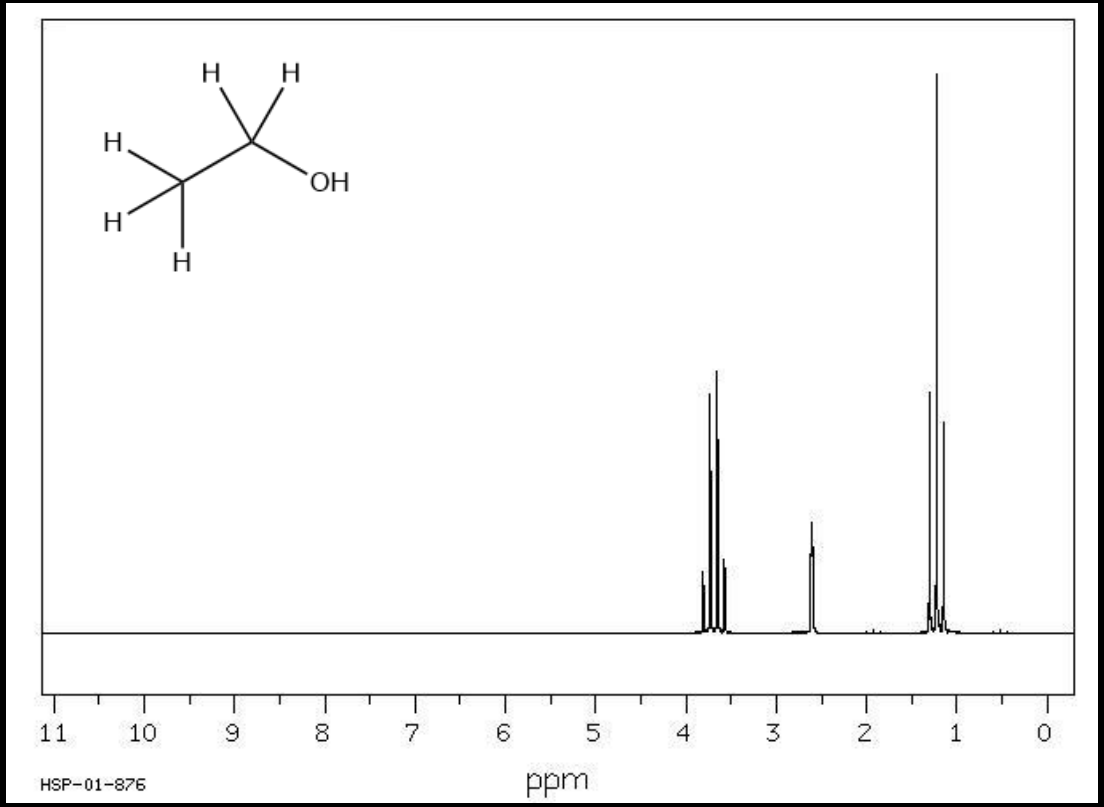
Predicting ^{1}H NMR Chemical Shifts With Graph Neural Networks

By: Aidan McCrillis and Daniel Roche

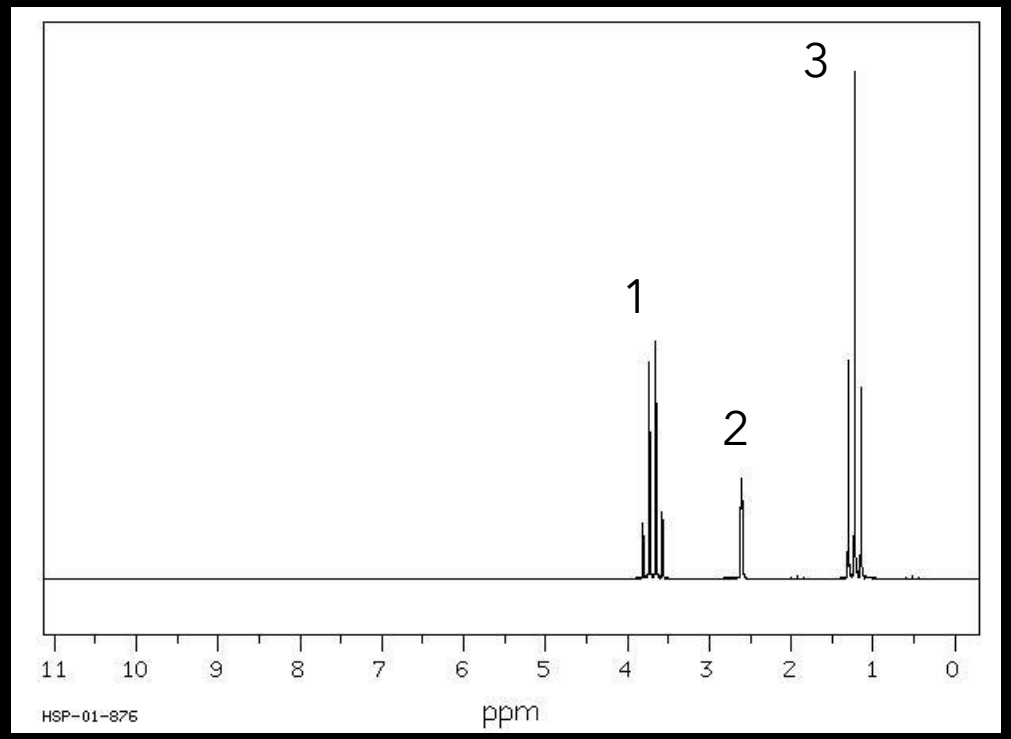
NMR Spectroscopy

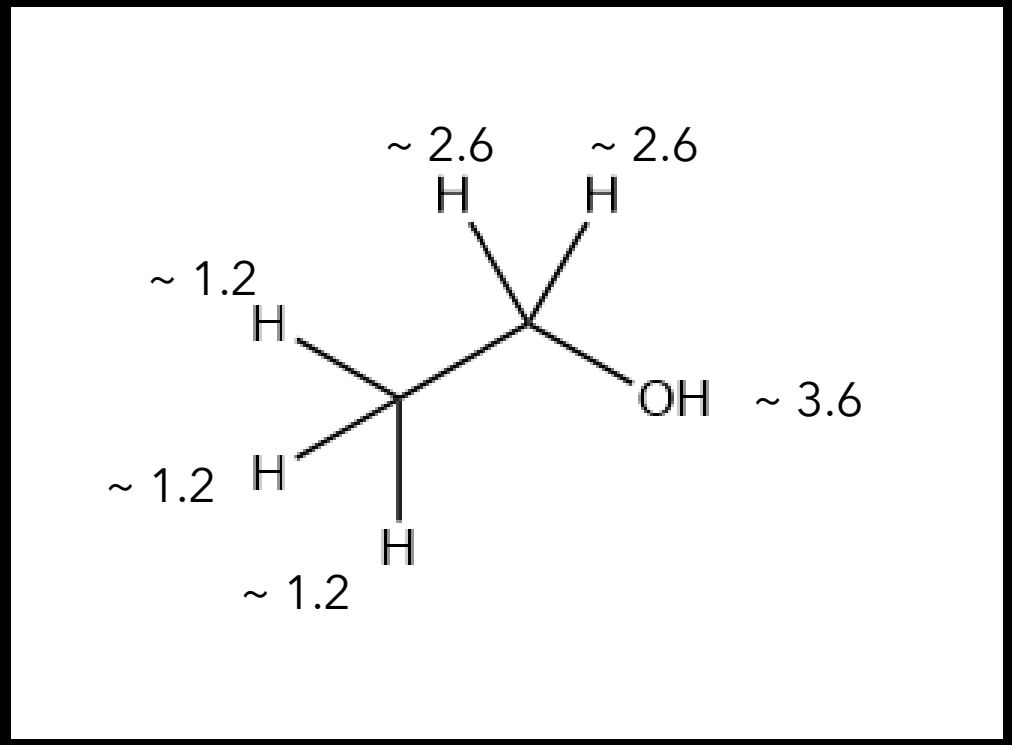
- Measure the nuclear spin of isotopes for a particular atom
- Good indicator of chemical environment
- This makes it good for identifying molecules

