*engineering proceedings*

MDPI

*Proceeding Paper*

# Peptide Sequencing Using Neural Machine Translation Based on Sequence-2-Sequence Architecture and Long- Short-Term Memory Networks †

**Sobhan Naderian*, Preslav Aleksandrov, Naveen Kumar and Vihar Georgiev**

DeepNano Group, University of Glasgow, Glasgow G12 8QQ, UK; sobhan.naderian@glasgow.ac.uk (S.N.); email2@email.com (P.A.); email3@email.com (N.K.); Vihar.Georgiev@glasgow.ac.uk (V.G.)

* Correspondence: sobhan.naderian@glasgow.ac.uk

† Presented at The 11th International Electronic Conference on Sensors and Applications (ECSA-11), 26–28 November 2024; Available online: https://sciforum.net/event/ecsa-11.

**Abstract:** Mass spectrometry is the most reliable and accurate approach for analyzing a complex biological sample and identifying its protein content, which is time-consuming and reasonably expensive. One possible option to overcome such limitations is to use potentiometric sensors based on transistors. However, for such technology to work, a protein database that contains information on billions of small peptides and amino acids (AA) is required. The only practical way to build such a database is to use machine learning and this study is going to show the initial steps towards achieving this aim. This study sheds light on the possibility of a new approach for peptide sequencing combining analytical simulations with Large Language Models (LLM) based on Sequence-2-Sequence (Seq-2-Seq) architecture built by Long Short-Term Memory (LSTM) networks. 11,573 tokenized data points (voltage and capacitance cross-over points) with a vocabulary size of 504 are fed to the model, 80% of data is used for training and validation, and 20% is used for testing. The model is tested on unseen data and the accuracy during the test is 71.74%, which is significant if compared to expensive and time-consuming conventional methods, i.e., spectrometry. In conclusion, the output results of this study show that the proposed Seq-2-Seq LLM architecture could be used to build a material database for a potentiometric sensor to replace the mass spectrometry method.

**Keywords:** neural machine translation; large language models; peptide sequencing; amino acids; long-short-term memory

**Citation:** Naderian, S.; Aleksandrov, P.; Kumar, N.; Georgiev, V. Peptide Sequencing Using Neural Machine Translation Based on Sequence-2-Sequence Architecture and Long-Short-Term Memory Networks. *Eng. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 26 November 2024



## 1. Introduction

Currently, mass spectrometry is considered the most reliable and accurate method for analyzing complex biological samples and identifying their protein content. However, this technique is both time-consuming and expensive. One potential solution to these limitations is the use of potentiometric sensors based on transistors [1]. For such technology to be effective, it would require a comprehensive protein database containing information on billions of small peptides and amino acids (AA). The most practical way to construct this extensive database is by employing machine learning, and this study explores the initial steps toward that goal.

This study tries to shed light on the possibility of a novel peptide sequencing method that integrates analytical simulations with Large Language Models (LLMs) based on a Sequence-to-Sequence (Seq-2-Seq) architecture built using Long Short-Term Memory (LSTM) networks [2,3]. The section begins by describing how to measure the potential and capacitance of peptides, and how to create a database for each peptide, including the corresponding zero cross-over points for potential and capacitance [4]. Next, it provides

a detailed overview of the LSTM neural networks and the Seq-2-Seq LLM architecture used in the proposed model. Following this, the paper presents the simulation results of the model, and finally, it concludes with a discussion of future directions for research.

## 2. Peptide Potential and Capacitance Measurement

Figure 1 shows the calculated potential and capacitance for two oligopeptides made of four AAs, using the analytical approach based on the Gouy-Chapman-Stern (GCS) and site-binding models [4]. The only difference between DYKD and DYND is the presence of mutation (change of AA) at the third position where K is replaced with N. The information presented in Figure 2 shows the cross-over points of the 2nd derivative of the surface potential ($d^2\Psi_0/dpH^2$) [4].

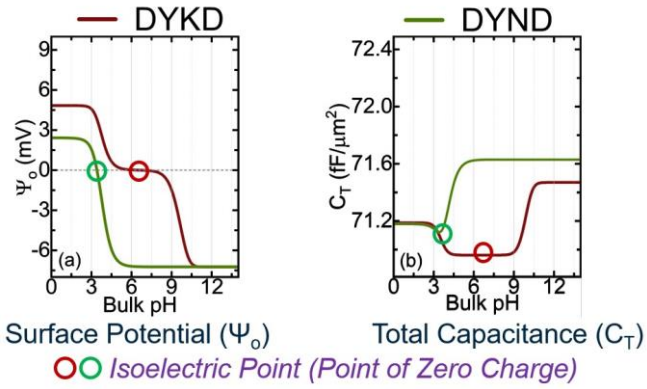


**Figure 1.** Analytical results for the calculated surface potential ($\Psi_0$) and total system capacitance ($C_T$) for two oligopeptides, DYKD (red line) and DYND (green line). The circles are the isoelectric points that can be compared with experimental values [4].

