Statement of Purpose

[REDACTED] *Ph.D. Application, Department of Computer Science, Stanford*

My objective is to pursue a Ph.D. in Computer Science with a focus on language learning in bioinformatics. In my past research I have investigated (a) language modeling for proteins and molecules [7, 8, 11, 14], (b) protein prediction tasks [3, 5, 12, 15], (c) continuous control [13], (d) biomedical feature predictors [1, 6], (e) character level neural machine translation [2], (f) optimizing model selection [4], and (g) a novel new way to perform and evaluate arithmetic computation in neural networks [10].

My motivation to pursue a Ph.D. is in large based on my accumulated research experience in ML, which, among other, includes: [REDACTED] honors program, a stay at [REDACTED] under Professor [REDACTED], [REDACTED] under [REDACTED], and my current position at the [REDACTED] under Professor [REDACTED]. These ventures into serious research, over the past four years, has affirmed my passion for technical research projects. I have found myself the most motivated when challenged with developing new algorithms, pushing state-of-the-art solutions, and uncovering structures in data. Ultimately, the aim is to build an academic career and become a Professor.

My interest in deep learning started when working along [REDACTED] on secondary structure prediction for proteins. Given instructions on methodologies, I independently obtained state-of-the-art in secondary structure prediction, which resulted in a second-author paper for Bioinformatics [3]. We later bested these results, in a first author paper [5], by adding a CRF layer to the classifier — improving over much more complex models from institutions such as Google Brain.

During my thesis, I found that word level hierarchies of characters can improve performance on neural machine translation [2]. At the time, TensorFlow was only a couple of months old and I submitted my code-base as the first dynamic sequence-to-sequence library; including a decoder, attention, and a loss function. My contributions has since been used by many researchers working on sequence-to-sequence models.

Most recently, through my research assistantship in 2019, I independently built a student research lab and managed to recruit over 30 M.Sc. students, mainly from the 2018 deep learning class, to investigate advanced topics (reinforcement learning) and work on my research ideas (13 M.Sc. thesis). I found the research proposals either; by myself, or by inquiring bioinformatics, chemistry, and biomedical researchers on their datasets and proposing deep learning solutions. The common denominator for applied projects has been a sequential problem that can be investigated with NLP based methodologies. My contribution as group leader has been idea generation, literature study, weekly supervision, code reviews, and paper writing. It is worth noting that the following submissions had first-authors with little to no prior research experience [9, 10, 11, 12, 15].

In an investigation of language modeling on molecules, last author [11], we made a novel adaption to the transformer such that it can encode weighted graphs (bond orders). We then performed masked language modeling over atoms in a discrete molecular graph. To our surprise this model, even with binary bond order, is able to infer bond order and learn to approximate the octet rule, relations in hypervalent molecules, and ions. This is useful for drug discovery as it can filter out molecules with low likelihood.

Language modeling in protein sequences is particularly interesting because of the high resemblance to text, which is why many researchers have recently investigated UniProt to build massive language models. However, little effort has been directed towards analyzing the properties of the datasets used to train these language models [7, 8]. We developed, in an upcoming co-first author submission for Bioinformatics [14], a new language modeling datasets[1] for proteins that highlights challenges with certain domains (in particular Archaea, Virus) and dataset quality. Moreover, we find that language models can be used to filter unlikely proteins, which is an important research direction as 99% of UniProt is of predicted quality.

Neural networks seldom generalize well outside of the range of numerical values encountered during training. In an in-depth analysis of recently propose Neural Arithmetic Logic Unit (NALU), we motivate a novel new interpretable arithmetic unit for exact multiplication, last author [10] under review at ICLR (top 15% ratings). This method can, with 65% convergence rate, in just 10k iterations, learn exact discrete function representations in spaces of over 10^{477} unique discrete combinations. As the solution is sparse our proposed solution can be easily interpreted. Moreover, to highlight the importance of consistent convergence we have developed a new success-criterion for measuring arithmetic extrapolation performance, which we

[1] https://github.com/alrojo/UniLanguage

will present at the SEDL Workshop at NeurIPS 2019, last author [9].