Functional Group	Chemical Shift Range
Alkyl (methyl-CH_3)	~ 1 ppm
Oxygen Adjacent (O-CH_2)	~ 3-4 ppm
Alkene (=CH_2)	~ 6 ppm
Alkyne (≡CH)	~ 3 ppm



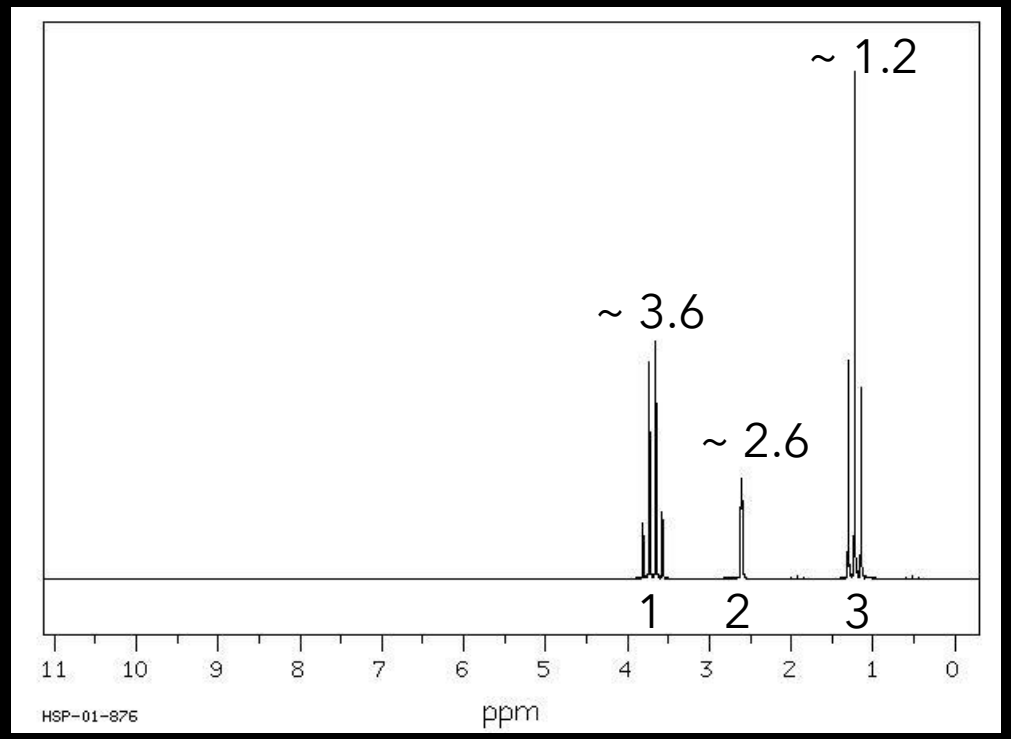
NMR Spectroscopy

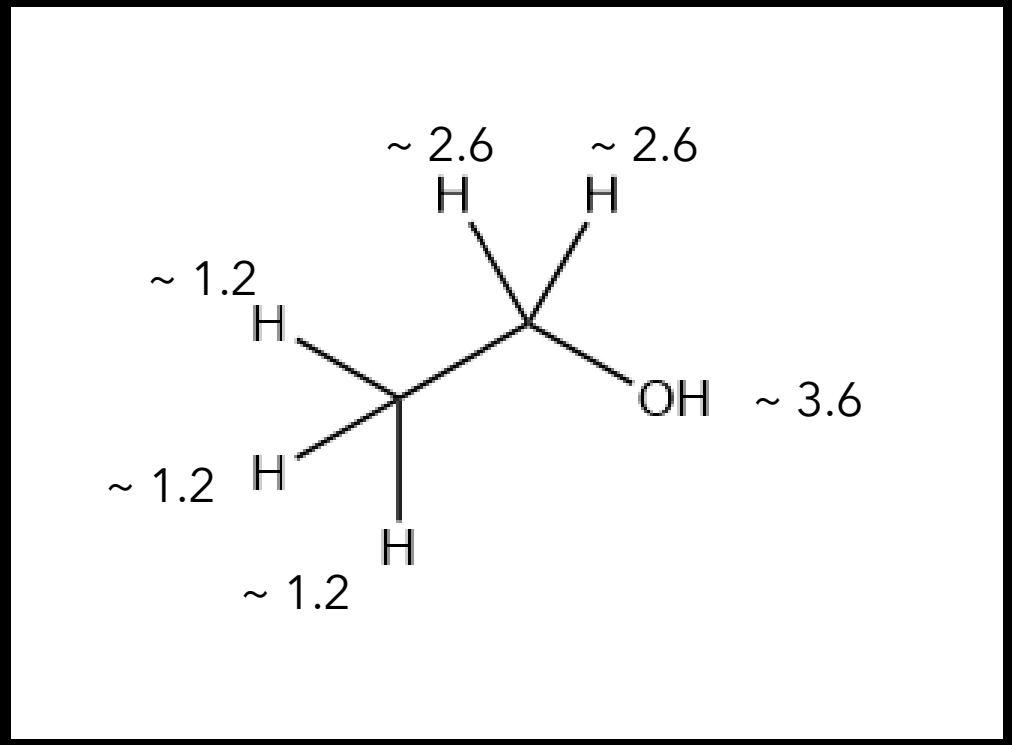




Functional Group	Chemical Shift Range
Alkyl (methyl-CH_3)	~ 1 ppm
Oxygen Adjacent (O-CH_2)	~ 3-4 ppm
Alkene (=CH_2)	~ 6 ppm
Alyne (≡CH)	~ 3 ppm

