Predicting RNA-Binding Preferences for Proteins

Millie Dwyer, Julie Helmers, Shasha Lin, Lihan Yao

New York University Center for Data Science, Fall 2017

Project Objectives

- Learn low dimensional representations of RRMs to identify proteins with similar binding preferences
- Apply deep learning methods to various representations of RRM sequences

Introduction

RNA Recognition Motifs (RRMs) are substructures of some RNA-binding proteins that specify which RNA sequences the protein will bind. For most identified RRMs, it is unknown which RNA sequences they bind, out of tens of thousands of possibilities. However, there are many RRMs that are homologous, meaning that they share sequence patterns as well as binding behaviors. If it is possible to generate a representation of these similarities and relate this representation to binding preferences, it may be easier to predict how a newly identified RRM will behave. This approach would be faster and less expensive than experimentally determining behavior for individual RRMs.

Data

RRMs are represented as sequences of letters drawn from a vocabulary of 20 amino acids. These sequences are at most around 90 letters long.

- Unlabeled data: 99,000 RRM sequences obtained from the Pfam [6] protein family database. Used for training unsupervised models.
- Labeled Data: 400 RRM-containing proteins for which the distribution of binding affinity over RNA sequences is known. Used to evaluate the success of our representations through similarity evaluation.

Sample Data

For some models, we aligned the original sequences from the labeled and unlabeled data to make similar structures appear in the same locations. Alignment is achieved through an external Hidden Markov Model (HMM) available on the Pfam database website. The Clustal Omega program [1] is used to unify labeled and unlabeled sequence alignments.

- Sample Aligned RRM Sequence:

```
> T117424||RNCMPT00259_RRM__0−−−−−−TPSTNVFINY−−IP−−−−−−−−−−P−
−RF−−−−−−−−−−−−−−−−−−−−−−−−−T−−−−−−−−−−E−−−−QD−−−−−−−−
−−L−R−−−N−−−−−−−−−−−−−−−−−−−−−−−−−−−−−L−−−−−−−CS−Q−−−Y−−−
−−−−−−−−−−−G−−−−−−−−−−−−−−−−−−−−E−−I−IS−−−−−−−−−−−−−−−−−−−−−−−
−−−−−−−−−−−−S−−−−−K−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
−−−−−−−−−−−−−−−−−−−−−−−−−−−−IM−−−−−−−−−−−−−−−−−−−−−−−−−−−−
−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−I−NL−
−−E−−−−−−−−−TG−−−−−QSKCFG−−−F−−−−−−−−−−−−−V−−−−−−−−−−−−−K−
−−−−−−−−−−F−−−−−−−−−R−−−−−−−E−−−−−−−−−L−−−−−−−−−−−−−S−−−−−−
−Q−−−−−−−−−A−−H−−−−−−−−A−−−−−−−−−A−−−−−−−−I−QA−−−−−−−−−−−−I−
−−−D−−−G−−−−M−−−−−−−SIGN−−−−−−−−−K−−−−−−R−−LLAKYAESQE−−−−−
```

- Sample Unaligned RRM Sequence:

```
> T117424||RNCMPT00259_RRM__0
TPSTNVFINYIPPRFTEQDLRNLCSQYGEIISSKIMINLETGQSKCFGFVKFRELS
QAHAAIQAIDGMSIGNKRLLAKYAESQE
```

Methods

We extracted our RRM representations from the hidden states of three models inspired by natural language processing (NLP):

- Seq2Vec: Applies Doc2Vec to biological sequences to learn distributed representations for similarity prediction [3].
- CNN+LSTM: Combines a ResNet encoder and Long-Short Term Memory (LSTM) decoder. Two sets of representations are learned: one from the features extracted by ResNet, and the other from the hidden state of LSTM decoder at the last time stamp. Inspired by an image captioning model [2].

