PepMPNN: Ion Mobility prediction for any Post Translational Modification

Garrett P (1) , Park R (1,2), Titus J (1), Yates J (1). 1 Scripps Research Institute, La Jolla California, United States. 2 Bruker Corporation, Bremen, Germany

ABSTRACT

A peptide's Collisional Cross Section (CCS) represents its gas-phase structure, influenced by both amino acid composition and sequence. These effects lead to complex interactions, making it difficult to accurately predict a peptide's CCS. Recently Neural Networks (NN) have been shown to be an effective tool for modeling peptides. These networks often rely on Natural Language Processing (NLP) architectures which tokenize each amino acid in the sequence. While these models have shown remarkable performance, they fail to understand the molecular makeup of each amino acid and cannot predict CCS values for unseen tokens, such as unencountered post-translational modifications (PTMs). For this reason, we propose to use a Message-passing neural network (MPNN) that learns to encode each amino acid from a molecular graph and thus can predict CCS values for any amino acid or PTM.

METHODS

The experiments used were downloaded from the pride repositories: PXD019086 and PXD010012 and an external private repository. These were then searched utilizing the Prolucid search engine, resulting in over 1,00,000 unique sequence-charge pairs. ~60% of peptides were charged 2, ~30% were charged 3, and ~10% were charged 4.

To test the model's ability to predict unseen PTMs, we separated peptides with and without PTMs. The training and validation sets were created from the unmodified peptides, while the unseen PTM set was created from peptides containing oxidation and acetylation modifications. The model was trained for 24 epochs, using Adam optimizer and an MSE loss function. An early stopping callback was implemented to stop training once the validation Pearson R decreased for three continuous epochs. We then used this model to predict the CCS values for the unseen PTM set.