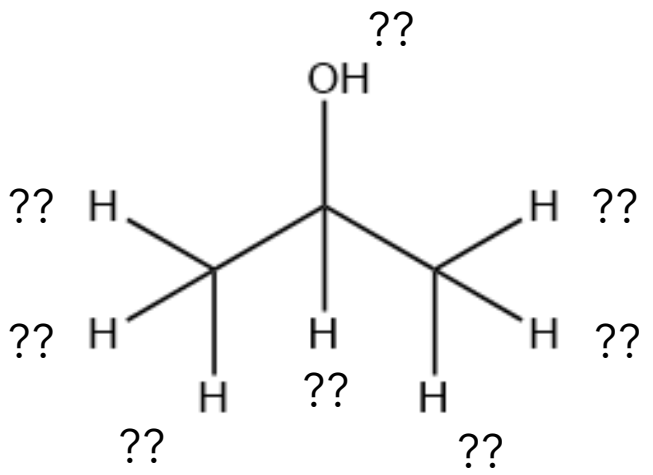
NMR Spectroscopy



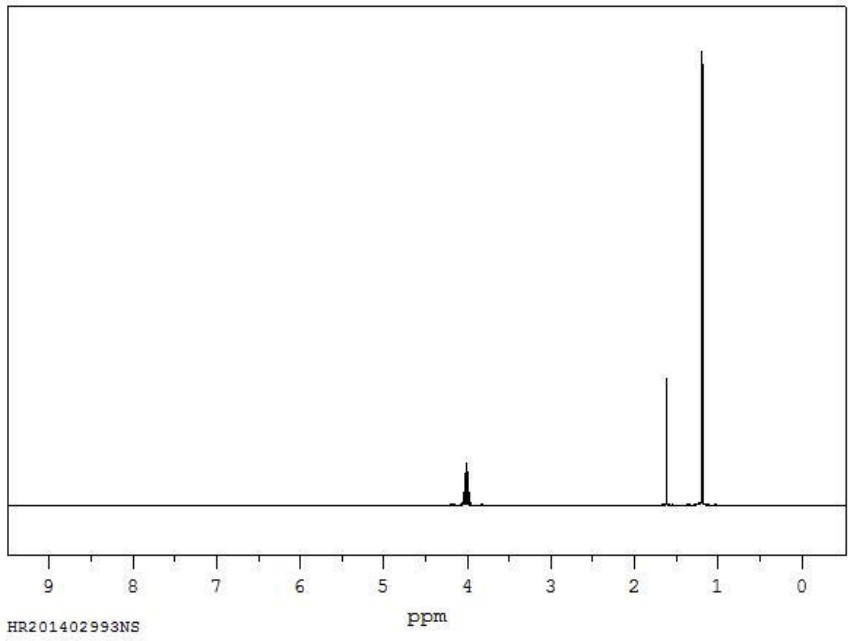


Functional Group	Chemical Shift Range
Alkyl (methyl-CH_3)	~ 1 ppm
Oxygen Adjacent (O-CH_2)	~ 3-4 ppm
Alkene (=CH_2)	~ 6 ppm
Alyne (≡CH)	~ 3 ppm

Problem Definition

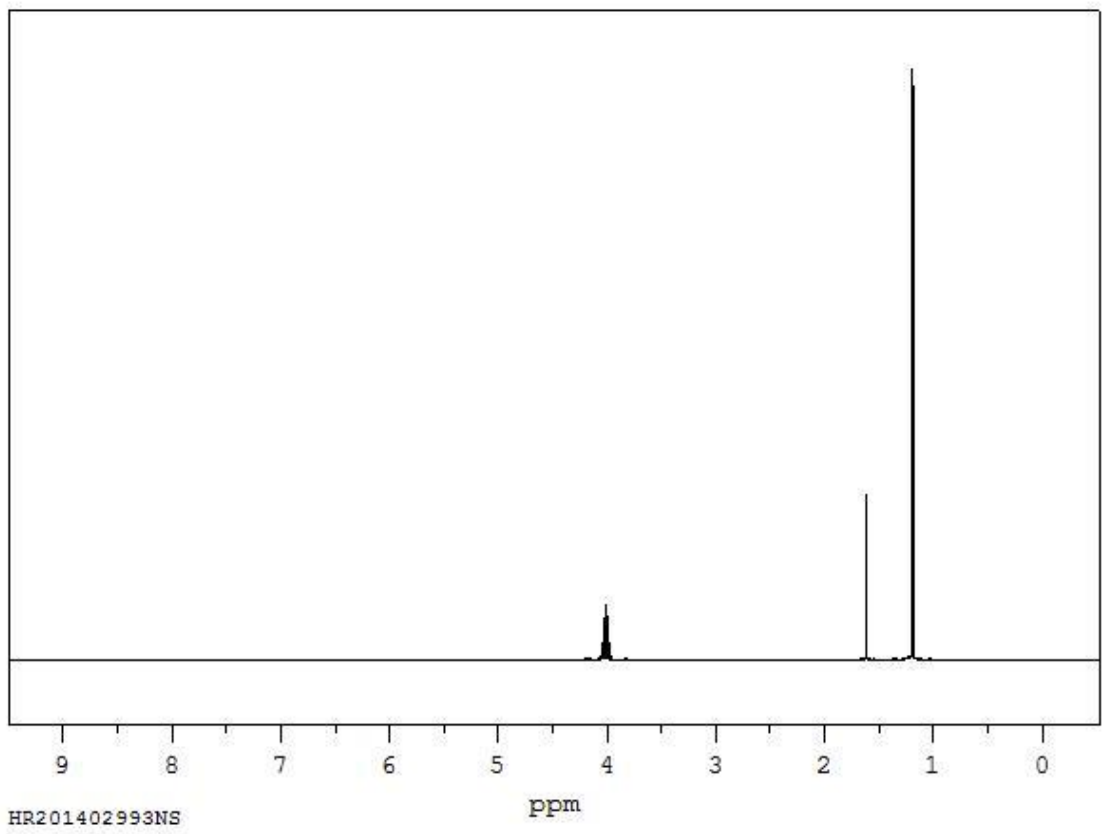


Do these match?



Functional Group	Chemical Shift Range
Alkyl (methyl-CH_3)	~ 1 ppm
Oxygen Adjacent (O-CH_2)	~ 3-4 ppm
Alkene (=CH_2)	~ 6 ppm
Alyne (≡CH)	~ 3 ppm