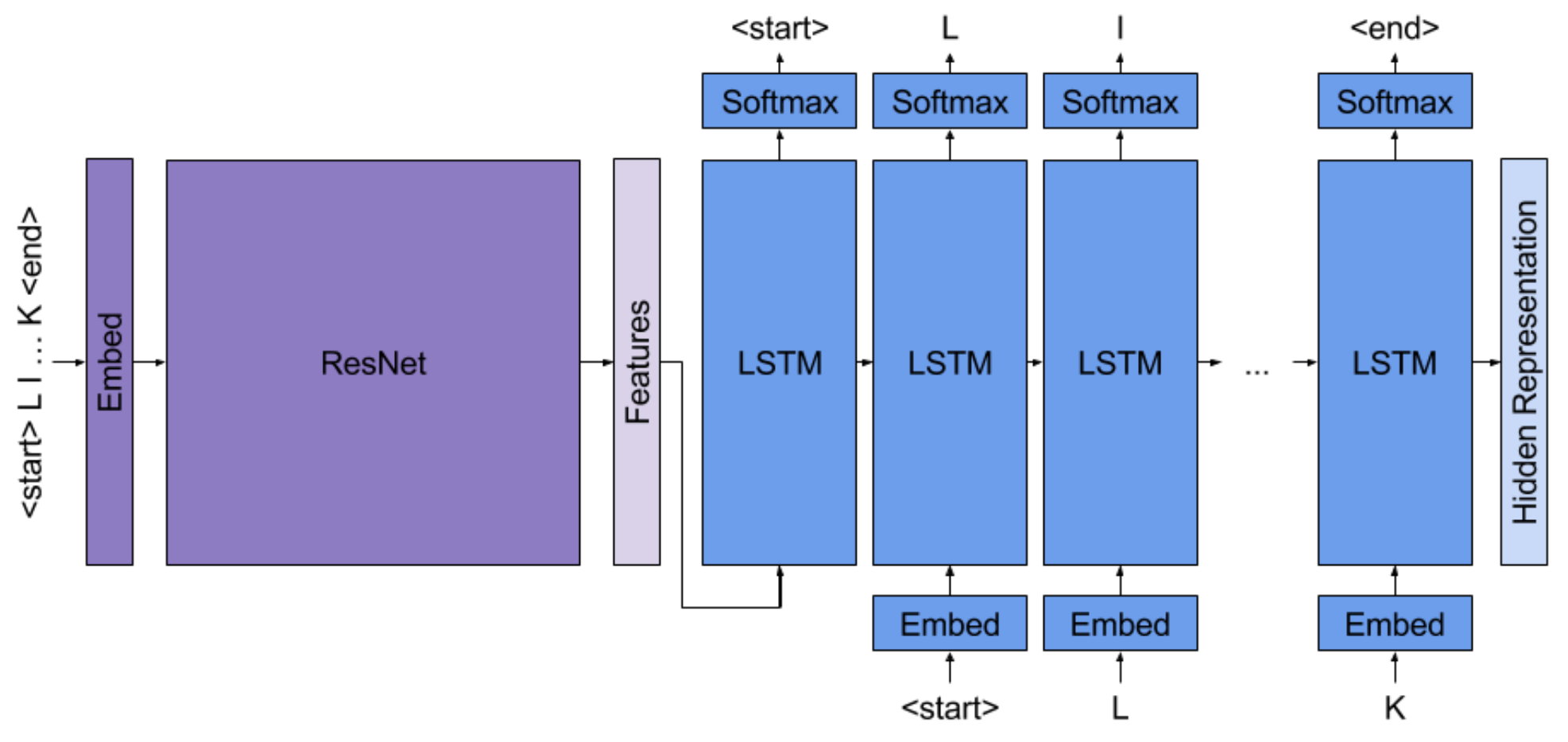
