PimMDN: Peptide Ion Mobility Mixture Density Network

Patrick Garrett, Robin Park, Titus Jung, Casimir Bamberger, John R. Yates, III
Department of Chemical Physiology, The Scripps Research Institute, La Jolla, CA, USA

Abstract

Database searching and spectral library matching are the leading methods for quantifying and identifying peptides in mass spectrometry-based proteomics. Though these methods can be improved through incorporating other peptide specific variables, such as ion mobility. We created PimMDN to maximize the contribution that ion mobility measurements can make to proteomic search methods. Current mobility models predict a single value but PimMDN predicts the entire mobility distribution.

Introduction

Due to the dynamic nature of an ionized peptide, a unique peptide sequence can adopt a range of possible mobilities [1] as well as multimodal or non-parametric distributions. Standard machine learning methods, based on linear regression, are not suited to fit such distributions. They will learn to predict the mean of the distributions in order to minimize the mean squared error (MSE). Previous attempts to train deep learning models to predict a peptides ion mobility relied on pre-processing the distribution into a single mean/median value. While these models demonstrate great capabilities, they lose valuable information regarding the effects that the peptide sequence has on its mobility distribution. Furthermore, these approaches cannot learn the effects which sequence pose on the standard deviation of the mobility distribution. Without a method to interpret the specificity, the true potential of mobility predictions cannot be utilized. To combat these limitations, we created pimMDN which utilizes a Mixture Density Network (MDN) [2] to model nonparametric and multimodal distributions, as a well as provide a metric of prediction specificity.