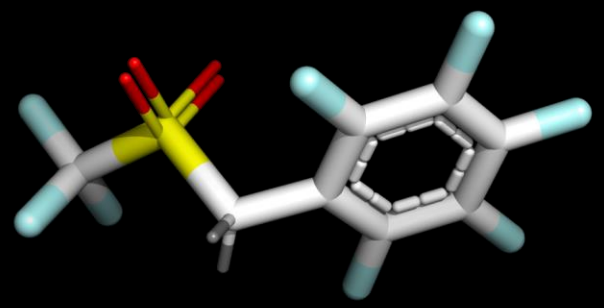
Data Extraction

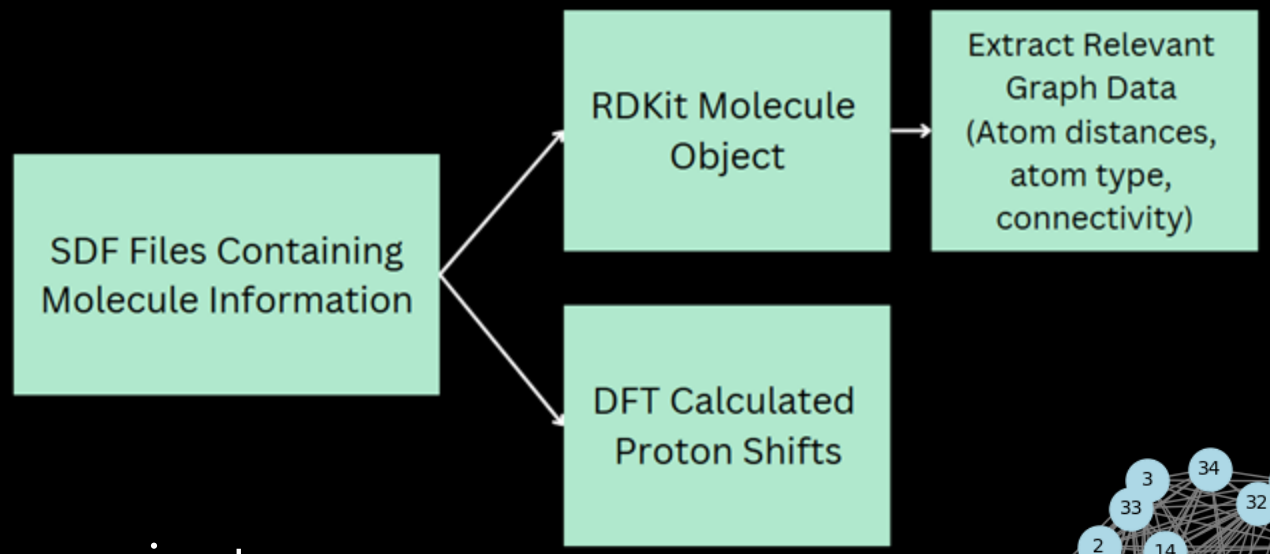


Get Peak Values →

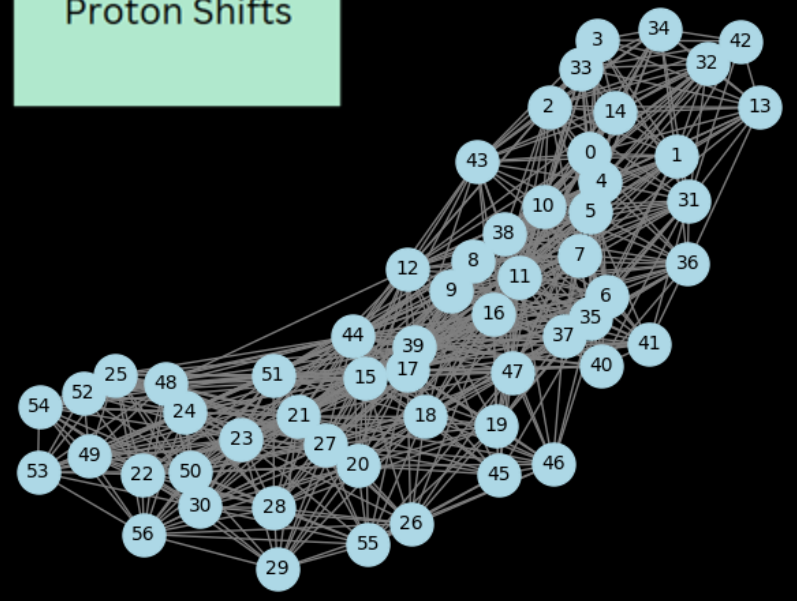
[1.2,
1.2,
1.2,
1.2,
1.2,
1.2,
1.6,
4.0]

Preprocessing Molecules

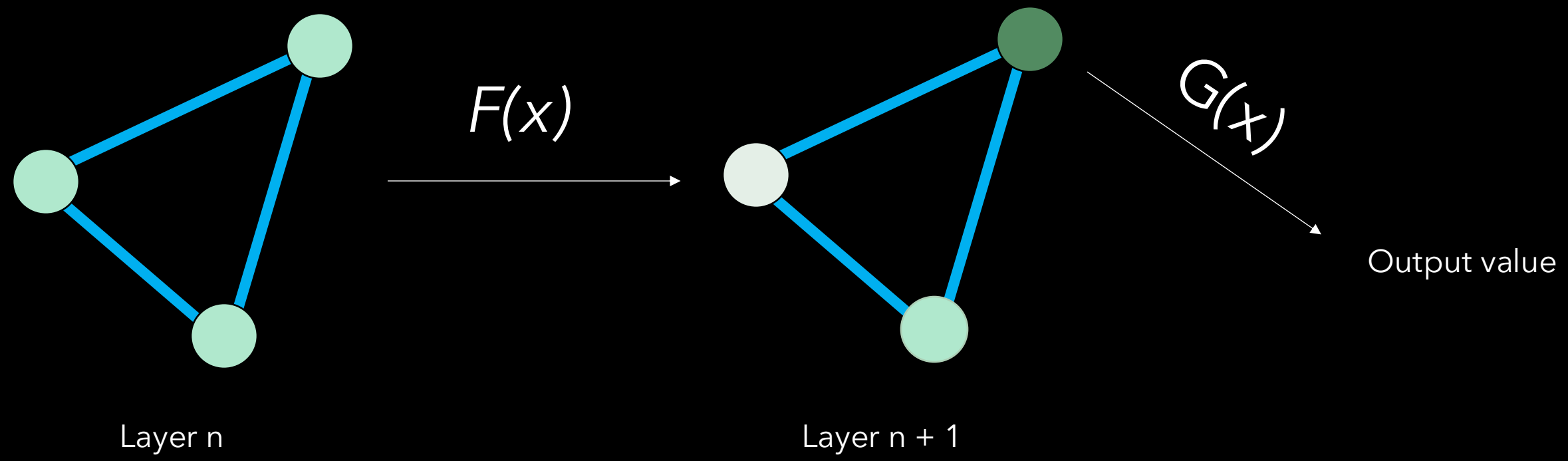




- The dataset that was used for this project contains 7,449 molecules
- Each of these molecules has an average of ~16 protons
- Atom distances under 10 Å treated as bonds (distances are indicative of bond types)

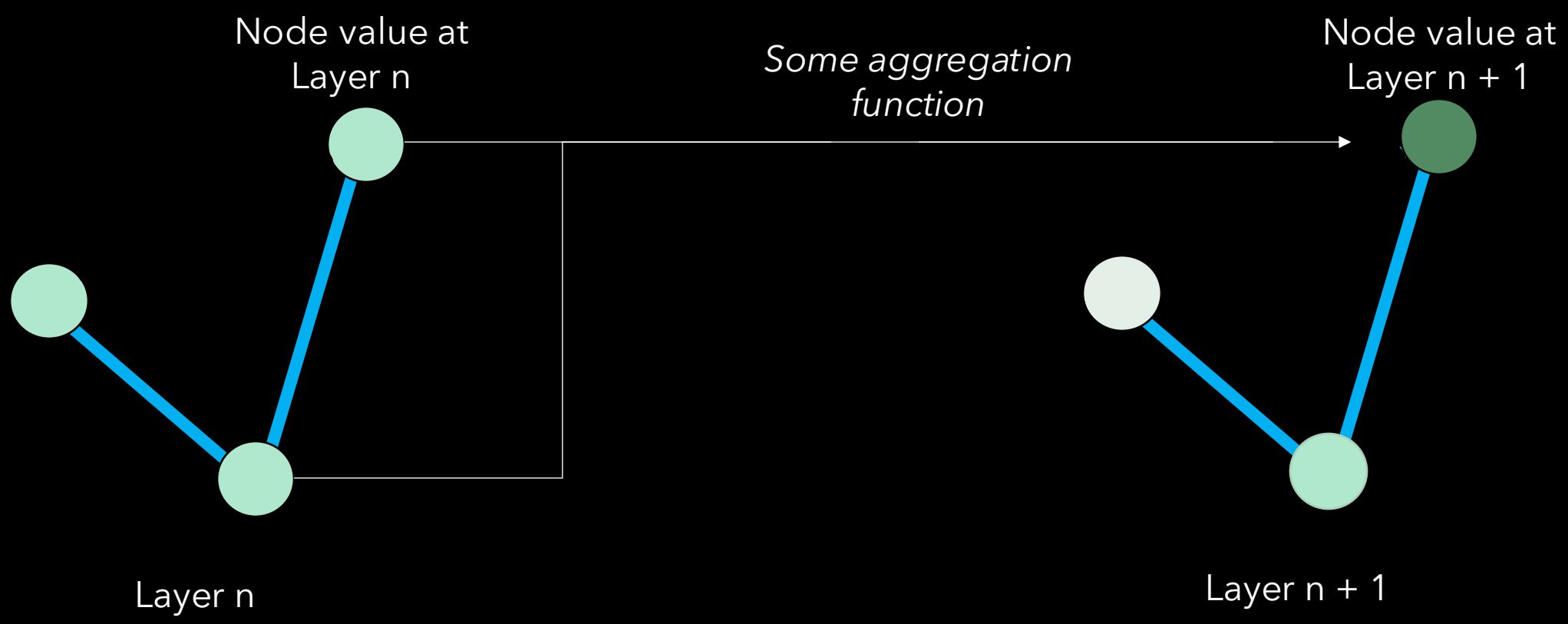


Graph Neural Networks (GNNs)



