. Comparison of different theory models and basis sets in the calculation of 13C NMR chemical shifts of natural products. *Magnetic Resonance in Chemistry*, 42(S1):S26–S33, October 2004. ISSN 0749-1581. doi: 10.1002/mrc.1410.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), 2015.

Iván Cortés, Cristina Cuadrado, Antonio Hernández Daranas, and Ariel M. Sarotti. Machine learning in computational NMR-aided structural elucidation. *Front. Nat. Prod.*, 2:1122426, January 2023. ISSN 2813-2602. doi: 10.3389/fntpr.2023.1122426.

Mikaela DiBello, Alan R. Healy, Herman Nikolayevskiy, Zhi Xu, and Seth B. Herzon. Structure Elucidation of Secondary Metabolites: Current Frontiers and Lingering Pitfalls. *Acc. Chem. Res.*, 56(12):1656–1668, June 2023. ISSN 0001-4842. doi: 10.1021/acs.accounts.3c00183.

A.M. El-Samman, S. De Castro, B. Morton, and S. De Baerdemacker. Transfer learning graph representations of molecules for pKa, 13C-NMR, and solubility. *Can. J. Chem.*, 102(4):275–288, April 2024. ISSN 0008-4042. doi: 10.1139/cjc-2023-0152.

Samuel G. Espley, Elliot H. E. Farrar, David Buttar, Simone Tomasi, and Matthew N. Grayson. Machine learning reaction barriers in low data regimes: A horizontal and diagonal transfer learning approach. *Digital Discovery*, 2(4):941–951, 2023. ISSN 2635-098X. doi: 10.1039/D3DD00085K.

William Falcon and The PyTorch Lightning team. PyTorch lightning, March 2019.

Frankie J. Fan and Yun Shi. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorganic & Medicinal Chemistry*, 72:117003, October 2022. ISSN 0968-0896. doi: 10.1016/j.bmc.2022.117003.

Abolfazl Farahani, Behrouz Pourshojae, Khaled Rasheed, and Hamid R. Arabnia. A Concise Review of Transfer Learning, 2021.

Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric, 2019.

Peng Gao, Jun Zhang, Qian Peng, Jie Zhang, and Vassiliki-Alexandra Glezakou. General Protocol for the Accurate Prediction of Molecular ^{13} C/ ^{1} H NMR Chemical Shifts via Machine Learning Augmented DFT. *J. Chem. Inf. Model.*, 60(8):3746–3754, August 2020. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.0c00388.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural Message Passing for Quantum Chemistry, 2017.

Christopher P. Gordon, Christophe Raynaud, Richard A. Andersen, Christophe Copéret, and Odile Eisenstein. Carbon-13 NMR Chemical Shift: A Descriptor for Electronic Structure and Reactivity of Organometallic Compounds. *Acc. Chem. Res.*, 52(8):2278–2289, August 2019. ISSN 0001-4842. doi: 10.1021/acs.accounts.9b00225.

Yanfei Guan, S. V. Shree Sowndarya, Liliana C. Gallegos, Peter C. St. John, and Robert S. Paton. Real-time prediction of 1H and 13C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.*, 12(36):12012–12026, 2021. ISSN 2041-6520. doi: 10.1039/D1SC03343C.

Thomas A. Halgren. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry*, 17(5-6):490–519, April 1996. ISSN 0192-8651. doi: 10.1002/(SICI)1096-987X(199604)17:5/6⟨490::AID-JCC1⟩3.0.CO;2-P.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs, 2017.

Herim Han and Sunghwan Choi. Transfer Learning from Simulation to Experimental Data: NMR Chemical Shift Predictions. *J. Phys. Chem. Lett.*, 12(14):3662–3668, April 2021. doi: 10.1021/acs.jpclett.1c00578.

Jongmin Han, Hyungu Kang, Seokho Kang, Youngchun Kwon, Dongseon Lee, and Youn-Suk Choi. Scalable graph neural network for NMR chemical shift prediction. *Phys. Chem. Chem. Phys.*, 24 (43):26870–26878, 2022. ISSN 1463-9076. doi: 10.1039/D2CP04542G.

Esther Heid, Kevin P. Greenman, Yunsie Chung, Shih-Cheng Li, David E. Graff, Florence H. Vermeire, Haoyang Wu, William H. Green, and Charles J. McGill. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.*, 64(1):9–17, January 2024. ISSN 1549-9596. doi: 10.1021/acs.jcim.3c01250.

