PREDICTING AND ANALYZING RNA AND PROTEIN MODIFICATIONS BY COMBINING DEEP PROTEIN LANGUAGE MODELS WITH TRANSFORMERS

A Thesis

by

Necla Nisa Soylu

Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in the
Department of Artificial Intelligence

Özyeğin University
January 2024

PREDICTING AND ANALYZING RNA AND PROTEIN MODIFICATIONS BY COMBINING DEEP PROTEIN LANGUAGE MODELS WITH TRANSFORMERS

Approved by:

Assistant Professor Emre Sefer
Dept. of Computer Science
Özyeğin University

Professor Olcay Taner Yıldız
Dept. of Computer Science
Özyeğin University

Associate Professor Ercüment Çiçek
Dept. of Computer Science
Bilkent University

Date Approved: December 22, 2023

To my family ...

ABSTRACT

Recent work on language models has resulted in state-of-the-art performance on various language tasks. Among these, Bidirectional Encoder Representations from Transformers (BERT) has focused on contextualizing word embeddings to extract the context and semantics of the words. Besides, their protein-specific versions such as ProtBERT generated dynamic protein sequence embeddings which resulted in better performance for several bioinformatics tasks. On the other hand, Post-transcriptional 2'-O-methylation (Nm) RNA modification and a number of different protein post-translational modifications are prominent in cellular tasks and related to a number of diseases. The existing high-throughput experimental techniques take longer time to detect these modifications, and costly in exploring these functional processes. Here, to deeply understand the associated biological processes faster, we come up with two efficient methods: the first one is BERT2OME to infer 2'-O-methylation RNA modification sites from RNA sequences and the second one is DeepPTM to predict protein post-translational modification (PTM) sites from protein sequences more efficiently.

BERT2OME combines BERT-based model with convolutional neural networks (CNN) to infer the relationship between the modification sites and RNA sequence content. Unlike the methods proposed so far, BERT2OME assumes each given RNA sequence as a text and focuses on improving the modification prediction performance by integrating the pre-trained deep learning-based language model BERT. Additionally, our transformer-based approach could infer modification sites across multiple species. According to 5-fold cross-validation, human and mouse accuracies were 99.15% and 94.35% respectively. Similarly, ROC AUC scores were 0.99 and 0.94 for the same species. Detailed results show that BERT2OME reduces the time consumed in biological experiments and outperforms the existing approaches across different datasets and species over multiple metrics. Additionally, deep learning approaches such as 2D CNNs are more promising in learning BERT attributes than more conventional machine learning methods.

Different than the current methods, DeepPTM enhances the modification prediction performance by integrating specialized ProtBERT-based protein embeddings with attention-based vision transformers (ViT), and reveals the associations between different modification types and protein sequence content. Additionally, it can infer several different modifications over different species. Human and mouse ROC AUCs for predicting Succinylation modifications were 0.988 and 0.965 respectively, once 10-fold cross-validation is applied. Similarly, we have obtained 0.982, 0.955, and 0.953 ROC AUC scores on inferring ubiquitination, crotonylation, and glycation sites respectively. According to detailed computational experiments, DeepPTM lessens the time spent in laboratory experiments while outperforming the competing methods as well as baselines on inferring all 4 modification sites. In our case, attention-based deep learning methods such as vision transformers look more favorable to learn from ProtBERT features than more traditional deep learning and machine learning techniques. Additionally, the protein-specific ProtBERT model is more effective than the original BERT embeddings for PTM prediction tasks.

ÖZETÇE

Dil modelleri üzerine yapılan son çalışmalar, çeşitli dil problemlerinde son derece iyi performans gösterdi. Örneğin, Transformatörler'de Çift Yönlü Kodlayıcı Gösterimleri (BERT), kelimelerin bağlamını ve anlamını çıkarma konusuna odaklanmıştır. ProtBERT gibi proteine özgü versiyonlar da, çeşitli biyoinformatik çalışmalarında çok iyi sonuçlar elde etmiştir. Bu çalışmaların yanısıra, transkripsiyon sonrası 2'-O-metilasyon (Nm) RNA modifikasyonu ve bir dizi farklı protein translasyon sonrası modifikasyonu, sadece hücresel görevlerde öne çıkmakla kalmayıp, canlılarda meydana gelen pekçok hastalıkla ilişkilendirilmiştir. Mevcut yüksek verimli deneysel tekniklerin bu değişiklikleri tespit etmesi hem uzun zaman almakta hem de çok fazla maaliyete gerektirmektedir. Bu alanlarda yapılacak ilgili biyolojik süreçleri daha hızlı hale getirmek ve ilgili konularda daha iyi sonuçlara ulaşmak için iki etkili model tasarladık; ilk olarak, RNA dizilerinden 2'-O-metilasyon RNA modifikasyon bölgelerini çıkarmak için kullanılacak olan BERT2OME ve ikinci olarak da protein dizilerinden protein translasyon sonrası modifikasyon (PTM) bölgelerini daha verimli bir şekilde tahmin etmek için DeepPTM'dir.