- Graph neural networks keep data in graph allowing for nodes, and edges to have their own attributes and keep data in position relative to each other

Message Passing



- Message passing allows nodes or edges to receive “messages” that are functions of neighboring nodes or edges and aggregating
- Allows for local nodes to affect each other, and multiple message passing layers can be used for node influence over longer distances

Model Architecture

- Model architecture on CASCADE model
- Consists of embedding atoms and edges passing them through a message layer and updating them
- Then fitting a final NN on the proton atom embeddings
- Three message layers used in this model

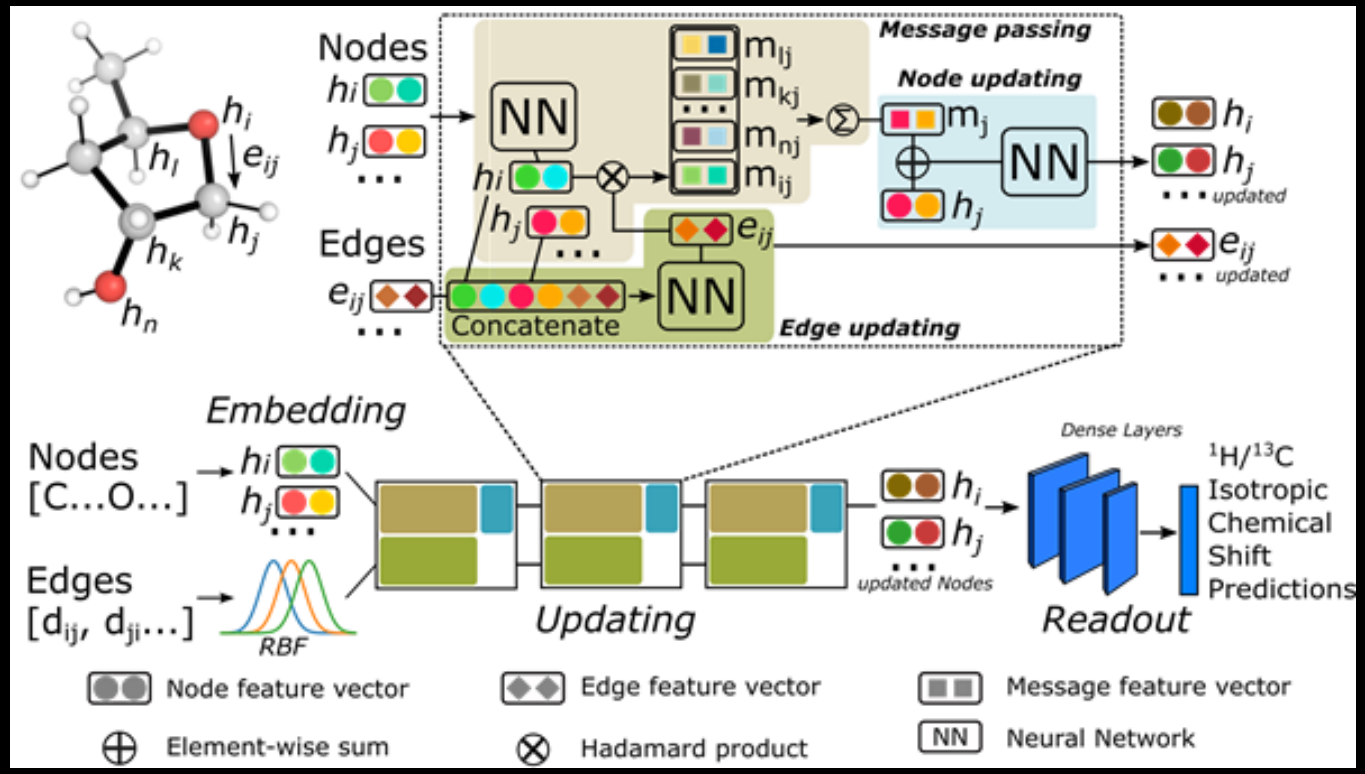
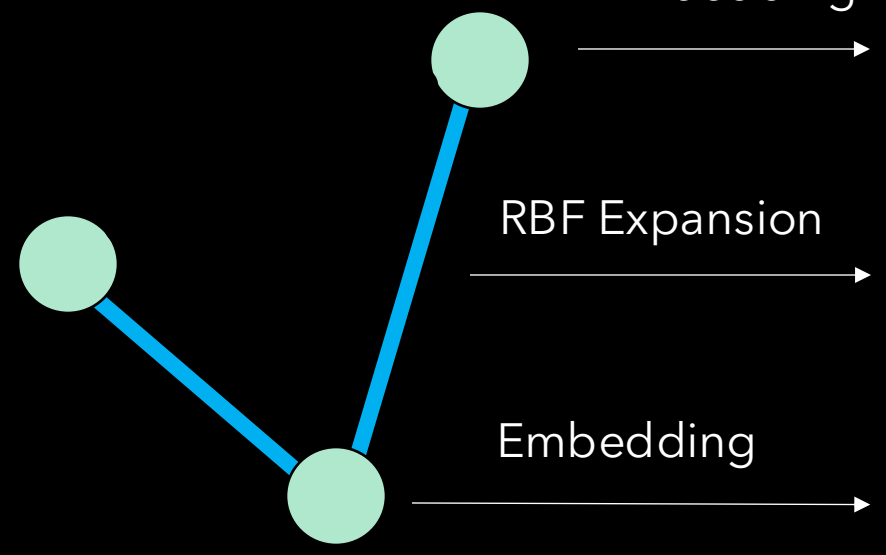


Diagram showing the message layer of the CASCADE model

Model Architecture

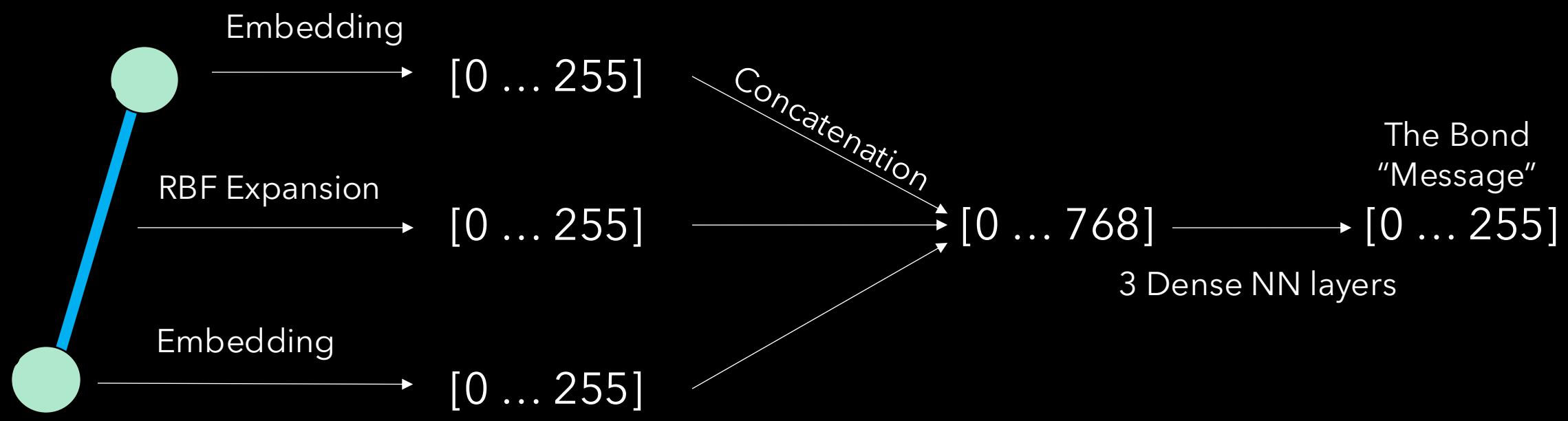


$$\widehat{e_{ij}^{0}} = \left[e^{\frac{-\left(d_{ij} - (\mu + \delta k)\right)^2}{\delta}} \right]_{k \in [0,1,2,\ldots 256]}$$