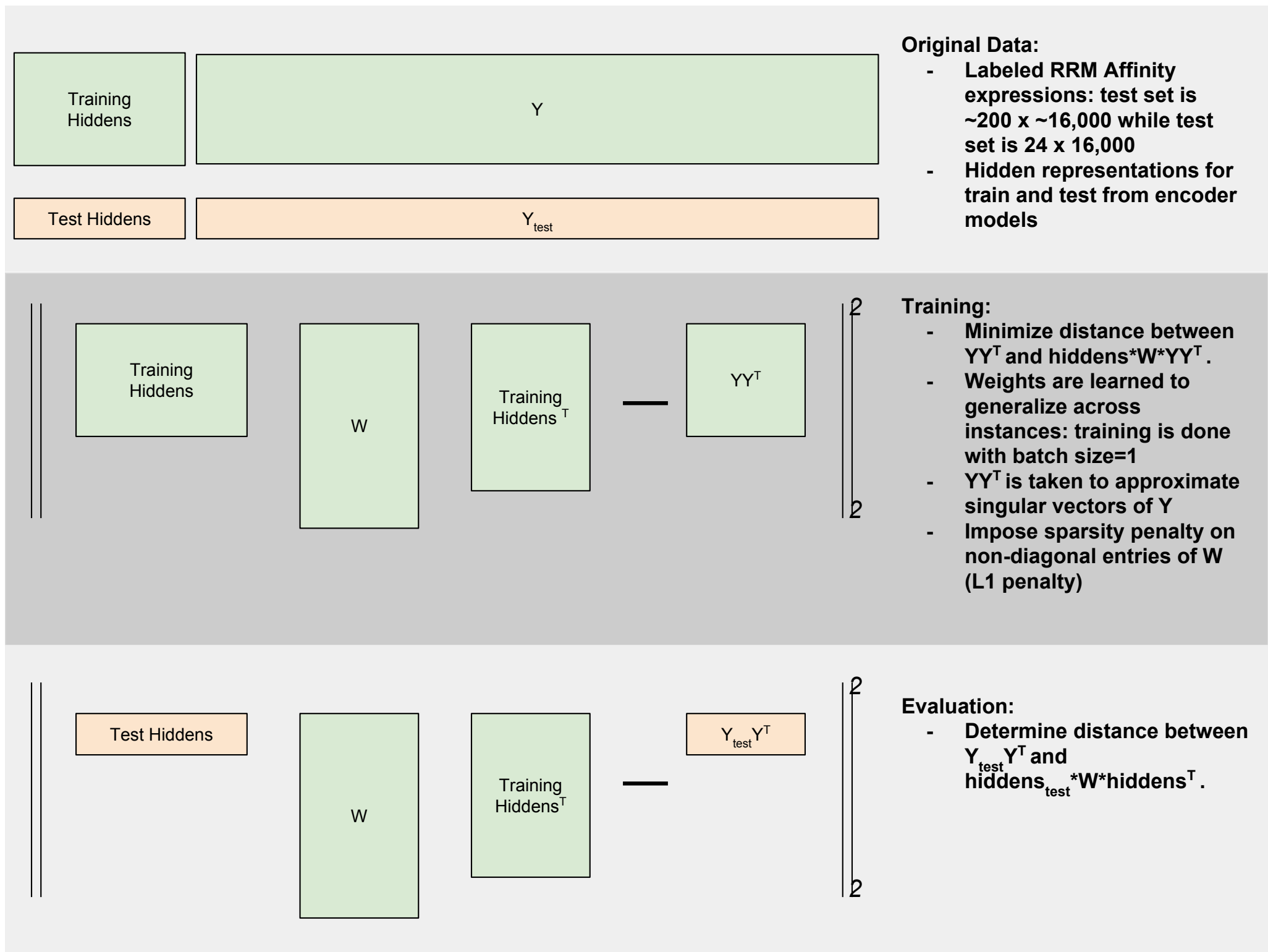