Mixture Density Network

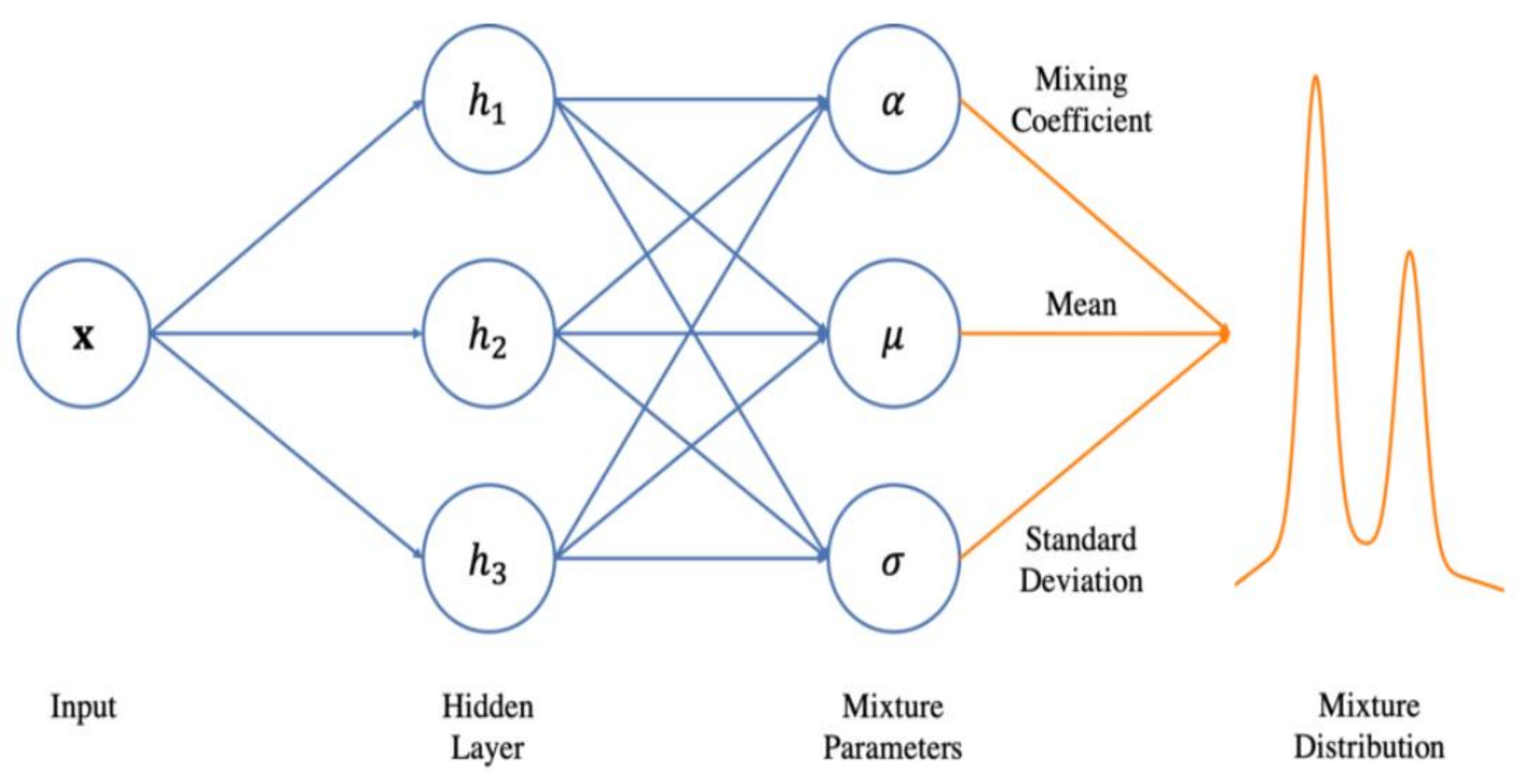

Fig. 1. The structure of an MDN is very simple. It's a combination of a Deep Neural Network (DNN) and a mixture model. It can theoretically model any probability distribution.

Data

Dataset: PXD010012 [3]

Experiments: HeLa_Fractions & HeLa_200ng_100ms

Search: Ip2 prolucid search engine [4]

Data Alignment

The reported mobilities can shift over time leading to misaligned experiments. It is important to correct any misalignment prior to training. To do this we selected the experiment with the most unique peptides as the reference and align all subsequent experiments to it. Fig 2. illustrates that the experiments mobility distributions become more similar with alignment.

