Predicting scalar coupling constant using dipole moment and magnetic shielding tensor values

Shri Vidhatri M M, Lokesh Paidi, Dhruvin Modi
2019113006 2019101062 2022900034

Abstract

This report explains the usage of dipole moments and magnetic tensor values in the calculation of scalar coupling constant using machine learning models and outlines a brief comparative analysis between the models that use these values and those that do not.

1 Introduction

NMR (Nuclear Magnetic Resonance) spectroscopy is used to investigate the chemical structures of an unknown molecule in solution.Scalar coupling constant (SCC) describes the interaction between two magnetic nuclei in NMR spectroscopy.

The value of SCC varies with the type of coupled atoms and the number of bonds between the coupled atoms. Thus, there is a need to accurately predict SCC values

The current project aims to do a comparative analysis between the models that utilise the dipole moment and magnetic shielding tensor values to understand the importance of these values on the accuracy of prediction of SCCs.

2 Dataset

The dataset for this is similar to the dataset that was used in the primary reference paper(Fang,2021). This dataset is from kaggle: https://www.kaggle.com/c/champs-scalar-coupling

The dataset contains molecule name, atom indices, scalar coupling constant values, dipole moments values, magnetic shielding tensor values, mulliken charge values and potential energy values.

3 Literature survey

The Fang 2021 paper mentions MPNN and transformer based architecture as some of the most striking solutions to calculate SCC. Our project compares these models with the other regression based models

The dipole moments and magnetic shielding tensor values are not just directly incorporated into the training of the architecture. They need to be not only pre-processed but also need to be included in the algorithm in specific ways. The project also outlines briefly the information regarding these ways of incorporation.

The architectures have also used a wide variety of methods such as utilising SchNet architectures, dual MPNN, XgBoost, Random Forest, Regression and many more.

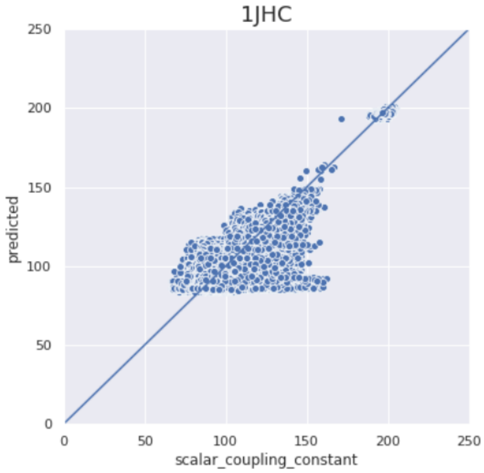
Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

4 Baseline model(s)

Multiple baseline models have been run for the comparative analysis. Primarily, there were three types of models that were run as a baseline to understand the different kinds of architectures. The first type is the models that do not utilise any chemistry information at all and purely go by black box machine learning methodologies of train and test data. The second type are the models that utilise some of the chemistry information given such as the structural data or dipole moments and utilise standard machine learning algorithms to predict the SCCs. The theird type is the models that utilise an encoder-decoder architecture with message parsing layers along with chemistry information like structures to obtain a prediction of SCCs.

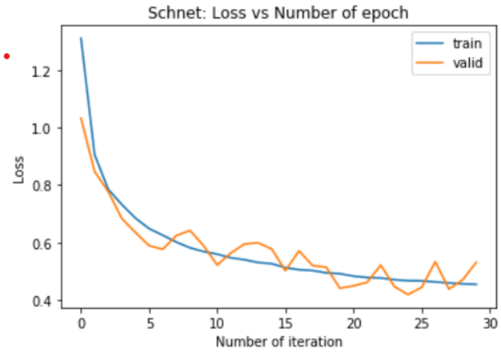
In the first type, https://www.kaggle.com/code/inversion/atomic-distance-benchmark/notebook utilises a random forest regressor model on the training data and then the model is used on the test data to obtain the predictions. The results that have been obtained by this architecture are, as expected low in accuracy due to the black box approach to the problem. The scatterplot between predicted and actual values for 1JHC is given below.

Figure 1: Predicted SCC values vs actual values



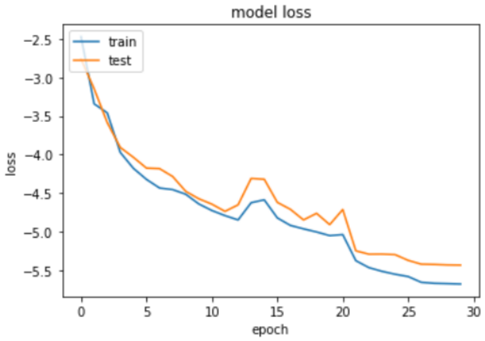
The second type of architecture includes the kind that utilises simple machine learning models by utilising some of the chemistry related data along with the test and train to get results. The example used here is a model that utilises the SchNet Architecture with train, test and structural data of the molecules to obtain the prediction of SCCs. https://www.kaggle.com/code/hau8899/schnet-1/notebook

Figure 2: Change in loss with every epoch on training and validation data



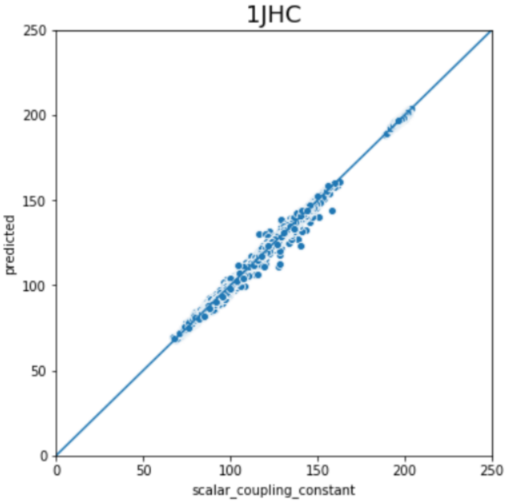
The third type of architecture is the one that utilises message-parsing neural networks. This apprach provides the most accuracy when compared to the above two types of architectures. This methodology utilises edges and nodes with message parsing layers and an Adam optimiser to predict SCC values. The change in loss with every epoch is given below.