BERT2OME, modifikasyon bölgeleri ile RNA dizisi içeriği arasındaki ilişkiyi ortaya çıkarmak için BERT tabanlı modeli evrişimli sinir ağlarıyla (CNN) birleştirir. Şu ana kadar önerilen yöntemlerden farklı olarak BERT2OME, verilen her RNA dizisini bir metin olarak kabul eder ve önceden eğitilmiş derin öğrenme tabanlı dil modeli BERT'i entegre ederek modifikasyon tahmin performansını iyileştirmeye odaklanır. Transformatör tabanlı bu yaklaşımımız, birden fazla türün modifikasyon bölgelerini de ortaya çıkarabilmektedir. 5 katmanlı çapraz doğrulamaya göre insan ve farelerde doğru tahminleme başarısı sırasıyla 99.15% ve 94.35% olarak ölçüldü. Benzer şekilde ROC AUC skorları da aynı tür için 0.99 ve 0.94 olarak bulundu. Ayrıntılı sonuçlar, BERT2OME'nin biyolojik deneylerde harcanan süreyi azalttığını ve birden fazla ölçüm üzerinden farklı veri kümeleri ve türler genelinde mevcut yaklaşımlardan daha iyi performans gösterdiğini kanıtlamaktadır. Ek olarak, 2 boyutlu CNN'ler gibi derin öğrenme yaklaşımları, BERT özelliklerinin öğrenilmesinde daha geleneksel makine öğrenme yöntemlerine göre daha umut vericidir.

Bir diğer modelimiz DeepPTM, mevcut yöntemlerden farklı olarak, ProtBERT tabanlı protein yerleştirmelerini dikkat tabanlı görüntü transformatörleri (ViT) ile entegre ederek modifikasyon tahmin performansını arttırıp, farklı modifikasyon türleri ile protein dizisi içeriği arasındaki ilişkileri ortaya çıkarmıştır. Süksinilasyon modifikasyonlarını tahmin etmek için insan ve fare ROC AUC'leri, 10 katmanlı çapraz doğrulama uygulandığında sırasıyla 0.988 ve 0.965 sonuçlarını vermiştir. Benzer şekilde, ubikuitinasyon, krotonilasyon ve glikasyon bölgelerinin çıkarılmasında sırasıyla 0.982, 0.955 ve 0.953 ROC AUC skorları elde edilmiştir. DeepPTM modelinin geliştirilmesinde, görüntü transformatörleri gibi dikkat odaklı derin öğrenme yöntemleri, ProtBERT özelliklerini öğrenme konusunda daha etkili olmuş ve daha geleneksel derin öğrenme ile makine öğrenme tekniklerine göre üstün bir performans sergilemiştir. Ek olarak proteine özgü ProtBERT modeli, PTM tahminlemesinde BERT modelinden daha etkili olmuştur.

ACKNOWLEDGEMENTS

Firstly, I want to extend my deepest appreciation to my advisor, Assistant Prof. Emre Sefer. His dedication and tireless efforts have been the cornerstone of my successful academic life.

I am sincerely grateful to my educators, especially Professor Reha Civanlar, Professor Olcay Taner Yıldız, Professor Ethem Alpaydın, Professor Hasan Sözer, Assistant Professor Reyhan Aydoğan, Assistant Professor İsmail Arı, Professor Okan Örsan Özener, Assistant Professor Erinç Albey, and Assistant Professor Sedat Özer. You have played a big role in shaping my academic path, and it is an honor to have been your student.

I want to express my sincere gratitude to my family. Their love and belief in my abilities have strengthened me and ensured that I never give up.

Our DeepPTM: DeepPTM Protein Post-translational Modification Prediction from Protein Sequences by Combining Deep Protein Language Model with Vision Transformers research was funded by TUBITAK (Scientific and Technological Research Council of Turkey) 1002 Project with grant number 122E589. I want to thank TUBITAK for their cooperation and support.

I have always been a hardworking one full of enthusiasm. Discovering the underlying logic of the subject I am working on, is very important for me. I have always believed that our true mentor in life is science. I hope that my academic achievements and determination will continue to grow in the upcoming academic periods and I will achieve many more successes.

Contents

List of Tables

List of Figures

Chapter I

INTRODUCTION

This section emphasizes the significance of Post-transcriptional 2'-O-methylation (Nm) RNA modification and protein posttranslational modifications for species. It discusses the limitations of current experimental methods, highlights the crucial need for more efficient approaches, and introduces our novel models designed to address these challenges.