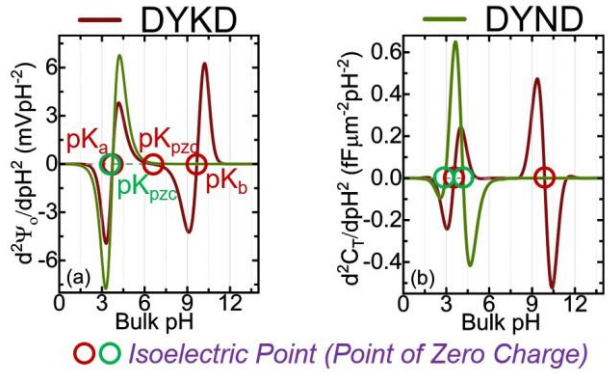


**Figure 2.** Analytically calculated 2nd derivative of the surface potential ($d^2\Psi_0/dpH^2$) and total capacitance ($d^2C_T/dpH^2$) as a function of pH. $pK_a$ and $pK_b$ are the dissociation constants and $pK_{pzc}$ is point of zero charges for both proteins. The values can be compared with experiments [4].

In this figure the curves cross with the zero dashed line, they represent the $pK_a$, $pK_b$, and $pK_{pzc}$ for each oligopeptide. All potential and capacitance curves have unique profiles [4].

## 3. LSTM

The sequential model for LSTM has been demonstrated in Figure 3, sliding on the predefined sequence of input data to generate an output sequence of data points or output time series. Each cell is fed by the sequence of input time-series data points [5]. Upcoming output data points and the cell's output are concatenated together to generate new input data points for the next data point. This repetitive procedure has been carried out to cover

whole data points. F. Gers introduced the LSTM network in 1999 which is a new type of RNN that consists of 4 main parts namely: input gate, input candidate gate, output gate, and forget gate [6]. Forget gate plays a key role in LSTM to forget former non-important cell's state and remember crucial state as expressed:

$$f_t = \sigma\left(W_f.[h_{t-1}, x_t] + b_f\right) \quad (1)$$

the $\sigma$ or sigmoid is an activation function widely used for adding nonlinear features to model as follows:

$$\sigma(X) = \frac{1}{1 - e^{-x}} \quad (2)$$

to generate a new cell state $C_t'$, the input gate and input candidate gate are activated simultaneously. This process is repeated across the entire data sequence [5]. The input gate utilizes a sigmoid function, while the input candidate gate applies a hyperbolic tangent function to compute the new cell state. Together, these gates work to update $C_t'$ as follows:

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_t) \quad (3)$$

$$C_t' = \tanh(W_c.[h_{t-1}, x_t] + b_c) \quad (4)$$

*tanh* function is a hyperbolic tangent function that maps input to a continuous number between −1 and 1.

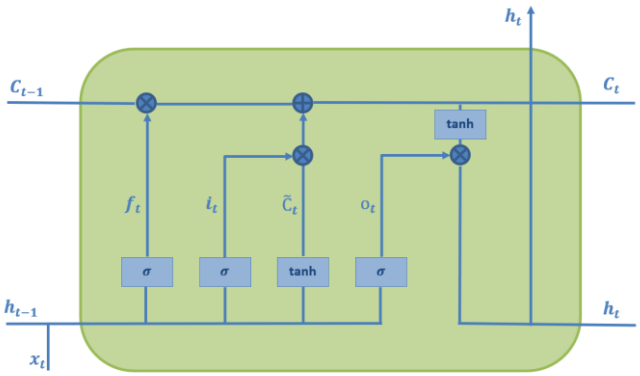


**Figure 3.** Internal structure of the LSTM cell.

## 4. Sequence-2-Sequence Architecture

Figure 4 represents the overall data, analyzing a pipeline we considered in our proposed approach used to translate potential and capacitance measurements into the equivalent protein sequence. First, tokenization converts measurements and proteins into tokens to create a vocabulary database and then feeds tokenized data to the proposed LLM model to train, validate and test the proposed approach.