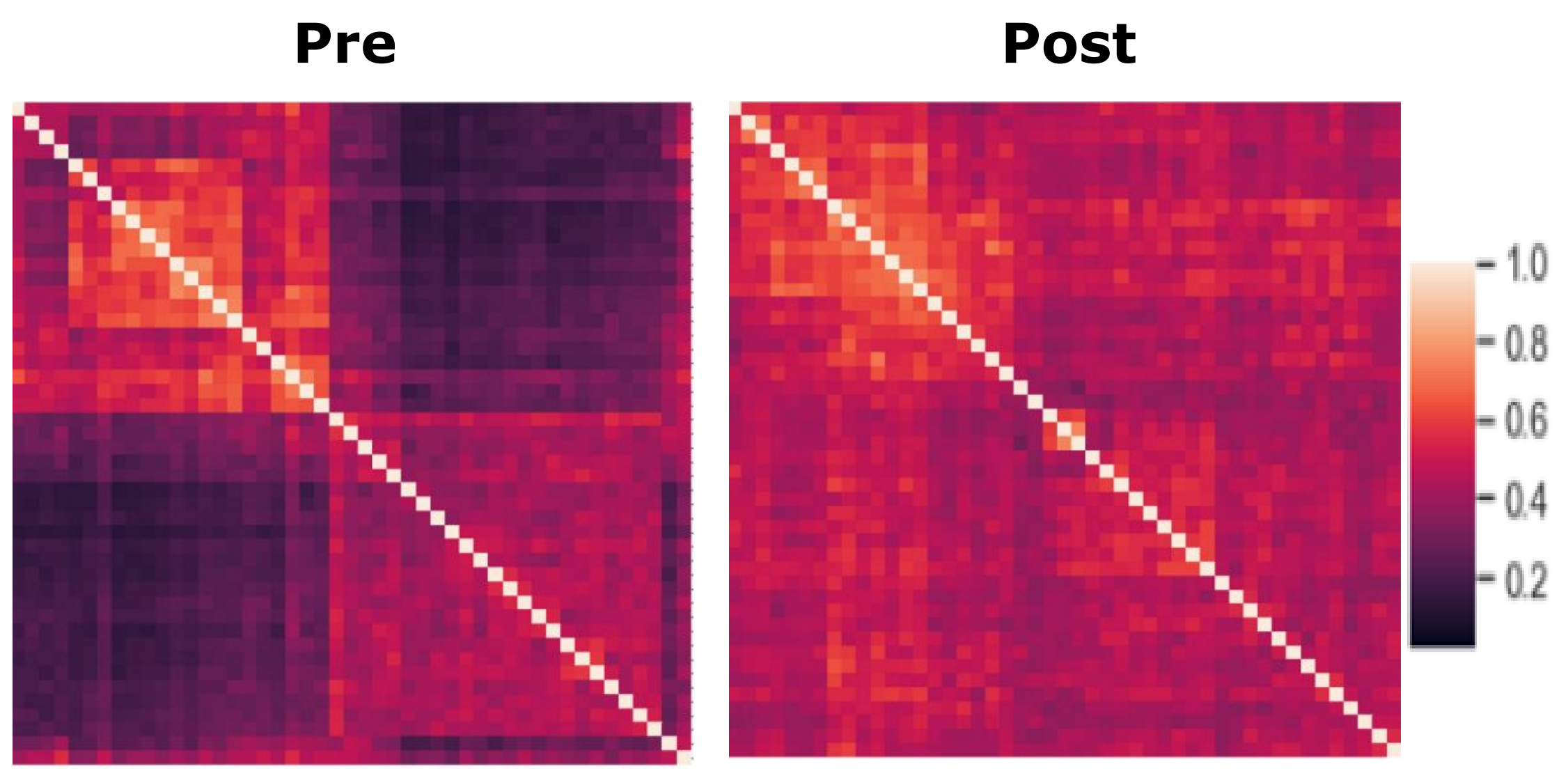

Fig. 2. Kolmogorov-Smirnov test (p-value) cluster map for experiments pre and post alignment.

Data Upscale

After alignment there still existed an uneven number of samples per peptide. Without correcting for this, the data would be unbalanced and lead to training instability. We decided to upscale peptides with fewer samples, and downscale peptides with many samples. Peptides with less than 10 samples were excluded from this process because their sample distribution is too small to model the parent distribution. A cutoff of 10 was chosen because it maximized the total number of samples while still maintaining a large proportion of unique peptides. Peptides were then up sampled/down sampled to 32 through fitting and then sampling a Kerel Density Estimation function (KDE).

Kernel Density Estimation

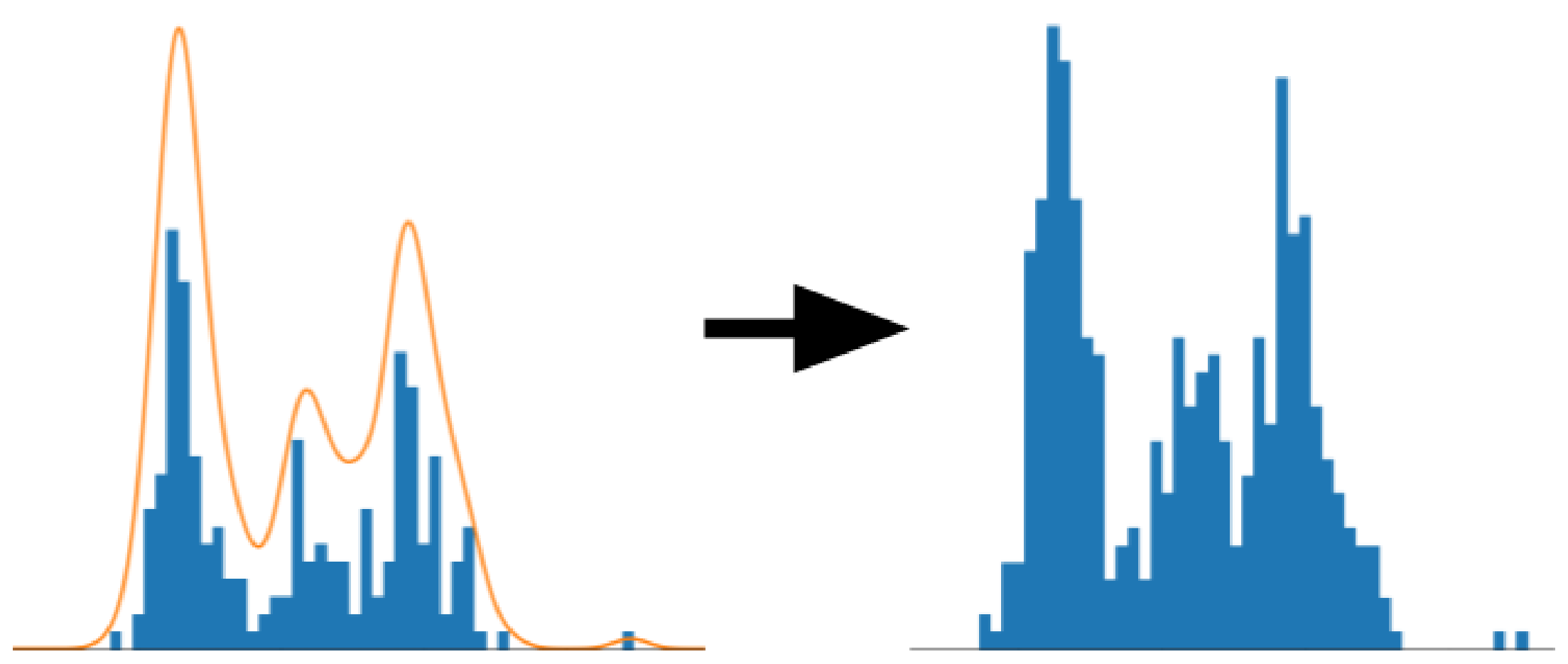

Fig. 3. Left image depicts the fitting of a KDE to the ion mobility distribution. Right image depicts the distribution after being sampled from the KDE.