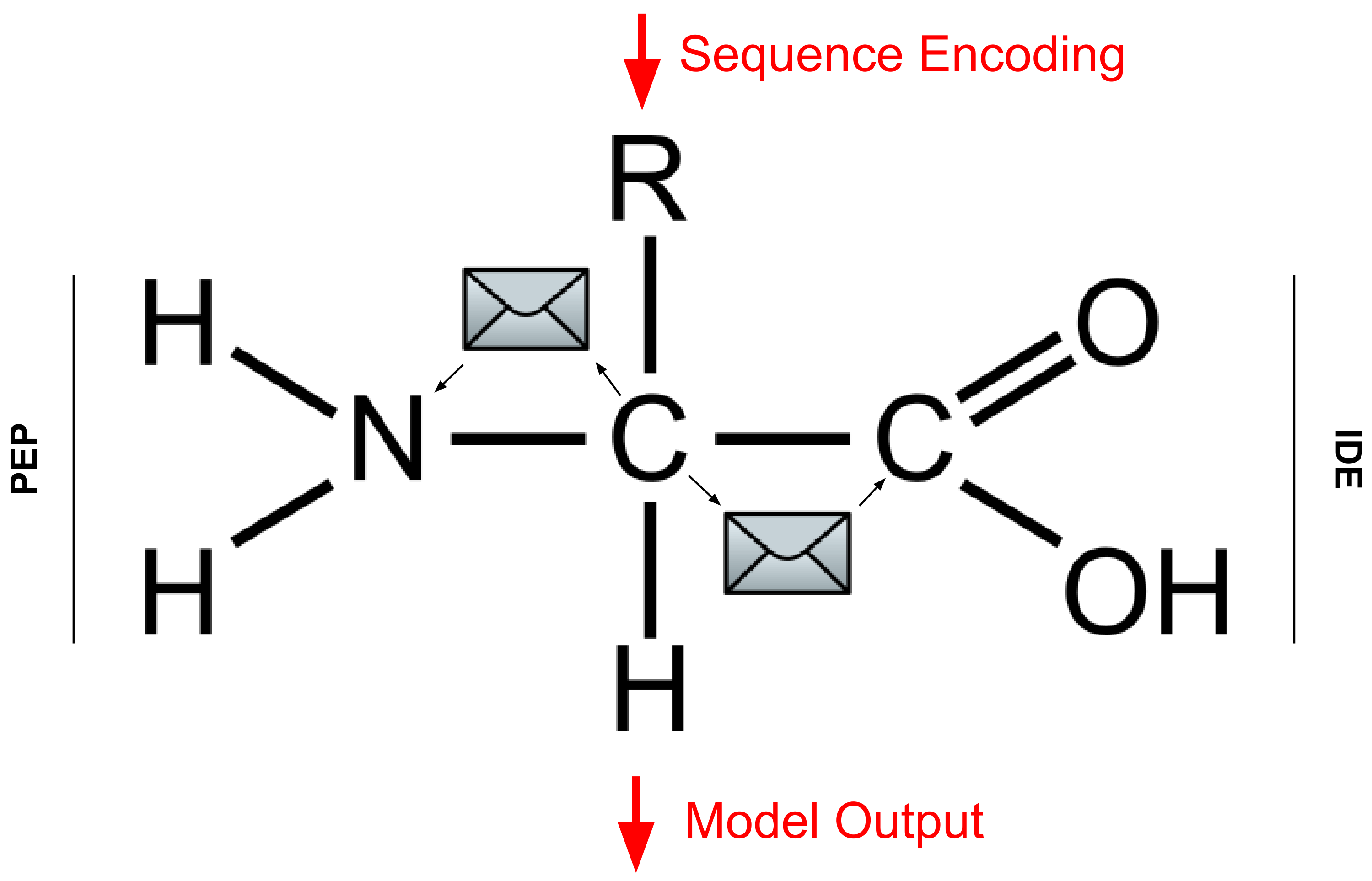
OBJECTIVES

- Predict CCS value for any peptide-like molecule
- Handle peptides of arbitrary length
- Learn the effects that peptide charge state has on Ion Mobility

MODEL

PepMPNN

Input Sequence: PEPTIDE


Sequence Encoding

Model Output

Predicted CCS: 450

Peptide sequences were converted into an undirected graph, where atoms represent nodes and bonds represent edges. Additionally, each atom and bond is represented by a feature vector based on its chemical properties. This graph-based molecular representation is processed by the MPNN and converted into a latent space feature vector of length 256. The encoded charge state is then concatenated with this vector to produce a final vector of length 260. This is then processed by dense layers, producing a predicted CCS value.

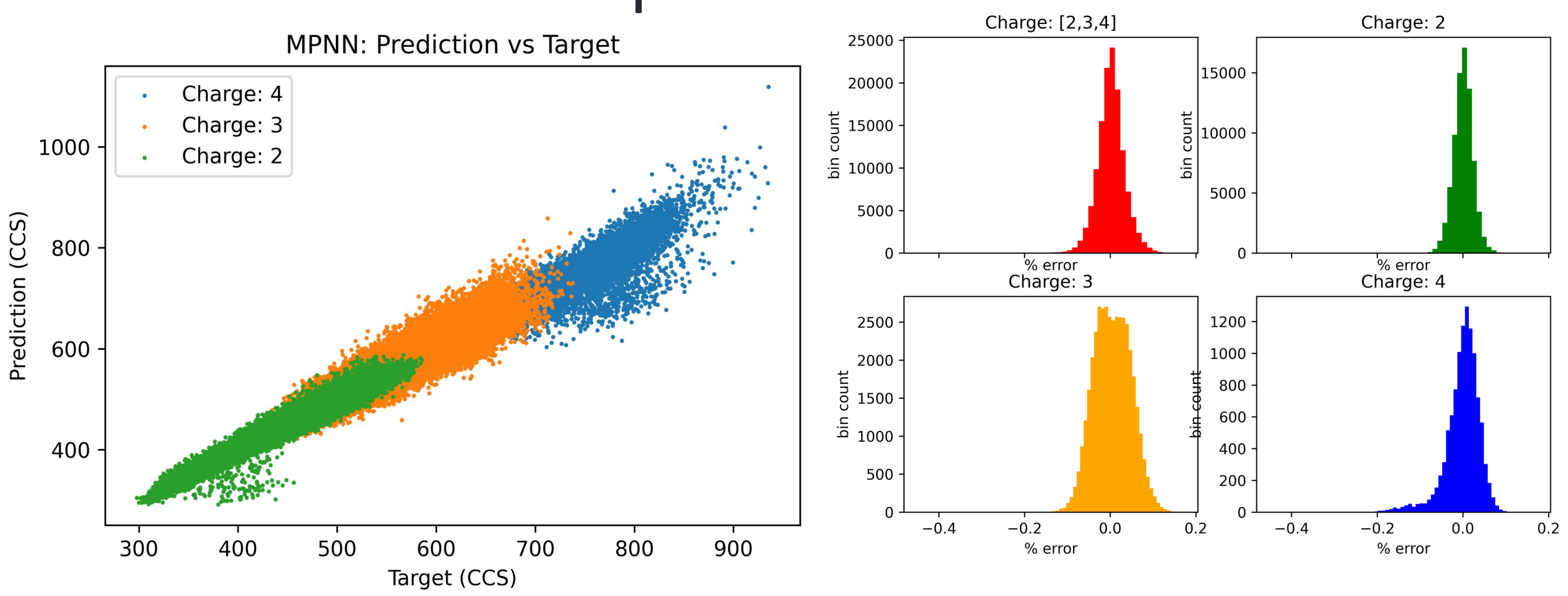
RESULTS

We found that the MPNN could predict the CCS values for peptides with unseen PTMs at a similar accuracy to that of unmodified peptides. This suggests that the MPNN architecture could interpret the graph-based representation of a peptide and could even infer the effect that unseen PTMs would have on the resulting CCS.