1.1 Significance of 2'-O-Methylation in Post-Transcriptional RNA Modifications Across Species

There are almost 160 types of post-transcriptional RNA modifications. Some common examples are 2'-O-methylation, m^1a, m^6a [2]. These RNA modifications are important in different cellular tasks and related to a number of diseases [2, 3, 4, 5, 6, 7]. Among these post-transcriptional RNA modifications, 2'-O-methylation is carried out by 2'-O-methyltransferase enzyme where the enyzme replaces hyrogen on the 2'-hydroxyl via the methyl group [8]. 2'-O-methylation (Nm) is observed in different RNAs such as mRNA, tRNA, miRNA [9, 10, 11] as well as across a number of species such as homo sapiens and mus musculus [12]. Nm modification is important in different biological mechanisms [13, 14, 15] including the regulation of gene expression, and its effect changes depending on the RNA type [12]. For instance, in messenger RNAs, 2'-O-methylation in ribosome is impactful in distinguishing among non-self and self mRNA [16, 17]. On the other hand, for ribosomal RNAs, ribosomal function and associated shape are affected by the degree of Nm modification's enrichment around the ribosome. Additionally, a number of diseases such as lung adenocarcinoma, hepatocellular carcinoma, congenital muscular dystrophy have been found to be correlated with 2'-O-methylation [18]. A number of experimental approaches have been developed to infer additional Nm modifications and relatedly to understand further biological functions. Some well-known examples are PCR-based approaches, reverse transcription-based approaches [19], and RNaseH-based approaches [20]. However, taking remarkably long time is a common disadvantage of all those experimental techniques. In the future, the number of available RNA sequence datasets for multiple species will grow since the existing sequencing technologies will continue to be applied and newer sequencing technologies will be developed. As a result, it is crucial to come up with efficient and robust methods to infer 2'-O-methylation modification locations for given sequences across different species.

1.2 Previous Work Related to 2'-O-Methylation in Post-Transcriptional RNA Modifications: Advances and Limitations

Several computational tools have been implemented to predict RNA 2'-O-methylation modification sites from RNA sequences without carrying out the costly biological experiments. [21] have utilized Support Vector Machine (SVM) for prediction site identification using nucleotide composition attribute encoding and nucleotide chemical characteristics. Their work constructs the prediction method only on human data, without showing Nm site classification results for remaining species. [22] have come up with iRNA-2OM, a SVM-based classifier based on sequences to predict Nm modification sites only for human. They have applied the attribute selection method to obtain optimal attributes for classification while mainly fusing nucleotide composition and chemical attributes. Recently, [23] have developed NmRF to predict RNA Nm modification sites across multiple species including human, yeast, and mouse. NmRF uses Light Gradient Boosting Machines integrated with incremental feature selection, and then predicts the Nm modification sites. Additionally, different than the aforementioned machine-learning based classifiers, deep learning-based methods have also been built to predict Nm modification sites such as iRNA-PseKNC (2-methyl) [24] and DeepOMe [25]. These two studies also focus on predicting the modification sites only for human. DeepOMe has come up with a hybrid approach combining Bidirectional Long Short-term Memory (BiLSTM) with CNNs. More recently, [26] has developed NmSEER2.0 to predict the modification sites over various human genomes. NmSEER2.0 uses random forest (RF) while using a mixture of encoding schemes such as K-nucleotide frequency encoding, position-specific dinucleotide sequence profiles, and one-hot encoding. Even though NmSEER2.0 performs reasonably accurately, it focuses on human to infer RNA Nm modification positions. Common to all these existing methods, they do not integrate contextual embeddings such as BERT into their 2'-O-methylation modification site prediction framework. As a result, they do not fully utilize the knowledge in transcriptome for Nm site prediction. Additionally, most of them are based on traditional machine learning (ML) techniques which have been significantly outperformed by deeper approaches across many tasks in bioinformatics [27, 28, 29, 30].

[31] has recently applied deep-learning based techniques to better represent contextualized sequences in m^6A modification prediction. Our work is different than their study in four ways: 1- [31] focuses on predicting RNA N6-methyladenosine sites whereas we focus on predicting 2'-O-methylation modification sites, 2- We have applied BERT to extract embedding-based contextualized feature encodings for prediction whereas they utilize ELMo, 3- We evaluate the performance of our methods over both balanced and more realistic imbalanced datasets whereas they only focus on balanced datasets, 4- Their prediction model following ELMo embeddings is quite deep with many parameters (a combination of number of CNNs and BiLSTMs) which may overfit given their relatively smaller dataset. However, our prediction methods

are simpler deep-learning methods which does not overfit as can be seen in the results.

According to recent studies, deep learning-based methods have performed quite accurately in multiple tasks including sequential biological data such as antiviral peptide identification [32], inference of electron transport protein [33], and predicting protein annotations [34, 35, 36]. Additionally, one can consider genomic sequences such as RNA sequences as textual knowledge, so these sequences have quite in common closeness with human languages. As a result, more recent deeper Natural language processing (NLP) methods have attained favourable results while learning practical and handy attributes over textual biological datasets. In this case, we can extract and represent biological sequence attributes efficiently by integrating embedding methods into deep learning.