Zhaorui Huang, Michael S. Chen, Cristian P. Woroch, Thomas E. Markland, and Matthew W. Kanan. A framework for automated structure elucidation from routine NMR spectra. *Chem. Sci.*, 12(46): 15329–15338, 2021. ISSN 2041-6520. doi: 10.1039/D1SC04105C.

Eric Jonas and Stefan Kuhn. Rapid prediction of NMR spectral properties with quantified uncertainty. *Journal of Cheminformatics*, 11(1):50, August 2019. ISSN 1758-2946. doi: 10.1186/s13321-019-0374-3.

Eric Jonas, Stefan Kuhn, and Nils Schlörer. Prediction of chemical shift in NMR: A review. *Magnetic Resonance in Chemistry*, 60(11):1021–1031, November 2022. ISSN 0749-1581. doi: 10.1002/mrc.5234.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, June 2021. ISSN 2522-5820. doi: 10.1038/s42254-021-00314-5.

Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules, 2023.

Stefan Kuhn and Nils E. Schlörer. Facilitating quality control for spectra assignments of small organic molecules: Nmrshiftdb2 – a free in-house NMR database with integrated LIMS for academic service laboratories. *Magnetic Resonance in Chemistry*, 53(8):582–589, August 2015. ISSN 0749-1581. doi: 10.1002/mrc.4263.

Stefan Kuhn, Heinz Kolshorn, Christoph Steinbeck, and Nils Schlörer. Twenty years of nmrshiftdb2: A case study of an open database for analytical chemistry. *Magnetic Resonance in Chemistry*, 62 (2):74–83, February 2024. ISSN 0749-1581. doi: 10.1002/mrc.5418.

Gaurav Kumar and Pradeep Kumar Bhatia. A Detailed Review of Feature Extraction in Image Processing Systems. In *2014 Fourth International Conference on Advanced Computing & Communication Technologies*, pp. 5–12, Rohtak, India, February 2014. IEEE. ISBN 978-1-4799-4910-6. doi: 10.1109/ACCT.2014.74.

Youngchun Kwon, Dongseon Lee, Youn-Suk Choi, Myeonginn Kang, and Seokho Kang. Neural Message Passing for NMR Chemical Shift Prediction. *J. Chem. Inf. Model.*, 60(4):2024–2030, April 2020. ISSN 1549-9596. doi: 10.1021/acs.jcim.0c00195.

Greg Landrum, Paolo Tosco, Brian Kelley, Ric, David Cosgrove, sriniker, Riccardo Vianello, gedeck, NadineSchneider, Gareth Jones, Eisuke Kawashima, Dan N, Andrew Dalke, Brian Cole, Matt Swain, Samo Turk, Aleksandr Savelev, Alain Vaucher, Maciej Wójcikowski, Ichiru Take, Vincent F. Scalfani, Daniel Probst, Kazuya Ujihara, guillaume godin, Axel Pahl, Rachel Walker, Juuso Lehtivarjo, Francois Berenger, strets123, and jasondbiggs. Rdkit/rdkit: Release_2023.09.5. Zenodo, February 2024.

Michael W. Lodewyk, Matthew R. Siebert, and Dean J. Tantillo. Computational Prediction of 1H and 13C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.*, 112(3):1839–1862, March 2012. ISSN 0009-2665. doi: 10.1021/cr200106v.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2017.

Theo D. Michels, Matthew S. Dowling, and Christopher D. Vanderwal. A Synthesis of Echinopine B. *Angewandte Chemie International Edition*, 51(30):7572–7576, July 2012. ISSN 1433-7851. doi: 10.1002/anie.201203147.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, 2019.

Erick da Silva Puls, Matheus V. Todescato, and Joel L. Carbonera. An evaluation of pre-trained models for feature extraction in image classification, 2023.

Sereina Riniker and Gregory A. Landrum. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, December 2015. ISSN 1549-9596. doi: 10.1021/acs.jcim.5b00654.

David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.*, 50(5): 742–754, May 2010. ISSN 1549-9596. doi: 10.1021/ci100050t.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-Supervised Graph Transformer on Large-Scale Molecular Data, October 2020.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties, December 2022.

Herman Rull, Markus Fischer, and Stefan Kuhn. NMR shift prediction from small data quantities. *Journal of Cheminformatics*, 15(1):114, November 2023. ISSN 1758-2946. doi: 10.1186/s13321-023-00785-x.

