Predict the scalar coupling using GAANN along with atom charges

...

By
Shri Vidhatri 2019113006
Dhruvin M. 2022900034
Lokesh P. 2019101062

Research gap

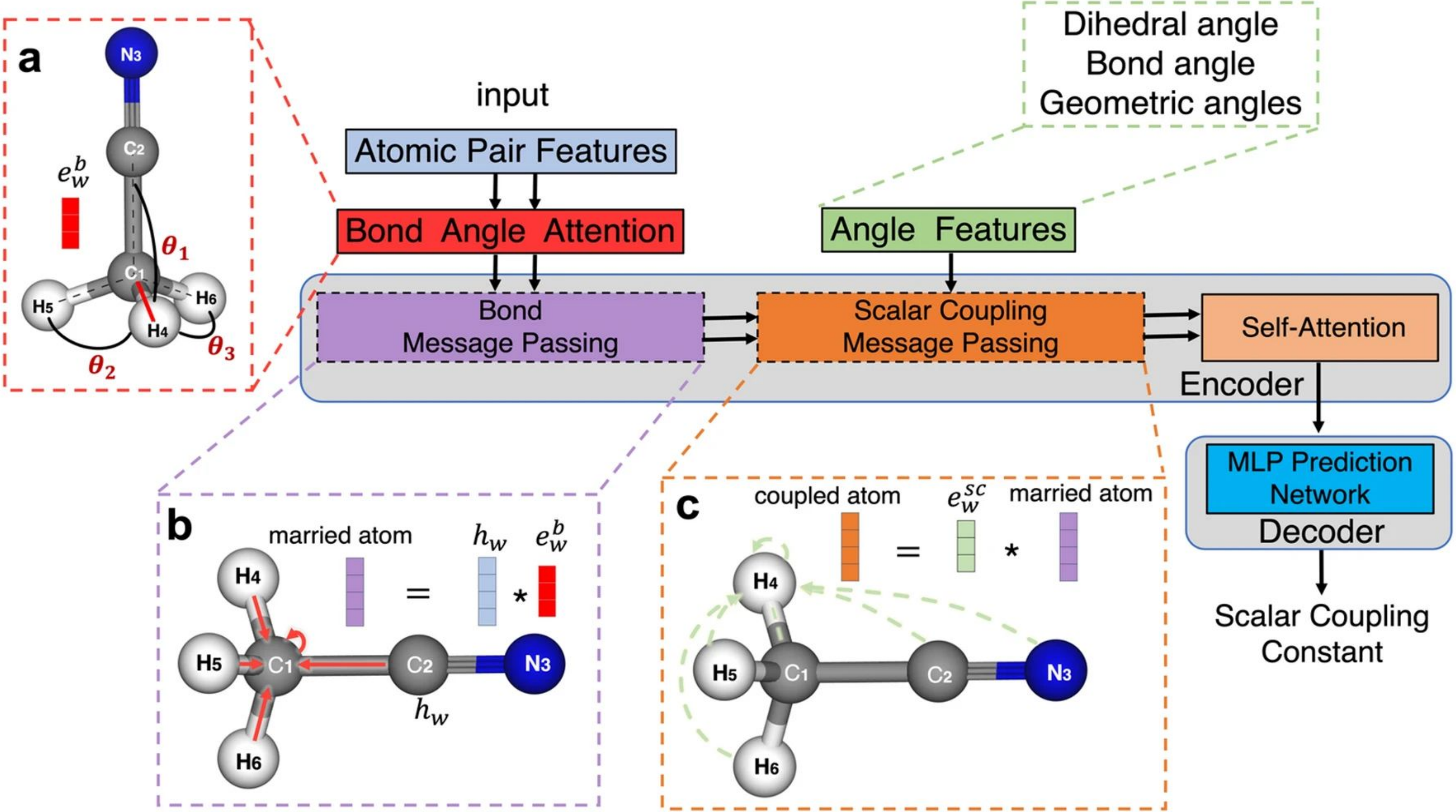
- NMR (Nuclear Magnetic Resonance) spectroscopy is used to investigate the chemical structures of an unknown molecule in solution.
- Scalar coupling constant (SCC) describes the interaction between two magnetic nuclei in NMR spectroscopy.
- The value of SCC varies with the type of coupled atoms and the number of bonds between the coupled atoms.
- Thus, there is a need to accurately predict SCC values

Statement

- As an extension from the reference paper(Fang,2021), the current project aims to utilise mulliken charge values and magnetic shielding tensor values in a GAANN model with a hypothesis that it will lead to a more accurate and robust prediction of SCC values.

Model

- Similar to the reference paper, a graph angle-attention neural network(GAANN) will be used in the project to get the SCC values
- The GAANN has a decoder and an encoder part
- The encoder includes two-layer message passing neural networks and self-attention neural network, while the decoder contains an MLP prediction network
- Atomic pair features and Bond angle attention are used as inputs for the bond message parsing whereas angle features are used in the scalar coupling message parsing layer.


a
N3
C2
e_w^b
θ_1
C1
H5
H6
H4
θ_2
θ_3
input
Atomic Pair Features
Bond Angle Attention
Dihedral angle
Bond angle
Geometric angles
Angle Features
Bond
Message Passing
Scalar Coupling
Message Passing
Self-Attention
Encoder
MLP Prediction
Network
Decoder
Scalar Coupling
Constant
b
married atom
h_w
e_w^b
=
*
H4
H5
C1
C2
N3
H6
h_w
c
coupled atom
e_w^{sc}
married atom
=
*
H4
H5
C1
C2
N3
H6

Model cont.