Transforming and inferring the contextual relations between RNA and amino acid sequences can be achieved by adopting the word embedding methods that are now frequently utilized in natural language processing tasks [37]. As an example, [38] has come up with GeneVec and ProtVec methods for gene and protein sequences respectively to extract and represent attributes. Similarly, fastText method for NLP has been used in representing various types of sequences such as promoter and DNA enhancer [39]. Additionally, [40] has focused on applying another NLP model ELMo in representation of protein sequences. Their results suggest the importance of transfer learning to infer knowledge from protein sequences. In their study, protein sequences have been represented as continuous vectors which define a novel sequence-specific language model that efficiently discovers the protein sequences biophysical attributes over untagged dataset such as UniRef50. [41] has used a CNN-BiLSTM model that incorporates deep learning methods in achieving better functionalities than more conventional models while identifying amino acid sequences contextual relations. Lastly, [42] has come up with transformer-based architecture using BERT to identify enhancers in DNA from DNA sequences.

Based on the research above, deep neural network method focusing on the embedding methods have a huge potential in interpreting RNA sequence knowledge to predict Nm modification sites. Additionally, most of the studies above integrate static word embeddings in various bioinformatics tasks by ignoring the context around each word. In static embedding, popular methods such as fastText or Word2Vec [43] obtain the identical embeddings for the identical words without considering the context. Once language models infer the extra information from the paragraph or sentence context, learned embeddings quality will increase.

1.3 Contextualized Deep Learning for Predicting RNA 2'-O-Methylation Modification Sites: Integrating BERT Embeddings and Convolutional Neural Networks

By using this non-static contextual embedding idea, we come up with BERT2OME to predict Nm modification sites from RNA sequences that is based on Bidirectional Encoder Representations from Transformers (BERT) [44] language model which can learn the word representation from both right and left context. Generally, one can transfer the formerly extracted information from huge textual corpora to carry out tasks with a smaller dataset in the identical or different domains. In this work, we have focused on taking BERT's pretrained models two major advantages: 1- The significant amount of training data used in BERT training, 2- The domain change from natural language text to RNA language. In BERT2OME, we use RNA segments as input language to pretrain BERT models where numerical vectorial embeddings output by BERT capture the desired knowledge identical to meaning, syntax, and context of human language. Afterwards, we feed this high-dimensional embedding dataset into a well-studied deep neural network, two-dimensional convolutional neural network (CNN), for extracting additional attributes. To our best knowledge, this is the first study to integrate newer deep learning models such as BERT and CNN to infer RNA 2'-O-methylation modification sites over RNA sequences.

We experiment with various machine learning and deep learning methods ranging from SVM to two-dimensional convolutional neural networks to fully utilize BERT embeddings. Among them, we found two-dimensional convolutional neural network to perform the best so that's why BERT2OME includes two-dimensional convolutional neural network in its design. In our study, we have used RNA 2'-O-methylation modification site datasets for human, yeast, and mouse. Additionally, we also infer modification sites across multiple species: By training BERT2OME with one species and predict the modification sites on a different species. According to 5-fold cross-validation, human, mouse and yeast accuracies were 99.15%, 94.35%, 97.37% respectively. Similarly, ROC AUC scores were 0.999, 0.9375, and 0.9783 for the same species. The prediction performance was still reasonable while predicting across species: When we train BERT2OME with human RNA dataset, we can predict the modification sites for mouse and yeast with 88.6% and 53.7% accuracies respectively. On the other hand, when we construct our model with yeast, the prediction accuracy of the Nm modification sites was lower. Such lower performance can be due to us having a larger dataset for human which incorporate more detailed and accurate sequence knowledge. As a result, such larger dataset for human will be more effective in 2'-O-methylation modification site prediction across species. The detailed results show that BERT2OME reduces the time consumed in biological experiments and outperforms the existing approaches across different datasets and species over multiple metrics.

Overall, our contributions can be summarized as follows: 1- To our best knowledge, this is the first study to integrate state-of-the-art contextual BERT embeddings of RNA sequences into Nm modification site prediction, 2- Our proposed method BERT2OME outperforms the existing approaches, 3- We show that Nm modification sites can be predicted by training on one species and predicting on another one, and 4- The detailed experiments show that deep learning-based methods hold a better potential in utilizing BERT attributes compared to the remaining more traditional machine learning approaches.