The architecture of the proposed approach has been presented in Figure 5, which shows the details of layers considered in the Seq-2-Seq approach. The first layer is an embedding layer, a type of hidden layer that takes high-dimensional input data and projects it into a lower-dimensional space to allow the network to identify the relationships between the inputs better and process the data more efficiently. This layer is connected to the two LSTM layers. A dropout layer is attached to the second LSTM layer to prevent overfitting. For the last layer, a dense layer with a *softmax* activation function is considered to generate meaningful outputs.

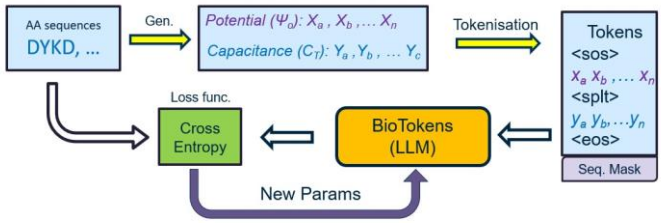


**Figure 4.** Flowchart of the data pipeline for the proposed machine learning architecture. Each protein is represented with potential and capacitance curves, and they are converted to tokens that are fed to the large language model.

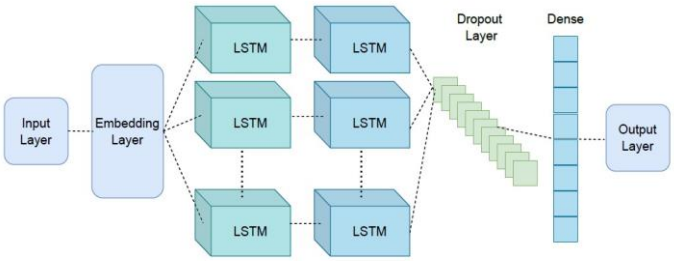


**Figure 5.** Proposed Seq-2-Seq model architecture for neural machine translation with LSTM layers.

## 5. Simulation Results

The configuration and parameters of the proposed architecture are presented in Table 1 which shows several parameters for each layer and a total number of trainable parameters. In each iteration all the trainable parameters, i.e., 4,468,247 should be trained and this process will end when a specified number of iterations passes.

**Table 1.** Network parameters of proposed Seq-2-Seq architecture for neural machine translation and LSTM network approach.

| Layer (Type) | Output Shape | #Param |
|---|---|---|
| embedding (Embedding) | (8, 512) | 258,048 |
| lstm (LSTM) | (512) | 2,099,200 |
| repeat_vector (RepeatVector) | (8, 512) | 0 |
| lstm_1 (LSTM) | (8, 512) | 2,099,200 |
| dropout (Dropout) | (8, 512) | 0 |
| dense (Dense) | (8, 23) | 11,799 |
| Total params: 4,468,247 | | |

11,573 tokenized data points (voltage and capacitance cross-over points) with a vocabulary size of 504 are fed to the first layer of the model, 80% of the data is used for training and validation, and 20% is used for testing. For the optimizer, RMS is considered with sparse categorical cross entropy function as a loss function to train and validate the model for 200 epochs, using a learning rate of 0.0001. The model is trained for 200 epochs, and loss values during each epoch for training and validation are measured and presented in Figure 6. The model is tested on unseen data, and the accuracy during the test is 71.74%, which is significant if compared to expensive and time-consuming conventional methods, i.e., spectrometry.

Table 2 compares the proposed model output and the actual output. The first column represents actual output, and the second column represents predicted output. From the data, the model output is accurate for most of the sequences of AA, specifically for

proteins with 2 and 3 AA. It is visible that for proteins with 3 amino acids, the model predicts them accurately in a different order. In Figure 7 we have compared the output data from ML and analytical solution for the DYND oligopeptide. Both curves are identical, and the models can reproduce the fingerprints for this protein. However, the ML model can be improved further by considering not only the zero cross-over points but also the magnitude of the peaks and the valleys and the slope of the curves. Indeed, these improvements are currently being implemented in a new version of our ML model.