Definition of the radial basis function where d_{ij} is the distance between nodes

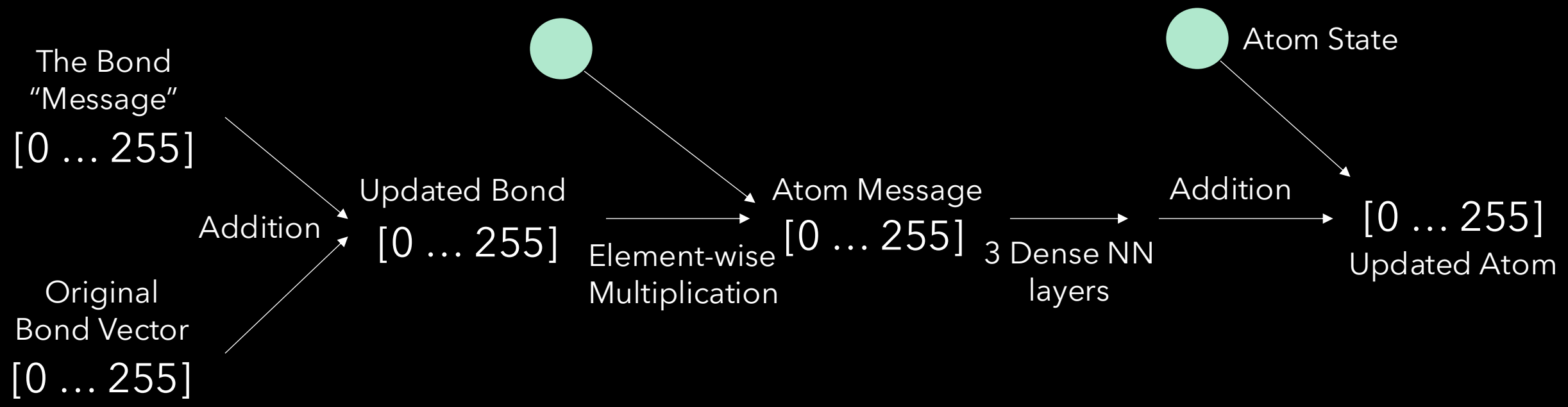
- All nodes embedded into 256 length vector (learnable) based on the atom type
- All edges are embedded into 256 length vector using a radial basis function (RBF)

Message Passing Layers



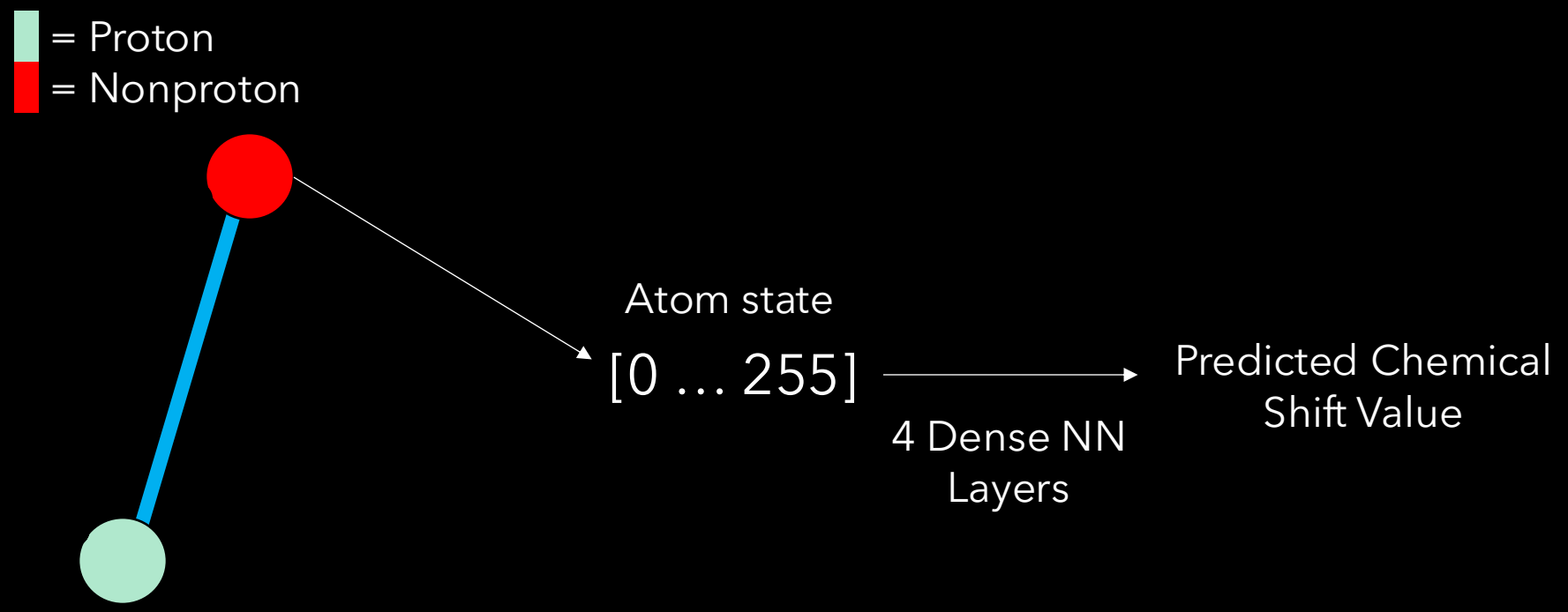
- After embedding, vectors for the atom receiving the message, the atom that's the source of the message and the edge vector are concatenated
- Layers are fitted to receive a bond "message"

Message Passing Layers



- The bond message is used to update the bond
- Bond message is multiplied "sender" atom to get a atom message
- Dense layers are used to get a final message and atom state is updated with addition