1.4 Significance and Diversity of Protein Post-Translational Modifications (PTMs)

Protein post-translational modification (PTM) refers to chemical changes that happen to a protein after it's made in a cell. These alterations can impact the protein's structure, its location in the cell, and how it interacts with other molecules [45, 46]. PTMs include processes like adding small molecules (such as phosphates or acetyl groups) to specific amino acids, breaking the protein into smaller pieces, or attaching other proteins to it. The specific type and location of these modifications influence the protein's role and behavior within the cell. Over 400 types of protein post-translational modifications have been discovered. Examples of PTMs include phosphorylation, where a phosphate group is added to regulate protein activity; acetylation, influencing gene regulation and DNA binding; glycosylation, essential for protein folding and stability; methylation, impacting gene expression and interactions; ubiquitination, regulating protein degradation; sumoylation, involved in cellular processes; hydroxylation, affecting stability; and nitrosylation, participating in cell signaling [46]. Each modification contributes to the intricate regulatory network governing cellular functions, ensuring precise control over biological processes. These modifications are crucial for diverse tasks across the cells such as the development of organisms through gene regulation, cellular differentiation, and DNA replication in the nucleus [47, 48].

In certain cells, abnormal protein modifications can lead to various diseases and pathophysiological conditions. Examples include the transformation of normal tissues into cancerous tissues and developmental defects [49]. Modern proteomics experiments have revealed diverse categories of protein post-translational modifications, such as ubiquitin-like protein modifications, acylation, methylation, glycation, and hydroxylation, across human, mouse, and yeast [50, 51, 52]. Notably, ubiquitin-like protein modifications encompass types like ubiquitination and sumoylation [53, 54]. The acylation category includes various modifications like acetylation, succinylation, crotonylation, glutarylation, etc [55, 56]. These modifications play a role in the onset of diseases and offer insights into the complexity of cellular processes.

1.5 Traditional Methods and State-of-the-Art Language Models in Post-Translational Modifications (PTMs) Detection

Various biological experimental techniques have been developed for the detection of post-translational modifications (PTMs). Among those techniques, Liquid Chromatography (LC) [57] is utilized to separate and analyze peptides or proteins based on their properties, Radioactive Chemical Labeling [58] is employed to study the presence or absence of PTMs by incorporating radioactive labels into specific amino acids or molecules involved in modifications, providing insights into the modification patterns, Chromatin Immunoprecipitation (ChIP) [59] is valuable for studying histone modifications and their role in gene expression regulation. Antibodies against modified histones enable the precipitation of chromatin fragments containing these modifications for subsequent analysis. Among these antibody-based methods, Western Blotting [60] is employed for detecting specific PTMs by using antibodies that recognize the modified form of a protein. This technique facilitates the identification and quantification of modified proteins in a sample. On the other hand, Mass Spectrometry (MS) [61] is important in PTM analysis, providing detailed information about the mass changes associated with modifications. It enables the identification of various PTMs, contributing significantly to the understanding of protein modification dynamics. However, these PTM discovery methods have drawbacks such as being time-consuming, costly, labor-intensive, and they often involve the use of hazardous chemicals. Particularly for large-scale identification of PTMs, these limitations become significant. Therefore, there is a crucial need for the development of techniques that can swiftly and cost-effectively analyze protein sequences to identify potential PTMs without the challenges associated with traditional experimental approaches.

Word embedding methods that are developed to model context-dependent dynamics over textual problems in natural language processing could also be used to infer the contextual relations between biological sequences [62]. For instance, [63] proposed ProtVec to model protein sequence features. Additionally, [64] applied text model ELMo to represent protein sequences, showing the significance of transfer learning while inferring domain information from the sequences. [64] has come up with a unique language model that infers the specific features of protein sequences, by representing the sequences as continuous vectors. Another paper [65] has focused on the first universal deep-learning model of protein sequence and function. Recently, [66] has come up with ProtBERT, which modified BERT architecture and trained it for protein sequences. Similarly, [67] has proposed a BERT-based neural network in identifying 2'-O-methylation modification over RNA sequences.

1.6 Previous Work Related to Protein Post-Translational Modifications (PTMs)

Transformers have been applied to several different bioinformatics problems, ranging from protein function annotation [68], predicting protein properties [69], protein sequence profile prediction [70], pre-miRNA prediction [71] to DNA language's interpretability [62]. Similarly, [72, 73] understand the intersection of NLP and proteome bioinformatics, by focusing on the transformative impact of transformer-based NLP models emphasizing their potential to enhance the accuracy and efficiency of various tasks.