Though I am open to a wide variety of research within Machine Learning, my experience in working with language models in bioinformatics and chemistry, as well as building learned interpretable arithmetic components, has inspired an interest for extracting latent structures in data. More concretely, I am interested in leveraging my previous work to build contextual representations that can used for protein prediction tasks and developing new methods which better generalize and interpret these models, which I believe proteins are particularly well suited for given their high degree of structure. Aside from this research, I am always open to interesting technical projects in other areas of deep learning, and hope to gain a deep understanding of the field during my Ph.D.

At Stanford, there are a few professors whose projects are especially interesting to me: [REDACTED], [REDACTED], [REDACTED], and [REDACTED]. After reading several papers in each of these groups, I see a clear fit for my skills and interests at Stanford and am confident that it is a great place for me to pursue a Ph.D.

[1] **A. Johansen**, J. Jin, T. Maszczyk, J. Dauwels, S. Cash, M. Westover. "Epileptiform spike detection via convolutional neural networks." in IEEE ICASSP 2016.
[2] **A. Johansen**, J. Hansen, E. Obeid, C. Sønderby, O. Winther. "Neural Machine Translation with Characters and Hierarchical Encoding." in RNN Symposium at NIPS 2016.
[3] V. Jurtz, **A. Johansen**, M. Nielsen, J. Armenteros, H. Nielsen, C. Sønderby, O. Winther, S. Sønderby. "An introduction to deep learning on biological sequence data - examples and solutions." in Oxford Bioinformatics 2017.
[4] **A. Johansen**, R. Socher. "Learning when to skim and when to read." in REPL4NLP Workshop at ACL 2017.
[5] **A. Johansen**, C. Sønderby, S. Sønderby, O. Winther. "Deep Recurrent Conditional Random Field Network for Protein Secondary Prediction." in ACM BCB 2017.
[6] A. Mohebbi, T. Aradóttir, **A. Johansen**, H. Bengtsson, M. Fraccaro, M. Mørup. "A deep learning approach to adherence detection for type 2 diabetics." in IEEE EMBC 2017.
[7] J. Armenteros, **A. Johansen**, O. Winther, H. Nielsen. "Learning the Language of Life." in ISMB/ECCB 2019.
[8] J. Armenteros, **A. Johansen**, O. Winther, H. Nielsen. "Language modeling for biological sequences — curated datasets and baselines." in LMRL Workshop at NeurIPS 2019.
[9] A. Madsen, **A. Johansen**. "Measuring Arithmetic Extrapolation Performance." in SEDL Workshop at NeurIPS 2019.
[10] A. Madsen, **A. Johansen**. "Neural Arithmetic Units." in ICLR 2020. **Currently Under Review, Ratings (8,3,6,6)**
[11] J. Olsen, P. Christensen, M. Hansen, **A. Johansen**. "Autoencoding undirected molecular graphs with neural networks." in JCIM. **Currently Under Review**
[12] M. Gíslason, H. Nielsen, J. Armenteros*, **A. Johansen*** (*equal contribution). "Prediction of GPI-Anchored proteins with pointer neural networks." in PROTEINS 2020. **Currently Under Review**
[13] A. Mohebbi, **A. Johansen**, N. Hansen, P. Christensen, M. Jensen, J. Tarp, H. Bengtsson, M. Mørup. "Short Term Blood Glucose Prediction Based on Continuous Glucose Monitoring Data." in IEEE EMBS 2020. **Planned Submission: 1st Jan 2020**
[14] J. Armenteros*, **A. Johansen***, O. Winther, H. Nielsen (*equal contribution). "Language modeling for biological sequences — curated datasets and baselines." in Bioinformatics 2020. **Planned Submission: Jan 2020**
[15] H. Martiny, J. Armenteros, **A. Johansen**, J. Salomon, H. Nielsen. "Predicting recombinant gene expression with deep learning techniques." in Biotechnology 2020. **Planned Submission: Jan 2020**