Data Summary

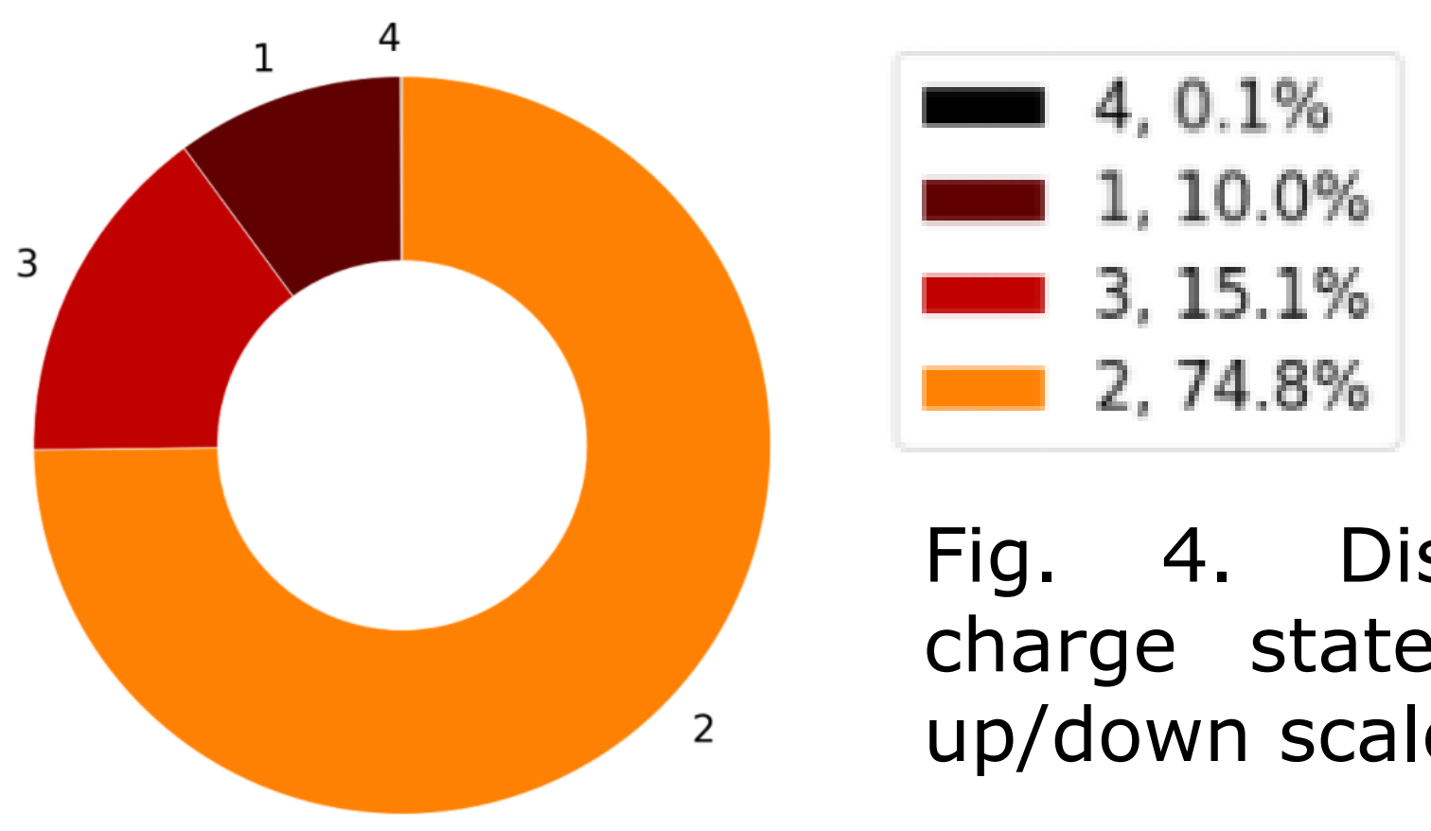

Fig. 4. Distribution of charge states across all up/down scaled peptides.