Recently, computational methods for inferring PTMs have taken considerable attention [74, 75]. These approaches may range from simpler methods to deep learning-based methods. For instance, GlyStruct [52] uses SVMs to predict glycated lysine residues using structural features of amino acid residues. Ubisite [76] incorporates two layered machine learning method to identify ubiquitin-conjugation sites. DeepTL-Ubi [77] comes up with a deep learning-based method to infer ubiquitination sites as well. LMSuccSite [78] and DeepSuccinylSite [79] propose LSTM and 2D CNN to infer succinylation sites. Similarly, Deep-Kcr [80] proposes a CNN-based approach to predict crotonylation sites. More recently proposed BERT-Kcr [81] and BERT-Kgly [82] methods take into account BERT embeddings as well. However, they do not include attention-based transformer structures such as ViT [83]. BERT-Kcr and BERT-Kgly combine BERT embeddings with BiLSTM and CNN networks to predict crotonylation and glycation modifications respectively. Both approaches use the general BERT model without training specifically for protein sequences.

The predictions made by all of these methods are limited to a certain type of PTM, whereas the proposed DeepPTM is capable of generating different types of PTMs for a given protein sequence. Recently, [84] proposed a conditional Wasserstein generative adversarial network-based method Multi-LyGAN to predict multiple types of modifications instead of a single PTM.

1.7 Integrating Vision Transformers and BERT for Efficient Protein Post-Translational Modification Prediction

By integrating transformer and dynamic contextual embedding concepts, we have developed DeepPTM to predict the presence of various post-translational modifications (PTMs) from protein sequences. DeepPTM is constructed based on protein language model ProtBERT (Protein BERT), which can be seen as an adaptation of Bidirectional Encoder Representations from Transformers (BERT) language model [85] to protein sequences. In this context, transferring knowledge gained from a vast textual or protein corpus to sequence-related problems on a smaller dataset, whether in an identical or nonidentical domain, is generally feasible. We leverage two key advantages of ProtBERT's pre-trained models: 1- ProtBERT's extensive training dataset contributes to its robust protein language understanding, and 2- ProtBERT's longer embedding vector for each amino acid models the sequence characteristics better. During DeepPTM's pretraining phase, protein sequences serve as the input language, and the vector embeddings generated by ProtBERT capture information similar to protein language. We feed this high-dimensional ProtBERT data into attention-based Vision Transformers (ViT) [83] to make predictions by learning from these features. To the best of our knowledge, this study represents the first attempt to combine Vision Transformers, featuring more complex attention mechanisms, with the pre-trained ProtBERT model to infer various types of PTMs in proteins.

After experimenting with various machine/deep learning approaches to maximize the utilization of BERT and protBERT embeddings, we identified Vision Transformer (ViT) [83] as the best performer, leading to its integration into the design of DeepPTM. In our study, we used 4 PTM datasets for human, mouse, and yeast. Moreover, DeepPTM infers PTMs across different species and across different modifications as well: 1- Training DeepPTM with a single species and predicting PTMs on a different species; 2- Training DeepPTM with one modification and predicting a different type of modification site. The accuracy of DeepPTM in predicting succinylation sites, measured by 10-fold cross-validation, with human, mouse, and yeast accuracies reaching 94.0%, 89.0%, and 90.7% respectively. The prediction performance is significantly higher than the competing approaches for glycation, crotonylation, and ubiquitination modifications as well. Our predictions maintain accuracy even across species and modifications. For instance, when DeepPTM is trained with the human succinylation dataset, it accurately predicts succinylation sites for mouse and yeast, achieving ROC AUCs of 0.955 and 0.954 respectively. Similarly, when DeepPTM is trained with known human succinylation predictions, it effectively predicts ubiquitination sites with a ROC AUC of 0.975. Through extensive computational experiments, DeepPTM not only reduces the time spent in laboratory experiments but also outperforms the competing methods across various species and modifications, as evidenced by multiple performance metrics.

In summary, our contributions can be outlined as follows: 1- To the best of our knowledge, this study represents the first attempt to integrate Vision Transformers (ViTs) with intricate attention mechanisms alongside pre-trained language models like BERT and ProtBERT to discern various types of PTMs within protein sequences. 2- Through extensive experiments, we demonstrate that advanced transformer-based models, such as ViTs, have the potential to effectively utilize BERT and more specialized ProtBERT features. 3- The proposed DeepPTM outperforms competing methods across four distinct types of modifications. 4- DeepPTM exhibits the capability to predict PTMs by training on one species and testing on another, as well as training on one type of PTM and testing on a different type. 5- We uncover consistent differential sequence motifs across various modifications and species.

Chapter II

MATERIALS AND METHODS

In this section, I will outline the process of collecting the datasets, discuss our approach for incorporating additional chemical properties, delve into the specifics of our models, and finally, explain the design of our models.

2.1 Dataset Collection for BERT2OME

We have obtained nucleotide sequences and the corresponding 2'-O-methylation modification sites from RMBase database [2] as well as from earlier papers [21, 22]. Even though our main focus is predicting Nm modification sites on human, we in general focus on 3 species in our analysis: Homo sapiens (human), Saccharomyces cerevisiae (yeast), Mus musculus (mouse). As part of data preparation, we have first downloaded the RNA sequences including the 2'-O-methylation modification sites across all species from the corresponding data sources. In RMBase database, the sequences are in their DNA structure so we transfer them to RNA sequences via converting T to U.