Scott D. Rychnovsky. Predicting NMR Spectra by Computational Methods: Structure Revision of Hexacyclinol. *Org. Lett.*, 8(13):2895–2898, June 2006. ISSN 1523-7060. doi: 10.1021/ol0611346.

Hugo A. Sánchez-Martínez, Juan A. Morán-Pinzón, Esther del Olmo Fernández, David López Eguiluz, José F. Adserias Vistué, José L. López-Pérez, and Estela Guerrero De León. Synergistic Combination of NAPROC-13 and NMR 13C DFT Calculations: A Powerful Approach for Revising the Structure of Natural Products. *J. Nat. Prod.*, 86(10):2294–2303, October 2023. ISSN 0163-3864. doi: 10.1021/acs.jnatprod.3c00437.

Ariel M. Sarotti and Silvina C. Pellegrinet. A Multi-standard Approach for GIAO 13C NMR Calculations. *J. Org. Chem.*, 74(19):7254–7260, October 2009. ISSN 0022-3263. doi: 10.1021/jo901234h.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Ulrich Sternberg, Raiker Witter, and Anne S. Ulrich. 3D Structure Elucidation Using NMR Chemical Shifts. In *Annual Reports on NMR Spectroscopy*, volume 52, pp. 53–104. Academic Press, January 2004. ISBN 0066-4103. doi: 10.1016/S0066-4103(04)52002-1.

J Stothers. *Carbon-13 NMR Spectroscopy: Organic Chemistry, A Series of Monographs, Volume 24*, volume 24. Elsevier, 2012. ISBN 0-323-14550-7.

Abdel Aziz Taha and Allan Hanbury. An Efficient Algorithm for Calculating the Exact Hausdorff Distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(11):2153–2163, November 2015. ISSN 0162-8828, 2160-9292. doi: 10.1109/TPAMI.2015.2408351.

Dean J. Tantillo. Walking in the woods with quantum chemistry – applications of quantum chemical calculations in natural products research. *Nat. Prod. Rep.*, 30(8):1079–1086, 2013. ISSN 0265-0568. doi: 10.1039/C3NP70028C.

Derek van Tilborg, Helena Brinkmann, Emanuele Criscuolo, Luke Rossen, Rıza Özçelik, and Francesca Grisoni. Deep learning for low-data drug discovery: Hurdles and opportunities. *Current Opinion in Structural Biology*, 86:102818, June 2024. ISSN 0959-440X. doi: 10.1016/j.sbi.2024.102818.

Rajeshwar P. Verma and Corwin Hansch. Use of 13C NMR Chemical Shift as QSAR/QSPR Descriptor. *Chem. Rev.*, 111(4):2865–2899, April 2011. ISSN 0009-2665. doi: 10.1021/cr100125d.

Florence H. Vermeire and William H. Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, August 2021. ISSN 1385-8947. doi: 10.1016/j.cej.2021.129307.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. Van Der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul Van Mulbregt, SciPy 1.0 Contributors, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius De Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods*, 17(3):261–272, March 2020. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0686-2.

Benjue Weng. Navigating the Landscape of Large Language Models: A Comprehensive Review and Analysis of Paradigms and Fine-Tuning Strategies, 2024.

Jake Williams and Eric Jonas. Rapid prediction of full spin systems using uncertainty-aware machine learning. *Chem. Sci.*, 14(39):10902–10913, 2023. ISSN 2041-6520. doi: 10.1039/D3SC01930F.

Anan Wu, Qing Ye, Xiaowei Zhuang, Qiwen Chen, Jinkun Zhang, Jianming Wu, and Xin Xu. Elucidating Structures of Complex Organic Compounds Using a Machine Learning Model Based on the 13C NMR Chemical Shifts. *Precision Chemistry*, 1(1):57–68, March 2023. doi: 10.1021/prechem.3c00005.

Jun Xia, Yanqiao Zhu, Yuanqi Du, and Stan Z. Li. A Systematic Survey of Chemical Pre-trained Models, 2022.

Dongyue Xin, C. Avery Sader, Om Chaudhary, Paul-James Jones, Klaus Wagner, Christofer S. Tautermann, Zheng Yang, Carl A. Busacca, Reginaldo A. Saraceno, Keith R. Fandrick, Nina C. Gonnella, Keith Horspool, Gordon Hansen, and Chris H. Senanayake. Development of a 13C NMR Chemical Shift Prediction Procedure Using B3LYP/cc-pVDZ and Empirically Derived Systematic Error Correction Terms: A Computational Small Molecule Structure Elucidation Method. *J. Org. Chem.*, 82(10):5135–5145, May 2017. ISSN 0022-3263. doi: 10.1021/acs.joc.7b00321.