Charge 1 and 4 peptides were excluded because they comprised a small portion of the overall data. Furthermore, peptide with < 7 residues, and > 30 residues were excluded. Training and testing sets were split 9:1 over unique peptides.

Total Peptides

Total unique peptides: 28,913

Unique +2 peptides : 23,946

Unique +3 peptides : 4,843

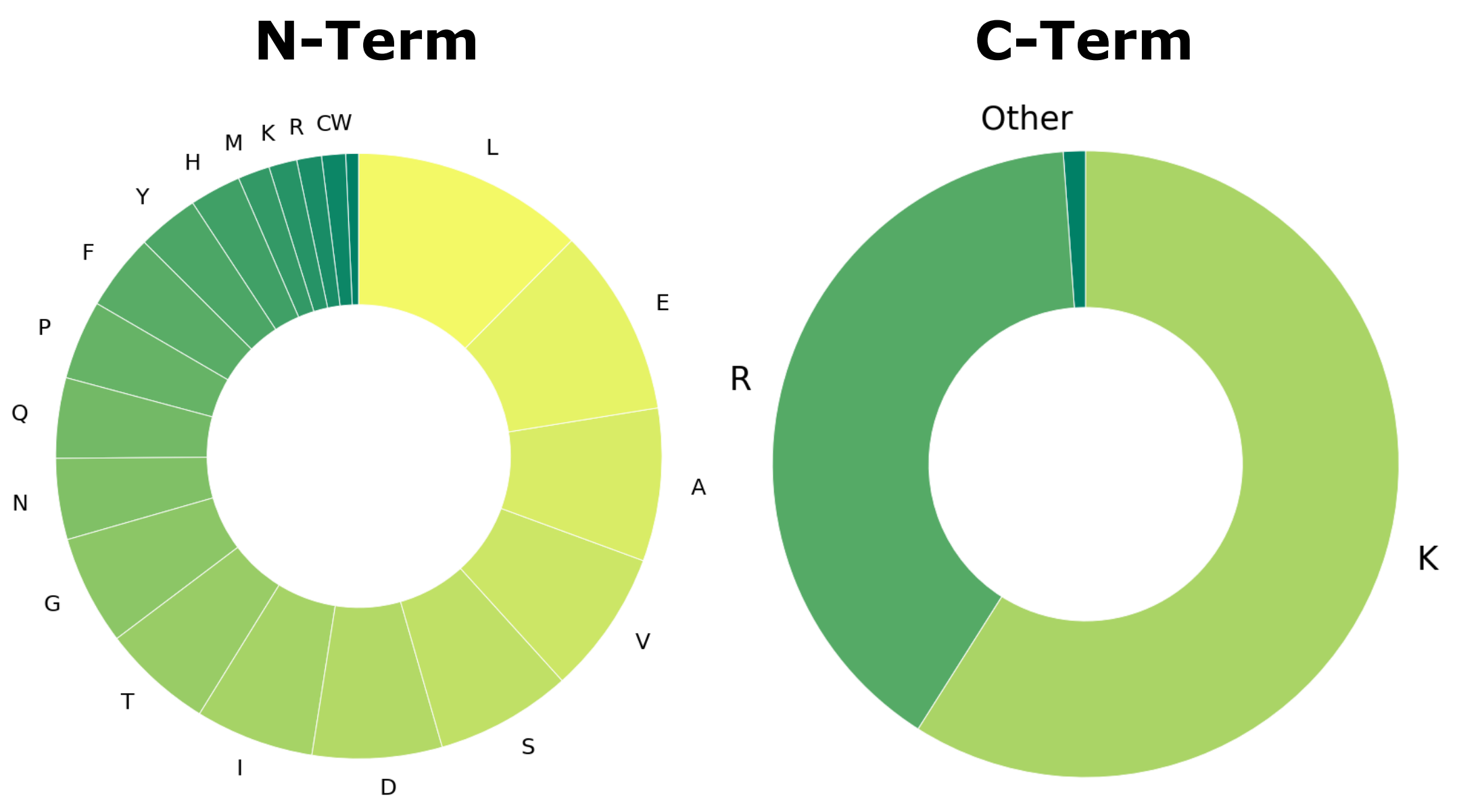

Fig. 5. The distribution of N-term/C-term amino acids.