Final Predictions



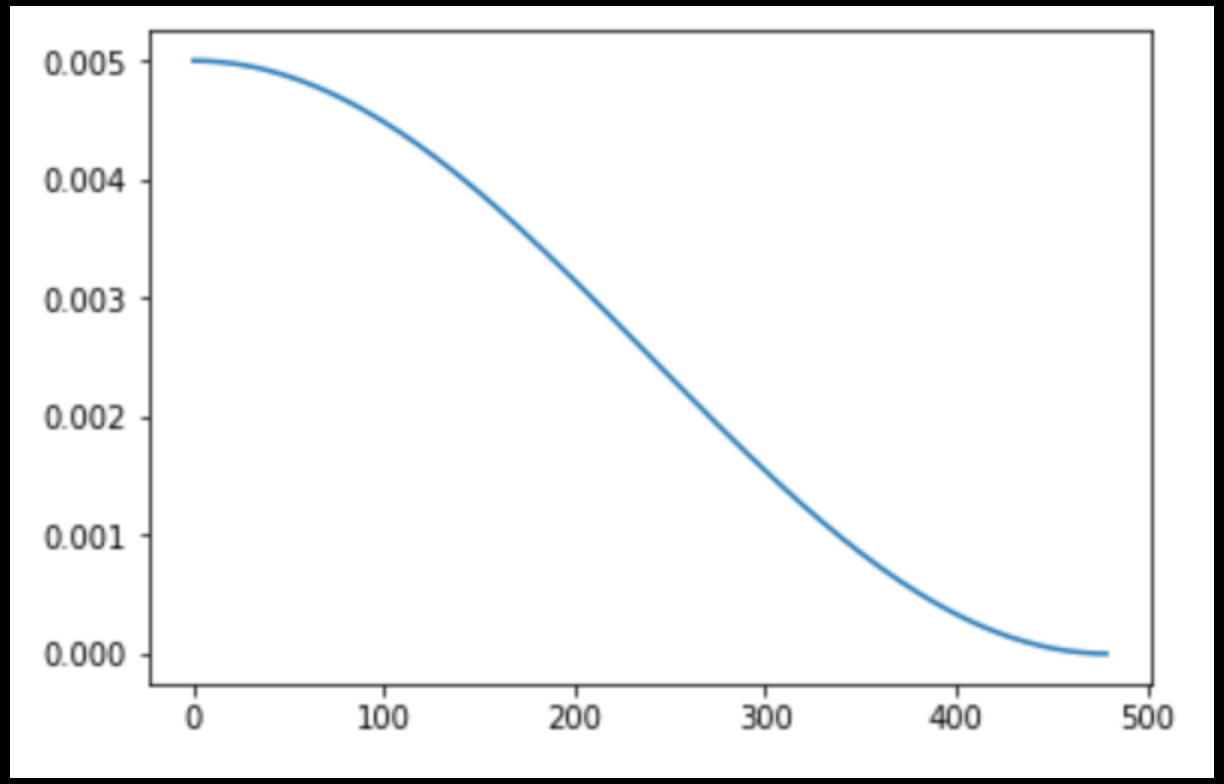
- Proton atom states are obtained and a final chemical shift prediction is obtained after using 4 dense layers

Model Details

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Mean Squared Error Loss

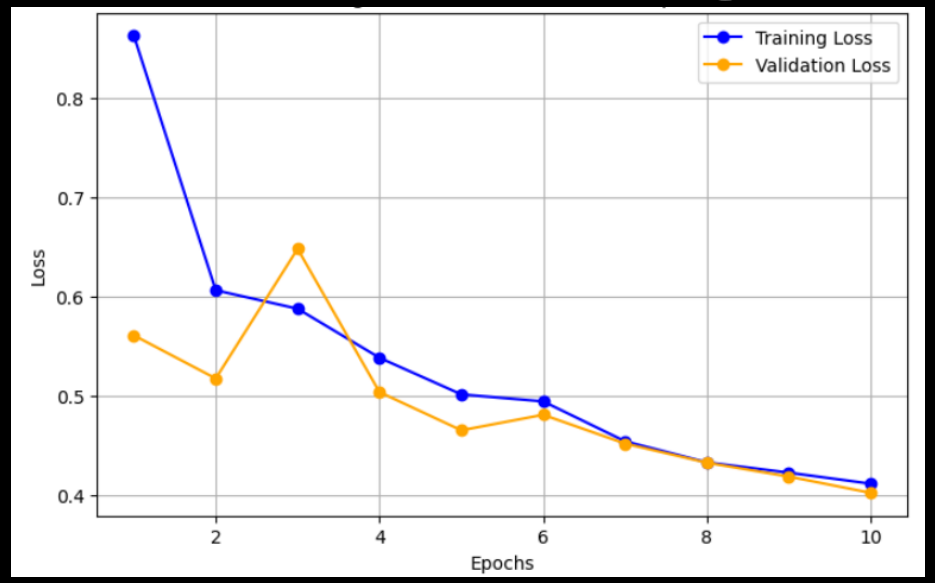
- Criterion – MSE Loss
- Optimizer - Adam
- Learning Rate Scheduler – Cosine Annealing learning rate
- Early Stopping – If there is no improvement over n epochs, then stop



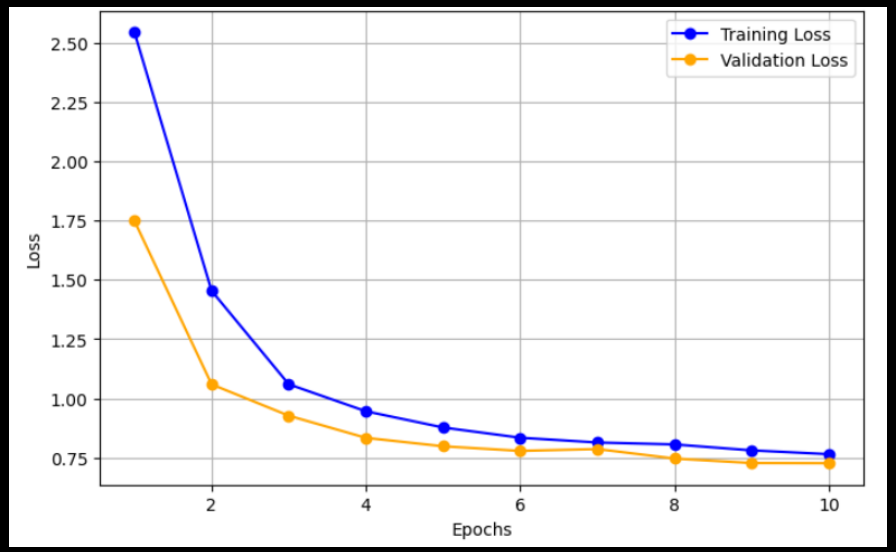
Cosine Annealing function

Hyperparameter Searching

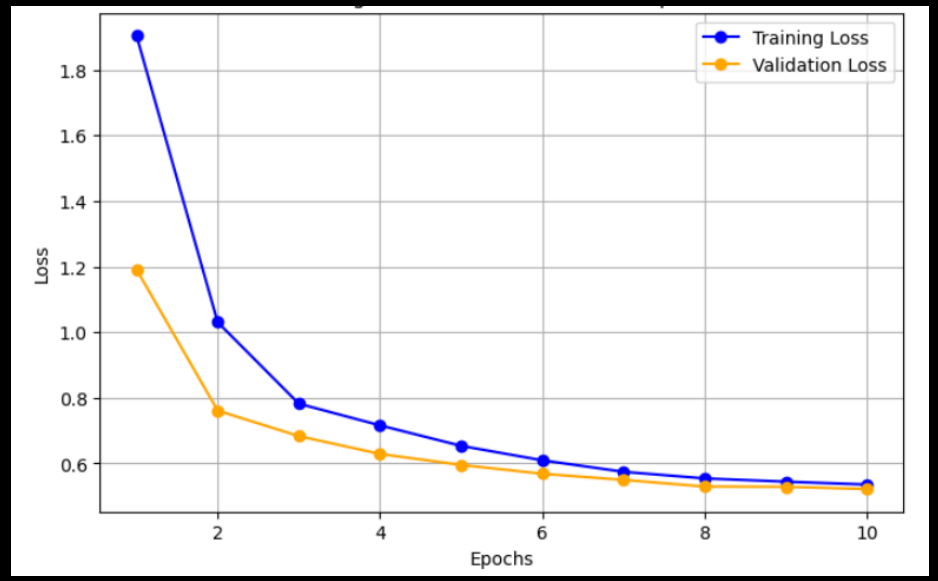
- Hyperparameters searched:
  - Epochs
  - Learning Rate
  - Message Layers
  - Batch size