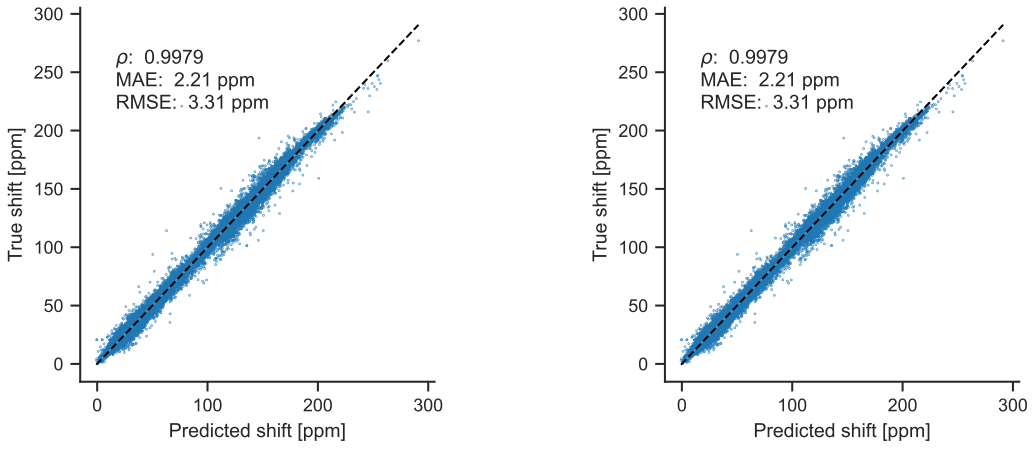
Yasemin Yesiltepe, Niranjan Govind, Thomas O. Metz, and Ryan S. Renslow. An initial investigation of accuracy required for the identification of small molecules in complex samples using quantum chemical calculated NMR chemical shifts. *Journal of Cheminformatics*, 14(1):64, September 2022. ISSN 1758-2946. doi: 10.1186/s13321-022-00587-7.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-Mol: A Universal 3D Molecular Representation Learning Framework, March 2023.

A 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 6: True against scaled DFT predicted chemical shifts of Exp5K dataset. More accurate DFT geometries don't result in better shift prediction

B ARCHITECTURES

Models architecture are shown in Figure 7 and 8. Dropout is applied after each layer.