Model Structure

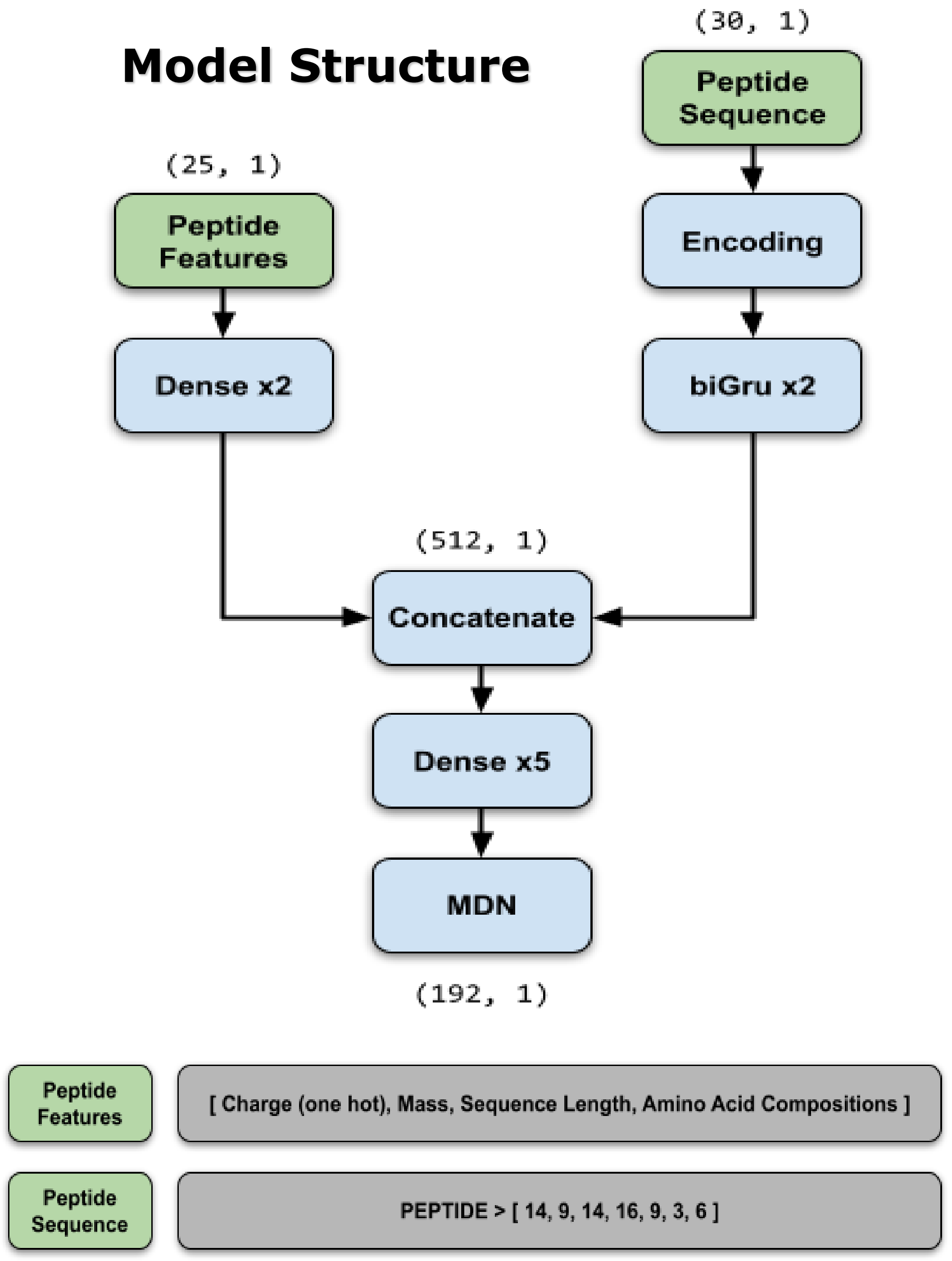

Fig. 6. Depicts PepMDN model structure. The peptide feature dense layers consist of relu activation and 256 units. The peptide sequence encoding layer uses 20 embedding units. The biGRU layers have 256 and 128 hidden units, respectively. All other dense layers have 512 units and relu activation functions. The MDN uses 64 mixture components. All hyper parameters were tuned using keras-tuner.

Training

PimMDN utilized Negative Loss Likelihood as the loss function and Adam as the optimizer. PepMDN was trained for 10 epochs on a Titan V GPU, with a batch size of 64. Total training time was 5 hours, achieving a testing loss of -2.67.