Further analysis will have to be done on peptides with larger and alternate PTMs as only oxidation and acetylation PTMs were investigated.

RESULTS

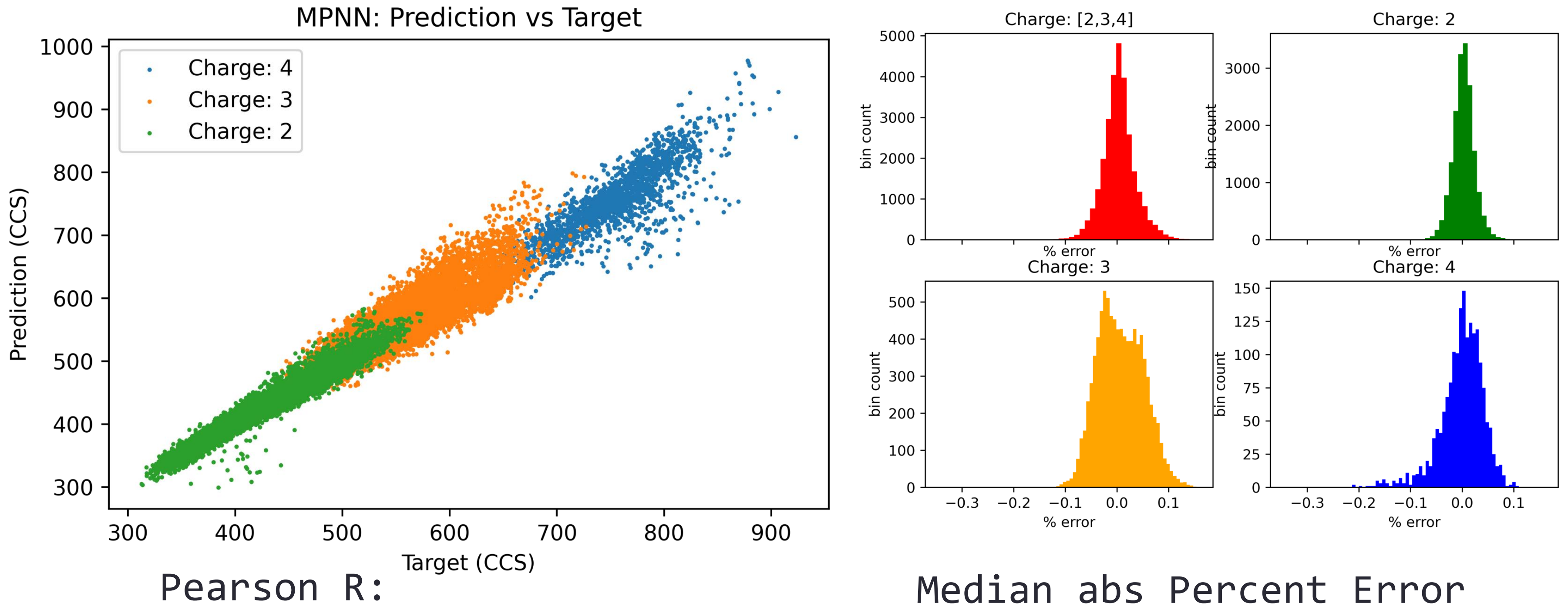
Unmodified Peptides



```
Pearson R:
All:        0.985
Charge 2: 0.977
Charge 3: 0.864
Charge 4: 0.840
```

```
Median abs Percent Error
All:       1.74%
Charge 2: 1.21%
Charge 3: 3.27%
Charge 4: 2.45%
```

Peptides with unseen PTMs



```
Pearson R:
All:        0.987
Charge 2: 0.979
Charge 3: 0.903
Charge 4: 0.858
```

```
Median abs Percent Error
All:       1.86%
Charge 2: 1.42%
Charge 3: 3.20%
Charge 4: 2.21%
```

CONCLUSIONS

- Successfully trained an MPNN model on peptides
- Model accuracy did not surpass that of current token-based variants
- Can predict CCS values for peptides with unseen PTMs