Figure 1: CNN+LSTM architecture

- Character-Level Neural Machine Translator (NMT): Sequence motifs are learned via convolutional layers, and then LSTM captures long range motif interactions. Borrowed from a machine translation architecture [4].

Evaluation

tSNE visualizations are used to subjectively evaluate our low dimensional representations: Similar RRMs should appear to cluster. In addition, Similarity Regression is used to evaluate predictive quality of the representations. It is based on Affinity Regression [5], where a weight matrix is learned to translate from a sequence representation to its likely binding. Our method of Similarity Regression differs in its use of learned hidden representations, allowing for a more direct comparison between hidden states for similar binding affinities.



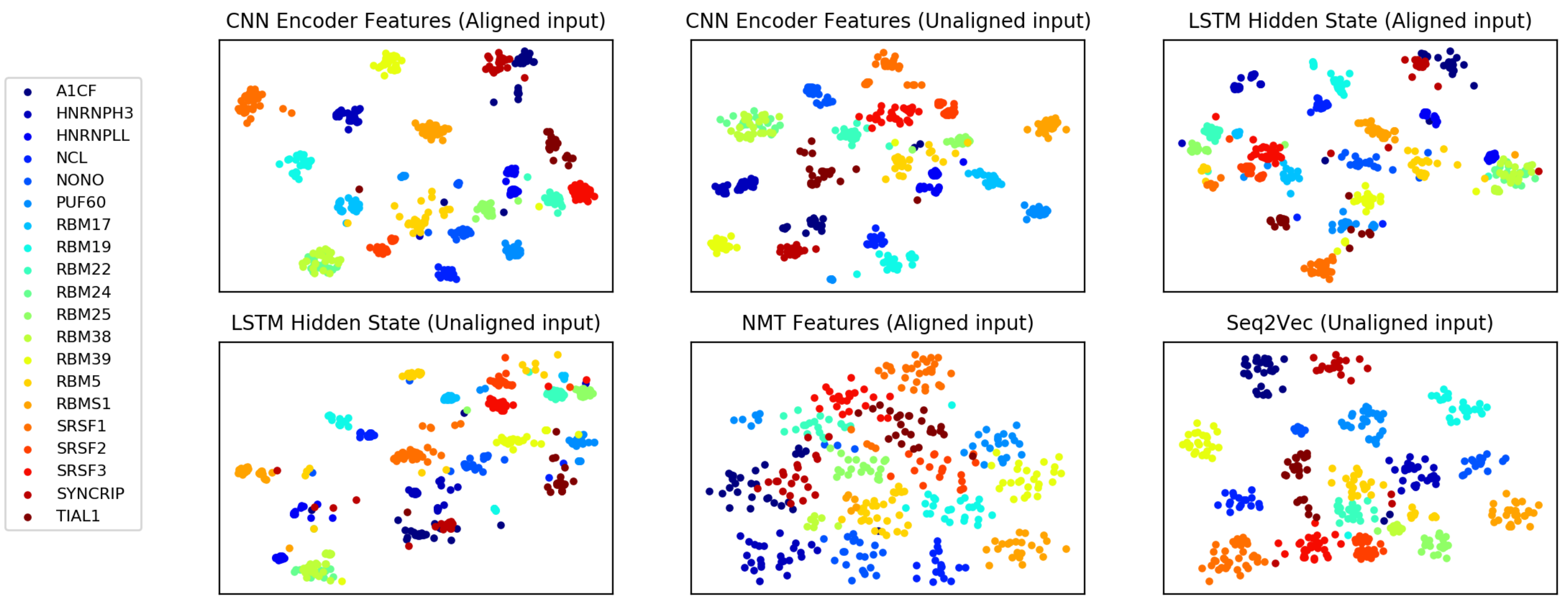
Original Data:
- Labeled RRM Affinity expressions: test set is ~200 x ~16,000 while test set is 24 x 16,000
- Hidden representations for train and test from encoder models

Training:
- Minimize distance between YY^T and hiddens*W*YY^T.
- Weights are learned to generalize across instances: training is done with batch size=1
- YY^T is taken to approximate singular vectors of Y
- Impose sparsity penalty on non-diagonal entries of W (L1 penalty)

Evaluation:
- Determine distance between $Y_{test}Y^T$ and $hiddens_{test}$*W*$hiddens^T$.

Figure 2: Similarity Regression method

Results



CNN Encoder Features (Aligned input)

CNN Encoder Features (Unaligned input)

LSTM Hidden State (Aligned input)

LSTM Hidden State (Unaligned input)

NMT Features (Aligned input)

Seq2Vec (Unaligned input)

Figure 3: Selected tSNE visualizations

Model	Dimension	Alignment	Training Loss	Validation Loss
CNN+LSTM Encoder	64	aligned	4556.46	4868.18
CNN+LSTM Encoder	64	unaligned	4665.3	4869.09
Seq2Vec	64	unaligned	4339.12	4879.70
CNN+LSTM Decoder	128	unaligned	4821.45	4973.89
Seq2Vec	128	unaligned	3604.97	4988.94
CNN+LSTM Decoder	128	aligned	4818.95	5000.74
Character-Level NMT	64	aligned	4330.47	5081.11
Character-Level NMT	128	aligned	3740.37	5082.57
Naive k-mer (ngram)	64	N/A	4017.92	17100.47
Naive k-mer (ngram)	128	N/A	3062.86	39558.71

Table 1: Similarity Regression results

Acknowledgments

We would like to thank Professor Quaid Morris and Alexander Sasse for their invaluable guidance and assistance.

Conclusions

- All three models inspired by NLP methods were able to learn meaningful representations of RRMs, either with unaligned sequences or aligned ones, as demonstrated by clear clustering of different protein types in tSNE visualizations.
- Features learned from the novel application of an image captioning architecture performed as well as or better than Seq2Vec features, and significantly better than naive ngram/k-mer features, in the downstream Similarity Regression task.
- Using unaligned RRM sequences was at least as successful as using HMM aligned sequences. While a highly useful technique in a large range of biology problems, HMM based sequence alignment may not be a necessity for analyzing protein sequences with deep learning.

Future Work

- Train models on entire protein sequences, and learn binding preferences without extracting the RRM substructures first.

Appendix

While preprocessing the aligned sequences, we eliminated positions that were populated in less than 1% of the data. The graphs below show the population frequency for each position.