Single Value Results

+3 R-value = **0.94** +2 R-value = **0.99**

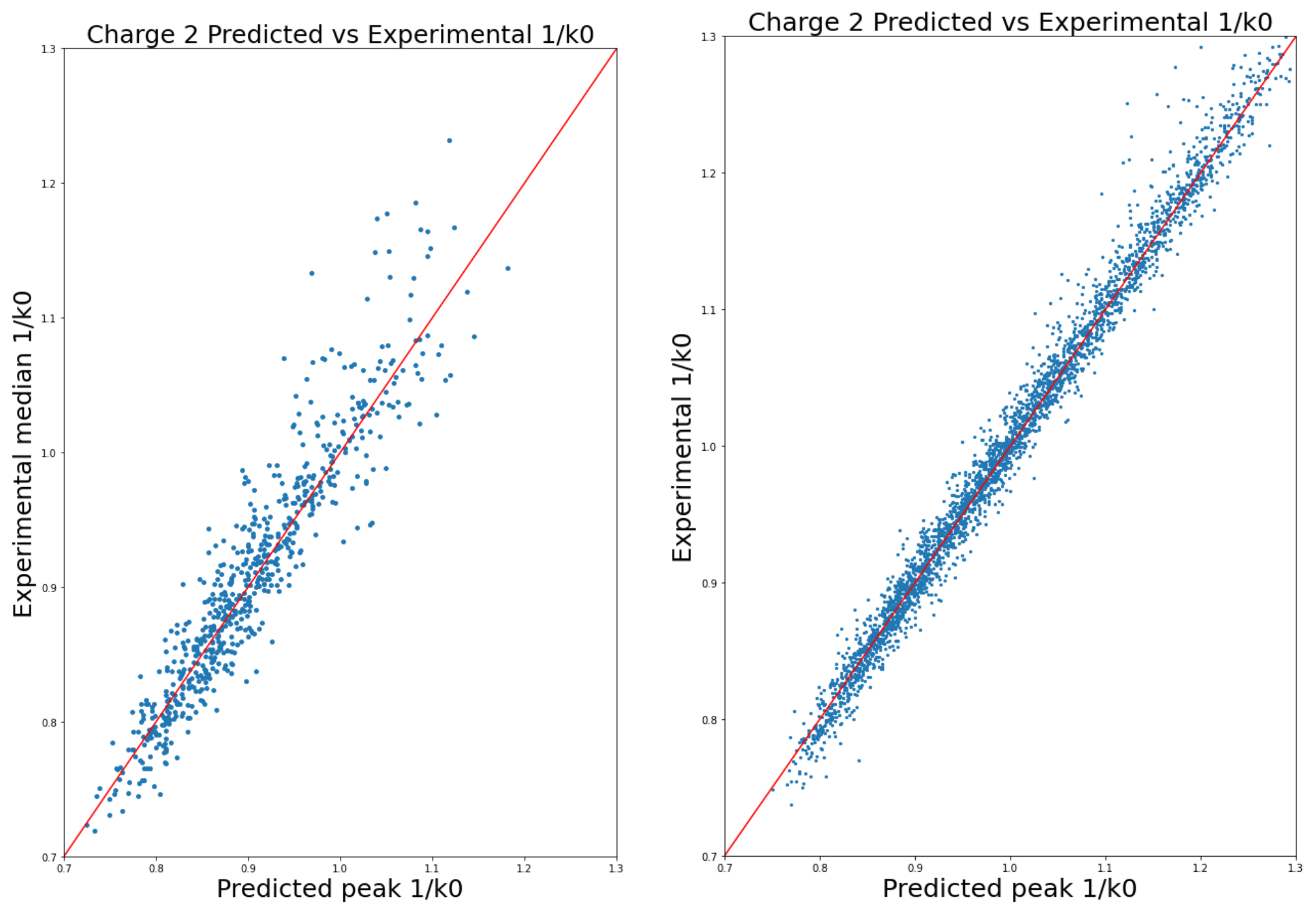

Fig. 7. Predicted vs experimental mobility values for both +2 and +3 peptides.