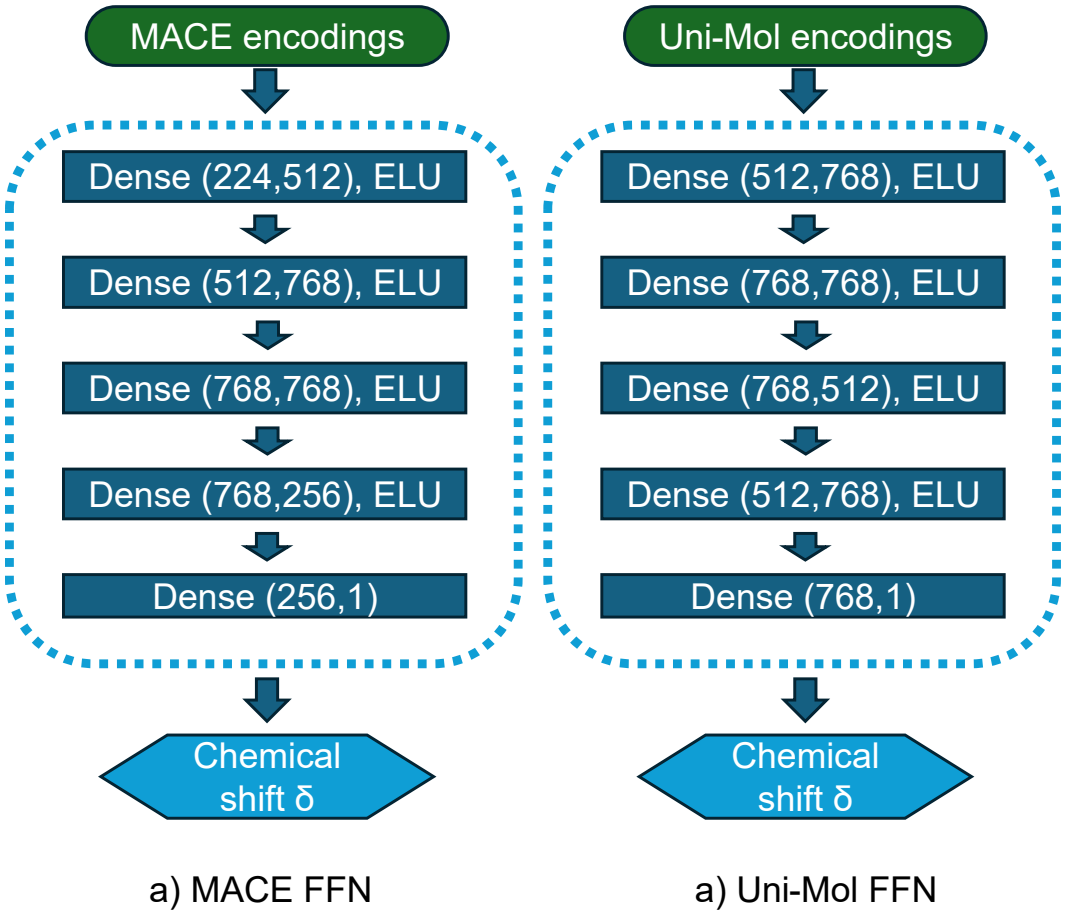


Figure 7: FFN models. Dropout layers are not shown. ELU = Exponential Linear Unit

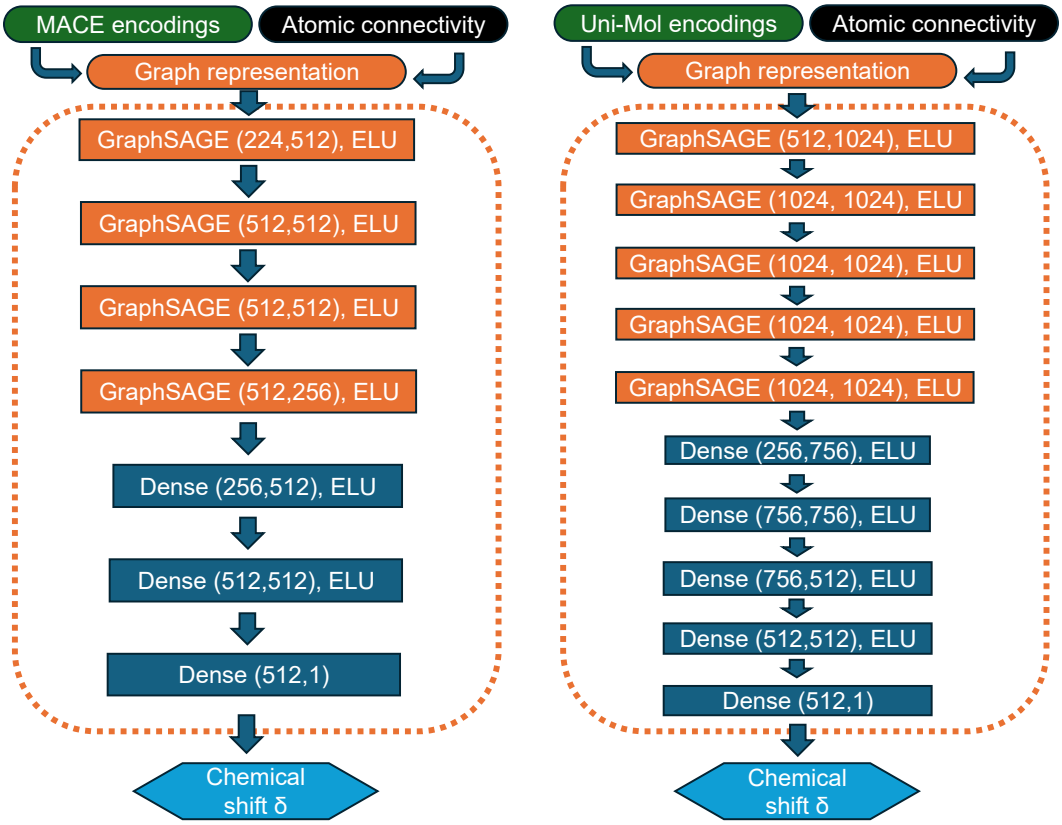


Figure 8: GNN models. Dropout layers are not shown. ELU = Exponential Linear Unit