We have checked the sequence similarity of our all datasets in order to minimize the risk of model overfitting, and we found them to quite reasonable, around 32%. In this case, the similarity of the most similar sequences were about 41%. In order to remove the redundancy in our datasets, we have removed the sequences with more than 30% similarity. As a result of such removal, the number of samples has reduced to 538 from 590 and 770 from 828 for first and second human datasets respectively. As discussed in Results section, reducing the number of samples by similarity threshold have not significantly changed our results. Such performance stability after removal is mainly due to the complex nature of the deep learning models as well as still keeping major percentage of samples.

All the sequences are 41 nucleotide long. In total, the first homo sapiens dataset contains 215 positive, 215 negative instances for the training part and 46 positive, 114 negative instances for the testing part. We also utilize another Homo sapiens dataset from RMBase database [2]. We use only a portion of this second homo sapiens dataset in our experiments. We use 499 positive instances. The remaining negative samples (329 samples) were taken from the first homo sapiens dataset.

Besides, in order to get more balanced human datasets, we have used SMOTE [86]. SMOTE is one of the most commonly preferred oversampling method and is used for increasing the number of the minority class by creating synthetic records via linear interpolation. These virtual records are randomly generated by selecting the k-nearest neighbors for each sample in the minority class instead of just creating copies of the minority samples to increase their numbers. As a result of applying SMOTE to the first human dataset (Human^1), we got 329 positive and 329 negative samples in total. On the other hand, 499 positive and 499 negative labeled 41 nucleotide long RNA sequences were generated for the second human dataset (Human^2). S. cerevisiae dataset contains 89 positive samples, 189 negative samples, and M. musculus dataset contains 10 positive samples, 35 negative samples in total. Besides, the number of samples in S. cerevisiae and M. musculus datasets were not adequate to efficiently train BERT2OME as well as these datasets were imbalanced. As a result, we use SMOTE to further expand the datasets. The number of positive and negative instances in our datasets are summarized in Table 1.

Table 1: Data summary for analyzed species and datasets.

Species	2'-O-methylation Dataset	Positive	Negative
Homo sapiens^1 (Human^1)	[22]	261	329
Homo sapiens^2 (Human^2)	[2]	499	329
Saccharomyces cerevisiae	[21]	89	189
Mus musculus	[21]	10	35

Here, we consider RNA sequences of 41 nucleotides as single input. We have experimented with natural language processing model via considering each single nucleotide as a single word corresponding to a unigram model. As a result, next, we have inserted spaces between pair of nucleotides to establish a consecutive set of bases where each base represents a word in human language. As part of the BERT model training, special tokens were added, namely CLS and SEP, to the start and end of the sentences.

2.2 Additional Chemical Properties

We also integrate chemical properties of RNA sequence nucleotides into our prediction method BERT2OME, that are quite common to be used in predicting the RNA modification sites [87]. Basically, RNA sequences are made up of 4 types of bases: Adenine (A), Guanine (G), Cytosine (C), and Uracil (U). Each of these 4 nucleotides exhibits a unique chemical structure and has its own internal binding properties. As a result of such unique structure and binding properties, all these four types of bases possess different chemical characteristics. RNA modifications sequence attributes generally express nucleotide sequences via 3 different chemical structural qualities. These 3 chemical structural qualities are ring structures, hydrogen bonds, and functional groups. In this case, 4 nucleotide types are partitioned into 3 groups: 1- In terms of ring structures, C and U have a single ring in its structure whereas A and G have double ring in its structure. 2- In terms of hydrogen bonds, as part of the hybridization, A and U could form 2 bonds whereas G and C could form 3 bonds. 3- In terms of chemical functional groups, G and U carry out ketone bases wheres and A and C carry amino groups.

Let X_i, Y_i, and Z_i represent the ring structure, hydrogen bonds, and chemical functional groups for RNA sequence's nucleotide i (s_i) respectively. We characterize RNA sequences further by incorporating these nucleotide chemical properties.

Nucleotide chemical property calculation formula are as below. According to these chemical properties equations, base A is expressed as (1,1,1). The remaining bases C, G and U are expressed as (0,1,0), (1,0,0) and (0,0,1) respectively. As a result, we obtain 123×1 vector after applying such nucleotide chemistry property encoding.

$$x_i = \begin{cases} 1, & \text{if } s_i \in \{A, G\} \\ 0, & \text{if } s_i \in \{C, U\} \end{cases} \tag{1}$$

$$y_i = \begin{cases} 1, & \text{if } s_i \in \{A, C\} \\ 0, & \text{if } s_i \in \{G, U\} \end{cases} \tag{2}$$

$$z_i = \begin{cases} 1, & \text{if } s_i \in \{A, U\} \\ 0, & \text{if } s_i \in \{C, G\} \end{cases} \tag{3}$$