Distribution Results

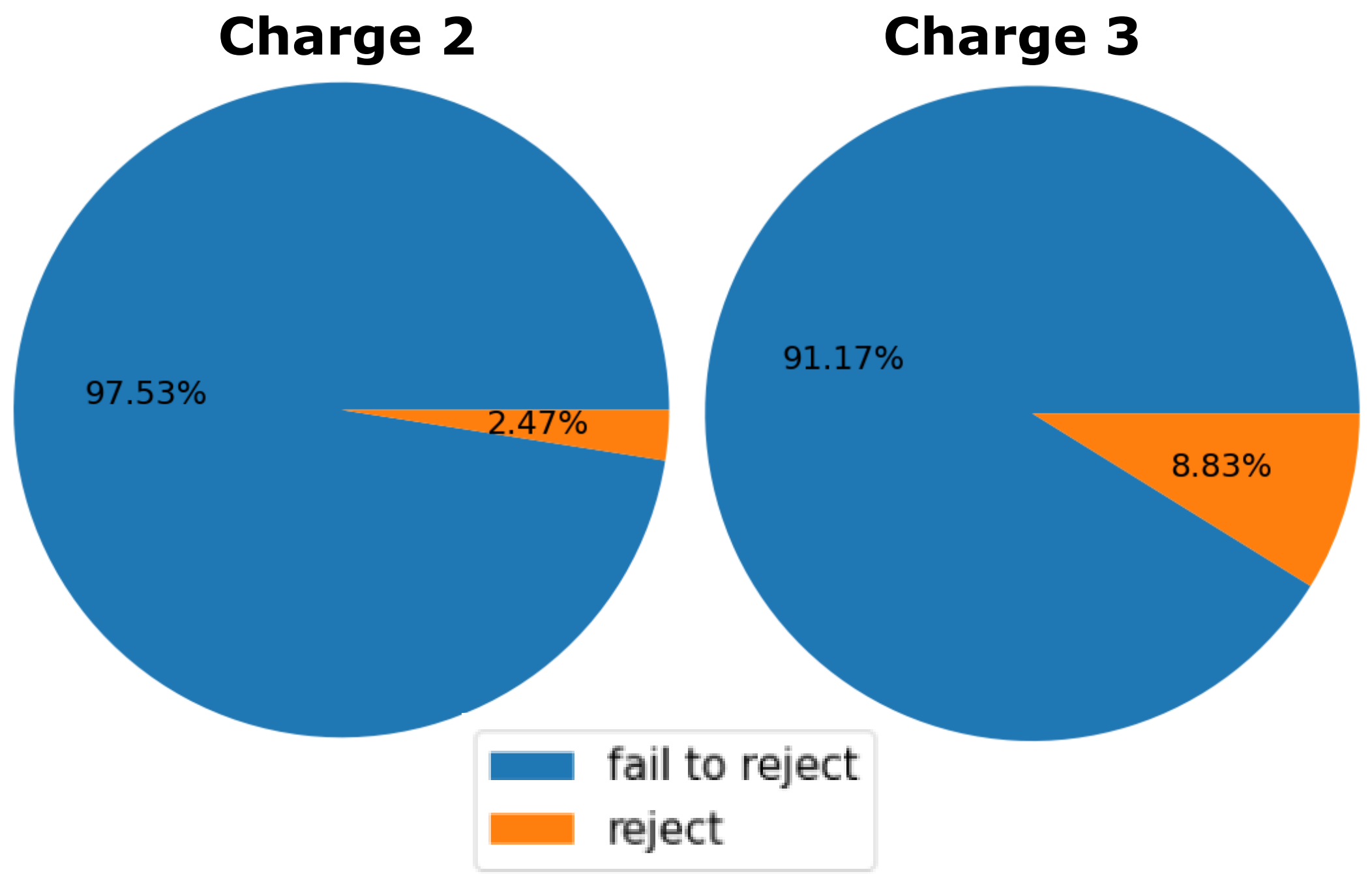

Fig. 8. Kolmogorov-Smirnov test for mean aligned experimental and predicted distribution. We used an alpha value of 0.05. The KS-test assumes that the distributions are the same. Failing to reject this hypothesis means that the experimental and predicted distributions are similar.

Conclusion

Overall PimMDN is highly accurate for +2, but slightly less accurate for +3. This can be attributed to there being less +3 peptides in the data. The "Single value results" demonstrates that PimMDN learned to place distributions correctly. Similarly, the "Distribution results" demonstrates that PimMDN learned to predict the overall shape of the mobility distributions.

References

1. Chih-Hsiang Chang, Darien Yeung, Victor Spicer, Oleg Krokhin, Yasushi Ishihama bioRxiv 2020.09.14.296590;
2. Christopher M. Bishop, Mixture Density Networks (1994)
3. Meier F, Brunner AD, Koch S, Koch H, Lubeck M, Krause M, Goedecke N, Decker J, Kosinski T, Park MA, Bache N, Hoerning O, Cox J, Räther O, Mann M. Online parallel accumulation - serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. Mol Cell Proteomics. 2018, PubMed: 30385480
4. Integrated Proteomics Pipeline (IP2) – http://manual.integratedproteomics.com/1/en/topic/14-1-for-your-publications