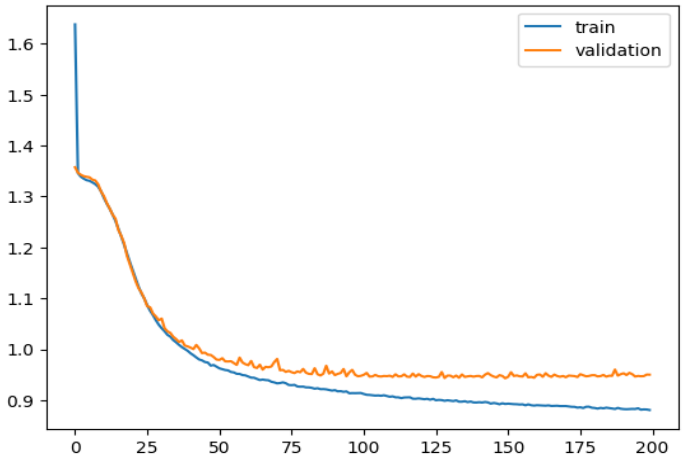


**Figure 6.** Train and validation loss values per each epoch step. Both curves show the expected saturated behaviour.

**Table 2.** Comparison of the actual and predicted values for amino acid (AA) content of some examples of protein sequences.

| Number | Actual | Predicted |
|---|---|---|
| 1 | HR | HR |
| 2 | RCU | CRU |
| 3 | HK | HK |
| 4 | DP | DP |
| 5 | EF | EF |
| 6 | CS | CS |
| 7 | EGP | EP |
| 8 | A E I | EI |
| 9 | KEM | EKM |
| 10 | H E C N | EH |

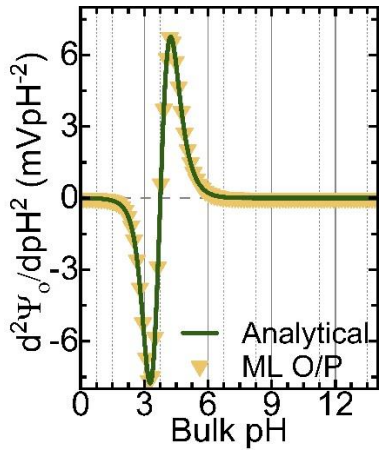


**Figure 7.** The 2nd derivative of the surface potential ($d^2\Psi_0/dpH^2$) for the DYND oligopeptide was calculated analytically (Green Solid Line) and from the Seq−2−Seq LLM model (Yellow Symbol−angle).

## 6. Conclusions

In this study, a novel method for peptide sequencing using neural machine translation has been proposed which is based on seq-2-seq LLM. To this end seq-2-seq architecture based on LSTM neural networks was implemented to translate voltage and capacitance measurements into amino acid combinations. In this study proteins with 2–4 amino acids were considered. 11,573 tokenized data points (voltage and capacitance cross-over points) with a vocabulary size of 504 were fed to the model, 80% of the data was used for training and validation, and 20% was used for testing. The model was tested on unseen data and the accuracy during the test was 71.74%, which is significant if compared to expensive and time-consuming conventional methods, i.e., spectrometry.

In conclusion, the output results of this study show that the proposed Seq-2-Seq LLM architecture could be used to build a material database for a potentiometric sensor to replace the mass spectrometry method. As a future work consideration of attention-based LLM would be a good option for improving the accuracy of the translation.

**Author Contributions:**

**Funding:**

**Institutional Review Board Statement:**

**Informed Consent Statement:**

**Data Availability Statement:**

**Conflicts of Interest:**

## References

1. Palacio Lozano, D.C.; Thomas, M.J.; Jones, H.E.; Barrow, M.P. Petroleomics: Tools, Challenges, and Developments. *Annu. Rev. Anal. Chem.* **2020**, *13*, 405-430.
2. Yang, K.L.; Yu, F.; Teo, G.C.; Li, K.; Demichev, V.; Ralser, M.; Nesvizhskii, A.I. MSBooster: Improving peptide identification rates using deep learning-based features. *Nat. Commun.* **2023**, *14*, 4539.
3. Yilmaz, M.; Fondrie, W.E.; Bittremieux, W.; Melendez, C.F.; Nelson, R.; Ananth, V.; Oh, S.; Noble, W.S. Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nat. Commun.* **2024**, *15*, 6427.
4. Kumar, N.; Aleksandrov, P.; Gao, Y.; Macdonald, C.; García, C.P.; Georgiev, V. Combinations of Analytical and Machine Learning Methods in a Single Simulation Framework for Amphoteric Molecules Detection. *IEEE Sens. Lett.* **2024**, *8*, 1–4.
5. Naderian, S. Machine learning approach for non-intrusive load monitoring in smart grids: New deep learning method based on long short-term memory and convolutional neural networks. In Proceedings of the 2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), Behshahr, Iran, 28–29 December 2022.

6. Petersen, N.C.; Rodrigues, F.; Pereira, F.C. Multi-output bus travel time prediction with convolutional lstm neural network. *Expert Syst. Appl.* **2019**, *120*, 426–435.