C MODIFIED TRAIN AND TEST SETS

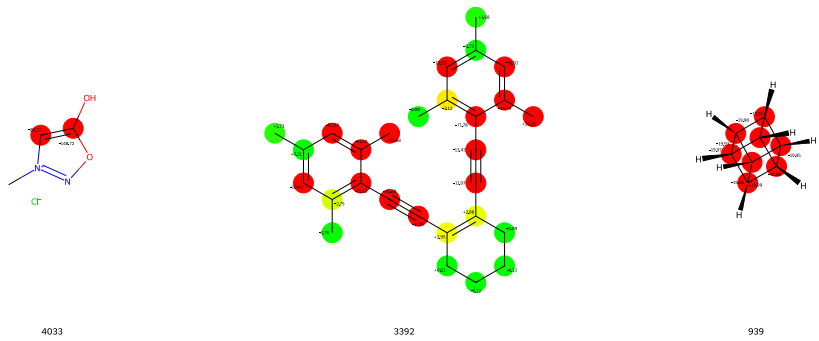


Figure 9: Molecules removed from the test set. The first molecule has the wrong graph connectivity(not shown), and the second molecule has the erroneous geometry with two methyl groups overlapping. The third molecule is removed only from the low-data regime test set due to erroneous behavior in MACE models when trained in low-data regimes, likely due to inaccurate geometry obtained from the forcefield.

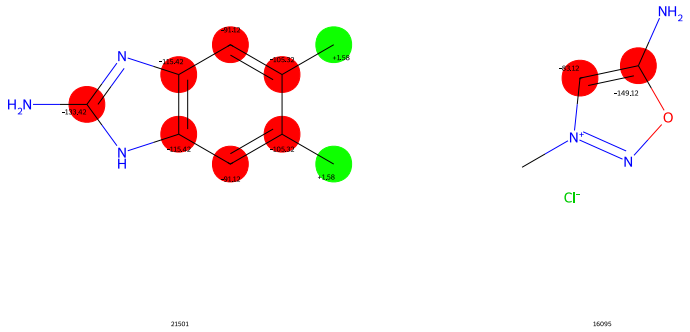


Figure 10: Molecules removed from the training set before sampling for low-data regimes. The first molecule has the wrong graph connectivity and misses one hydrogen atom in the structure(not shown), while the second molecule has the wrong graph connectivity (not shown).

D TRAINING AND HYPERPARAMETERS

The hyperparameters for all models are listed in Table 2. 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 2: 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	1000
MACE GNN	1e-3	48	0.1	1000
Uni-Mol GNN	6e-4	64	0.15	900

E DATASET SUMMARY

Table 3: 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

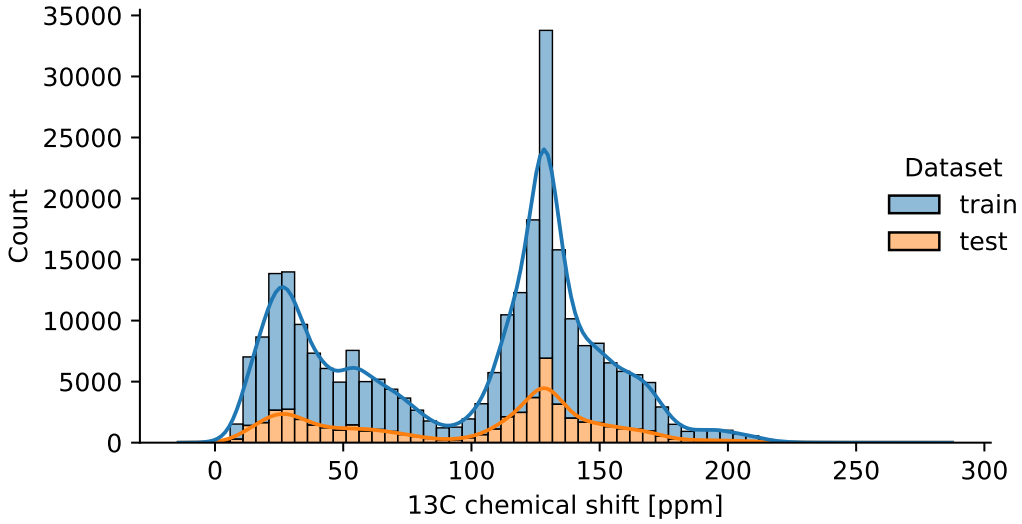


Figure 11: Distribution of chemical shifts