- The project aims to include mulliken charge values and magnetic shielding tensors in the encoder part of the GAANN.
- This can be done in two ways:
  - There can be a new message parsing layer in the encoder for these charge/magnetic values which will take the respective inputs and process them as features for the GAANN.
  - The mulliken charge values and magnetic shielding tensor values can be used as inputs in the already existing message parsing layers to incorporate them in the existing encoder architecture.
- Both of the above mentioned methods needs to be tested and analysed for accuracy of the SCC values and efficiency in terms of computational cost, which is what the project aims to do.

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.
- If possible/needed during the course of the project, it has been planned to add additional data into the dataset according to the requirements using DFT calculations and formulas from reference papers.

References

- 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
- 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.
- 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. doi:10.1021/ja003781c

Understanding the pre-existing model to make improvements

- The paper mentions MPNN and transformer based architecture as some of the most striking solutions to calculate SCC. Our project will compare these architectures also after the improvements are made
- Due to the way bond angles were incorporated into the MP layers of the encoder, the mulliken/magnetic tensor values cannot be added directly to the encoder
- The model needs to be implemented in a modular way and the data pre processing code needs to be slightly modified since the importing of data was given ambiguously
- There needs to be a correlation/similarity matrix done for the above mentioned properties (similar to the way it was done in the paper) which will be used for designing a new message parsing layer of the encoder based on these similarity values

Dataset understanding

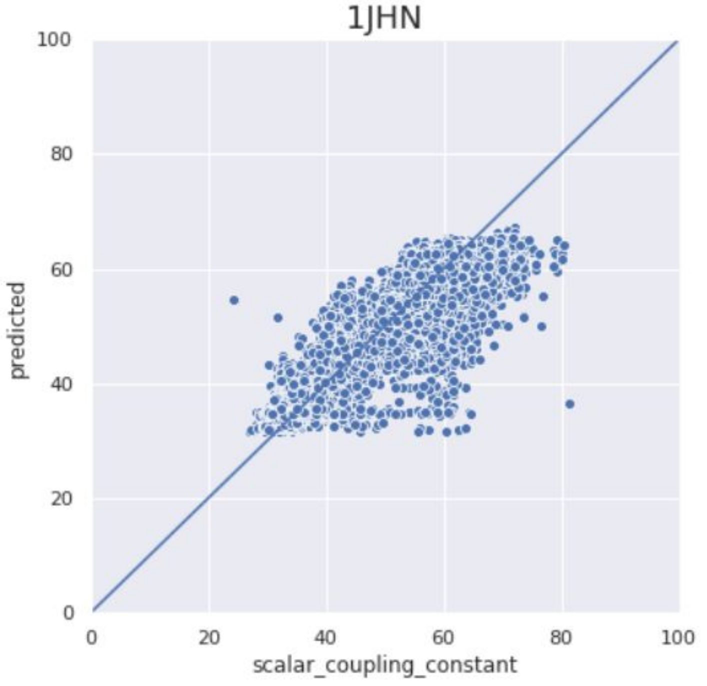
- The dataset contains 130789 molecules with their structural and other relevant information
- The part of the dataset containing magnetic shielding tensors and mulliken charges needs to be preprocessed to be used in the model
- The dataset also contains SCC’s for different types of bonds. It has been planned to understand the similarity/ correlation between magnetic tensor values and mulliken charges with these bond types before continuing with the modification of the encoder architecture

Comparison

- There have already been state of the art codes given for the dataset on kaggle. Once the proposed GAANN model is executed as mentioned earlier, the project aims to compare the accuracy and efficiency of this model with some of the models mentioned in kaggle.
- The model will also be compared against some benchmark methods as well as standard SCC calculation software efficiency for a detailed analysis.
- It is hypothesised that the efficacy and architecture of GAANN along with the proposed additions will be comparable and/or better than the existing models

Baseline model

- The baseline model was run successfully.
- The dataset had test and train separated, but the test set did not have the scalar coupling constant values, thus being not helpful in getting the error metrics
- Thus, the training set was split into train and validation set dynamically and tested for cross validation and rmse error metrics
- The results were obtained with the cross validation score -1.4973 and have been plotted separately for each SCC type.



Improvements

- Mulliken values and magnetic shielding tensor values will be added to the model.
- The model currently uses structure information along with scalar coupling constant contribution data.
- The mulliken values and magnetic shielding tensor values are currently being preprocessed before they can be added into the encoder of the model.
- Once the mulliken values and shielding tensor values will be added into the model, a comparative analysis will be done between the new model and the other architectures which do not use these two properties (models other than baseline)
- The comparative qualitative analysis will be done with architectures that include these two properties based on the way they are incorporated into the model