Figure 3: Change in loss with every epoch on training and validation data



To get an understanding of how accurate these predictions are using MPNN architecture, a scatterplot of the predicted values vs actual SCC values can be plotted. This plot is for 1JGC similar to the above plot(Fig1)

Figure 4: Scatterplot of predicted vs actual values of SCC



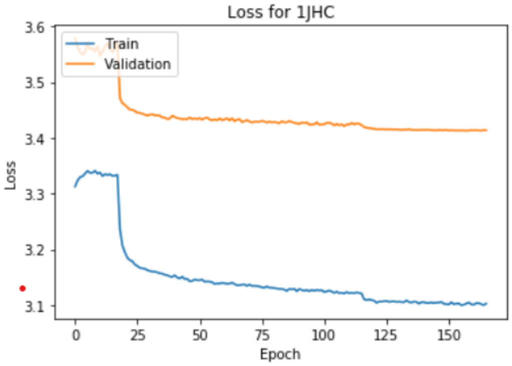
5 Comparitive analysis

Another architecture utilising one hot encoding and a bidirectional RNN on dipole moment data, structural data and potential energy data along with the test and train achieved a 97.7% accuracy on their validation data https://www.kaggle.com/code/manasghosal/brnn-using-additional-dataset.

A Keras neural network architecture was implemented with leaky ReLU and batch normalisation utilising mulliken charges, dipole moments and magnetic shielding tensors https://www.kaggle.com/code/jagannathrk/keras-neural-net-for-champs. This model was then used on the train which was split further into train and validation set for prediction. The change of loss with epochs for this architecture is given below(fig5)

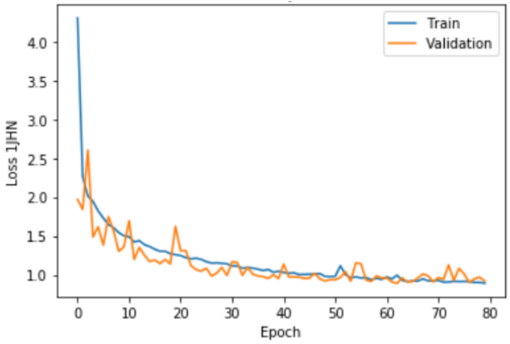
A model developed using keras neural networks with batch normalisation and Adam optimiser with multiple dense layers and hyperparameter tuning using dipole moments, Mulliken charges, magnetic

Figure 5: Loss vs Epochs



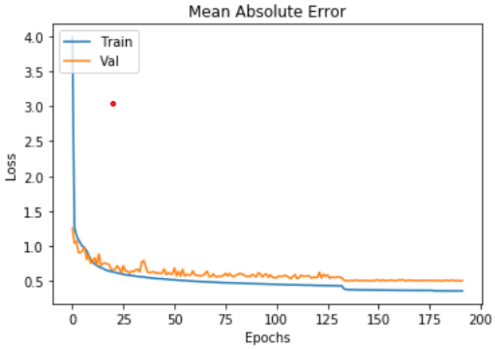
shielding tensor values, and test and train gave a -0.80 MAE loss value. The change in loss with epochs for 1JHC is given below.(fig6)

Figure 6: Loss vs Epochs



Another similar keras NN-based model achieved a similar accuracy over 200 epochs as given below.

Figure 7: MAE Loss vs Epochs



6 Conclusion

The above analysis clearly indicates that the utilisation of dipole moments and magnetic shielding tensor values in the machine learning model yeilds good results. It was also observed that regardless of the exact model used, either a keras NN or MPNN architecture or bidirectional RNN, the accuracy of these models is significantly higher than those models that are more algorithmically efficient but do not utilise these additional dipole moment and magnetic shielding tensor data. When it comes to the models within those that utilise these dipole moments and tensor values, the MPNN architecture offers the most accurate predictions with the predicted values being very close to the ground truth.

References

[1] Fang, J., Hu, L., Dong, J. et al. Predicting scalar coupling constants by graph angle-attention neural network. Scientific Reports 11, 18686 (2021). https://doi.org/10.1038/s41598-021-97146-1

[2] Guan, Y., Sowndarya, S. S., Gallegos, L. C., John, P. C. S., & Paton, R. S. (2021). Real-time prediction of 1 H and 13 C chemical shifts with DFT accuracy using a 3D graph neural network. Chemical Science, 12(36), 12012-12026.

[3] Barfield, M., Dingley, A. J., Feigon, J., & Grzesiek, S. (2001). A DFT Study of the Interresidue Dependencies of ScalarJ-Coupling and Magnetic Shielding in the Hydrogen-Bonding Regions of a DNA Triplex. Journal of the American Chemical Society, 123(17), 4014–4022. https://doi.org/10.1021/ja003781c

[4] Jian, Caiqing & Cheng, Xinyu & Zhang, Jian & Wang, Lihui. (2020). Scalar Coupling Constant Prediction Using Graph Embedding Local Attention Encoder.