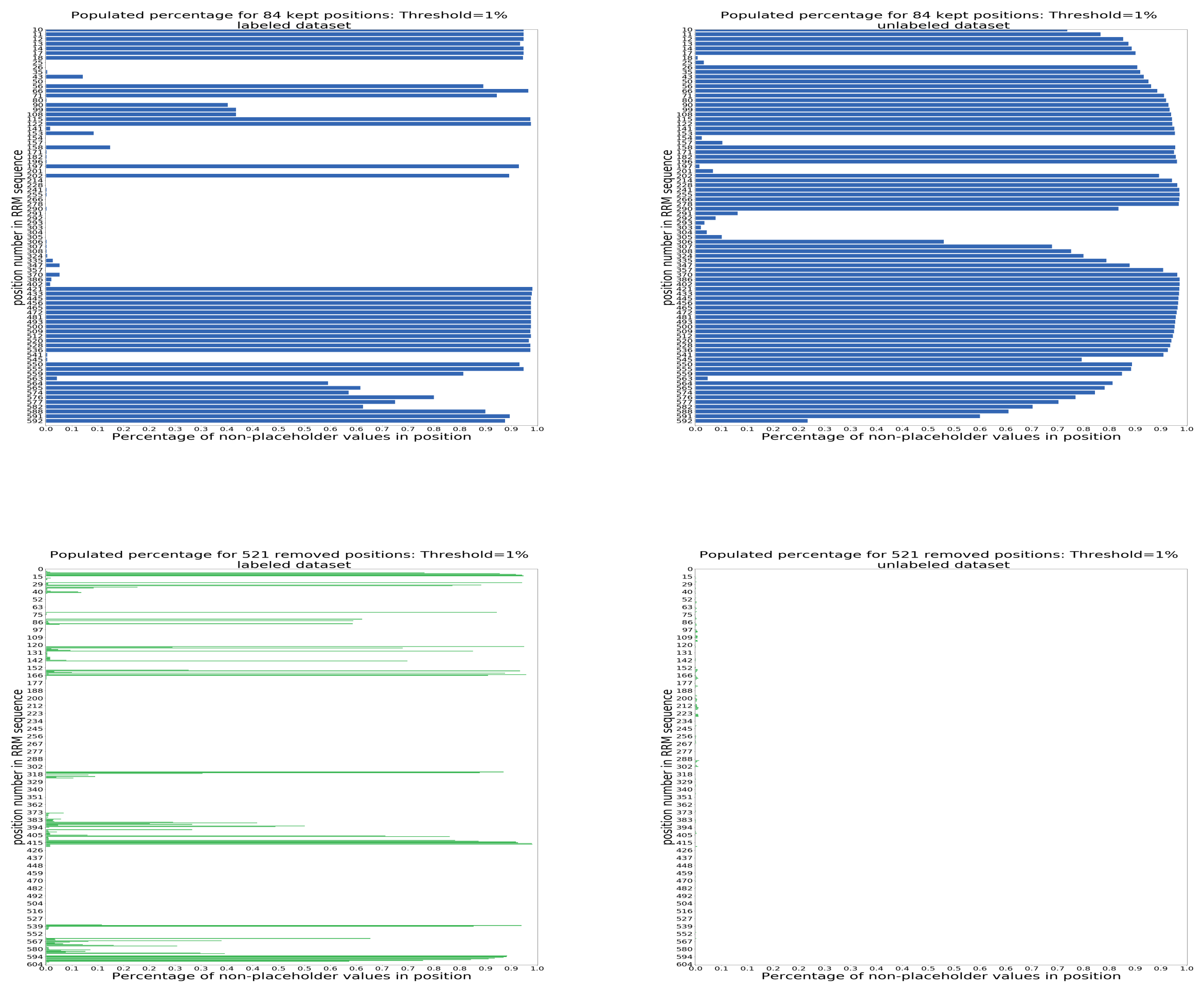


Figure 4: Population frequency by sequence position

References

[1] *Clustal Omega program.* https://www.ebi.ac.uk/Tools/msa/clustalo/. Accessed: 2017-12-10.

[2] Yunjey on GitHub. *Image captioning tutorial.* https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning. Accessed: 2017-11-09.

[3] Dhananjay Kimothi et al. "Distributed representations for biological sequence analysis". In: *arXiv preprint arXiv:1608.05959* (2016).

[4] Jason Lee, Kyunghyun Cho, and Thomas Hofmann. "Fully character-level neural machine translation without explicit segmentation". In: *arXiv preprint arXiv:1610.03017* (2016).

[5] Raphael Pelossof et al. "Affinity regression predicts the recognition code of nucleic acid-binding proteins". In: *Nature Biotechnology* (2015).

[6] *Pfam protein family database.* http://pfam.xfam.org/. Accessed: 2017-12-10.