Batch 32



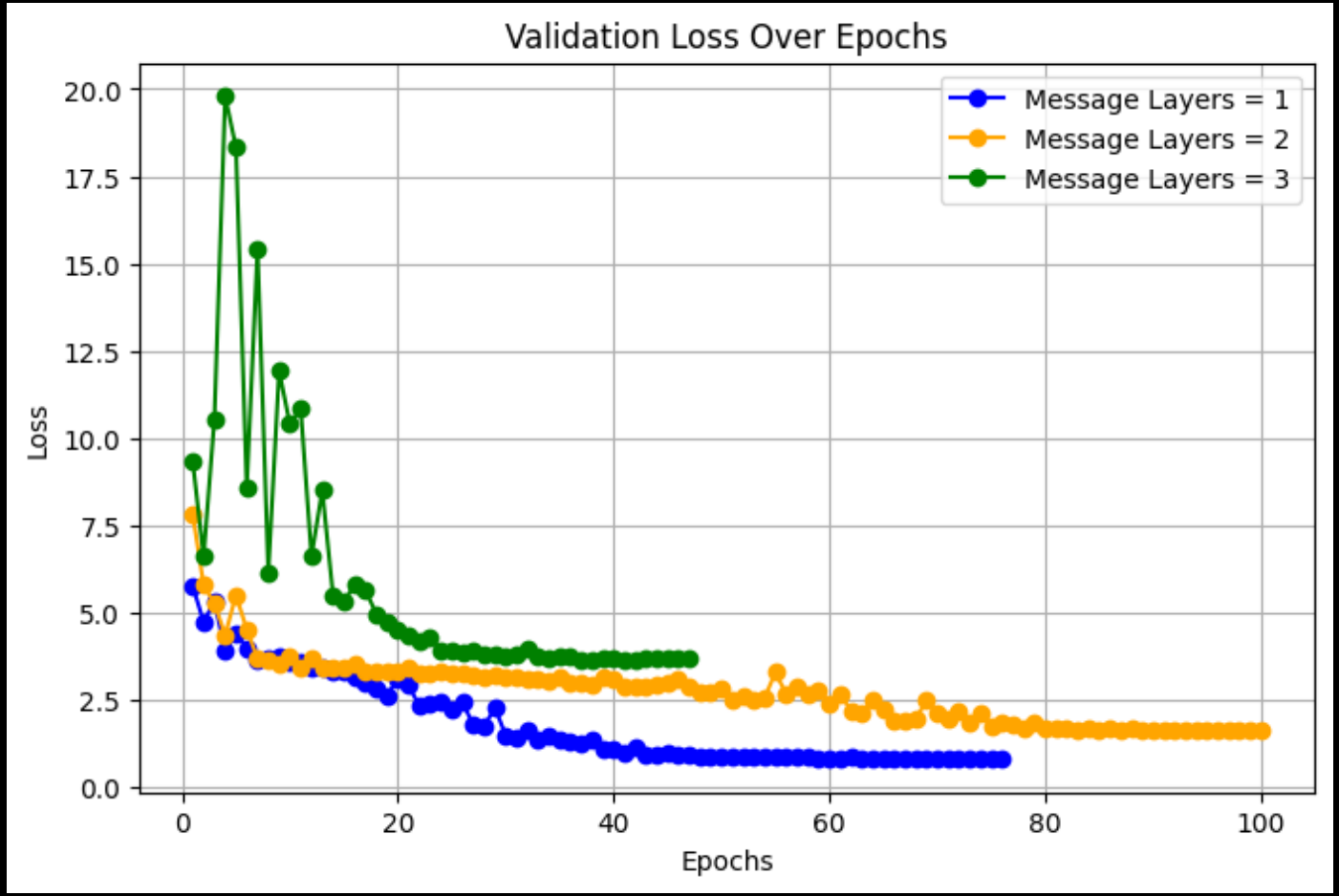
Batch 128



Batch 64

Hyperparameter Searching

- Results of testing different amount of message layers in our model



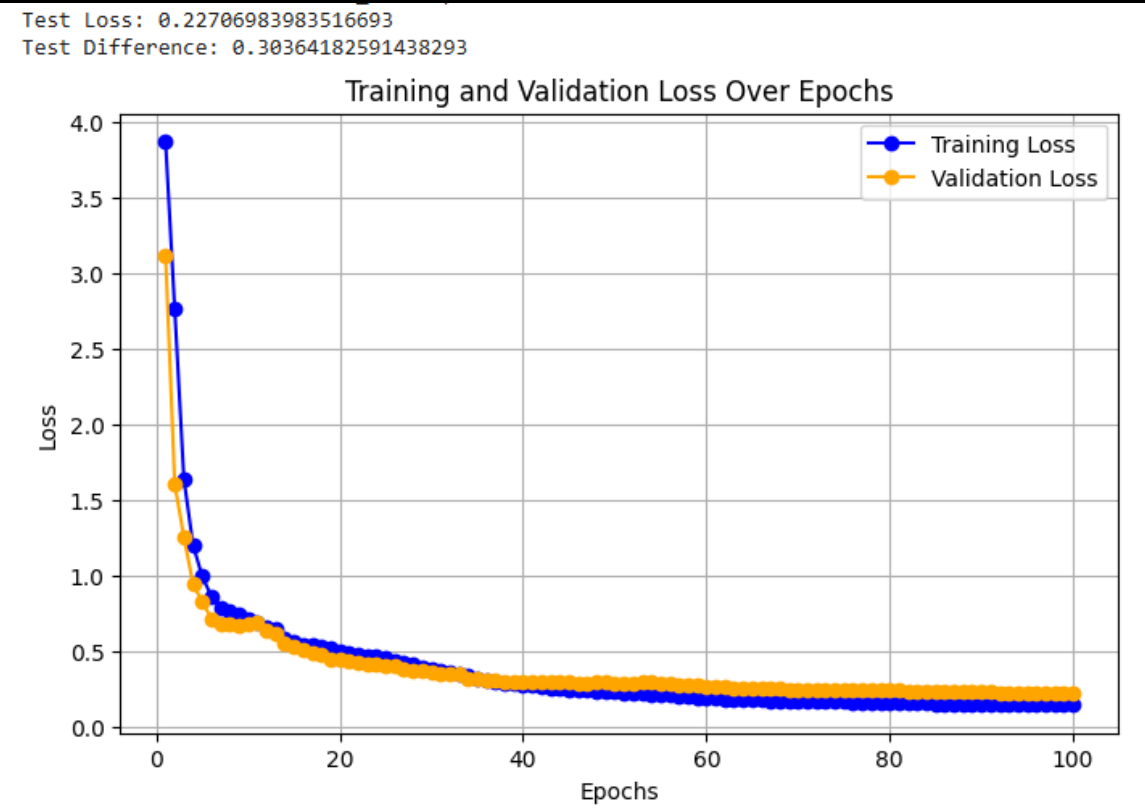
Final Model

Final Model parameters:

- 100 Epochs with early stopping
- Initial learning rate of 0.001
- 1 Message Passing Layer

Final Test Results:

- Mean Squared Loss: 0.2314
- Mean Absolute Loss: 0.3240



Sources

Cortés, I., Cuadrado, C., Daranas, A. H., & Sarotti, A. M. (2023). Machine learning in computational NMR-aided structural elucidation. *Frontiers in Natural Products*, *2*. https://doi.org/10.3389/fntpr.2023.1122426

Guan, Y., Sowndarya, S. V. S., Gallegos, L. C., St John, P. C., & Paton, R. S. (2021). Real-time prediction of 1H and 13C chemical shifts with DFT accuracy using a 3D graph neural network. *Chemical Science*, *12*(36), 12012–12026. https://doi.org/10.1039/d1sc03343c

Sanchez-Lengeling, B., Reif, E., Pearce, A., & Wiltschko, A. (2021). A gentle introduction to graph neural networks. *Distill*, *6*(8). https://doi.org/10.23915/distill.00033