2.3 Bidirectional Encoder Representations from Transformers (BERT) for BERT2OME

The degree of associations between words and context over a specified window can be analyzed via word embedding techniques. A number of traditional embedding approaches are still widely-used, such as Glove [88], neural network with continuous bag-of-word architectures (Word2vec) [43] or skipgrams. Other than more traditional word embedding approaches discussed above, newer approaches such as BERT and XLNet [89] have tried to tackle the existing word representation models' single directional training problem. Such single directionality assumption already limits the possible architectures that are part of pretraining. [44] come up with BERT to pretrain the model by using unlabeled text in a bidirectional way via combination of forward and reverse contexts across all of its layers. As a result of such novel better training, BERT generates dynamic word embeddings meaning that the identical nucleotides or words across different sentence positions take non-identical continuous real-valued vectors. Due to several advantages of BERT, such as ability to generate contextual embeddings, BERT could consider word position and has a support for words that do not appear in the vocabulary. We assume that BERT will better and more efficiently process the latent knowledge in RNA sequences.

Pretrained BERT models bring the best performance on variety of natural language processing tasks without remarkable changes specific to task, such as inferring languages from text, generating automatic answers to questions, etc. In our case, we have used BERT in our 2'-O-methylation modification site prediction problem to explain and analyze RNA sequences information. BERT is mainly made up of 2 steps: 1- Pretraining step: BERT model is trained over unlabeled dataset for 2 tasks, namely predicting the next sentence and masked language modeling (MLM); 2- Finetuning step: Pretrained parameters are then used to initialize the model, and these parameters are finetuned for the considered specific task.

In our prediction problem, we performed the first BERT step via MLM by substituting 15% of all tokens with (MASK) token. Since (MASK) token is not contained in the finetuning part of BERT, we incorporate the subsequent rules not to come across mismatch between pretraining and finetuning: 1- The token is substituted with (MASK) in 80% of the scenarios, 2- The token is substituted with a random counterpart in 10% of the scenarios, 3- The token is kept as it is in 10% of the scenarios.

There are a number of available pretrained BERT models. For instance, BERT-large cased and uncased models need approximately 340,000,000 parameters and include 16 heads, 24 layers, and 1024 hidden units. Similarly, BERT-base cased and uncased models need approximately 110,000,000 parameters as part of the training, and include 12 heads, 12 layers, and 768 hidden units. In our study, we carried out experiments with BERT-base uncased model. As part of the training, we use one-hot encoding of RNA sequences as our input as seen in Figure 1. Once the training is finalized, we transform every RNA nucleotide into a contextualized 768 dimensional word embedding vector. As a result, we translate each 41 nucleotide length RNA sequence into a single vector made up of $n = 41$ vectors of 768 dimensions appended consecutively. There are 12 layers in our BERT model, and every layer in the pretrained model is an encoder where input from one encoder is the output from the previous encoder. By trying to incorporate the knowledge from the whole set of encoders, we added up the attribute vectors of all these 12 layers (of each having dimension $n = 41 \times 768$) and used the resulting vectors as the input of the two-dimensional convolutional neural network discussed below.

2.4 BERT2OME: BERT + 2D CNN for 2'-O-Methylation Modification Site Prediction

BERT2OME integrates 2D CNN model to learn from BERT embeddings, which categorizes RNA sequences into with or without 2'-O-methylation modification. CNN has been mainly used in a number of computer vision applications such as image segmentation, object detection, image classification. etc. CNN has also been applied to a number of bioinformatics problems including extracting the knowledge in dinucleotide one-hot encoder [90], topology structure, etc. As a result, the performance of BERT2OME will be improved when CNN extracts attributes from contextual word embedding vectors from BERT.

First, we create vector embeddings from input RNA sequences by using BERT. For each given RNA sequence, we have added each 41 nucleotide long nucleotide sequence CLS (added at the beginning of the RNA sequence) and SEP (added at the end of the RNA sequence) tokens, making them 43 nucleotide long. We have used the "bert-base-uncased" model while creating vector embeddings of RNA sequences which has a neural network structure with 12 layers with 768 hidden units. The first layer has consisted of the input embeddings, and the remaining layers represent the output of the model. After getting the vector embeddings by using BERT, each nucleotide was represented by a $33024 = (43 * 768)$ long vector. Besides, by using chemical properties strategy, $123 = (41 * 3)$ numerical values have also been generated, and we have appended this additional vector into the vector from BERT. By such appending, we have added chemical properties vector to that embedding vector without changing the overall structure of the model.

While designing the CNN part of BERT2OME, we have used 2 hidden layers in the feature extraction stage of the 2D CNN model. Since BERT model's output is a two dimensional matrix with size 43×768, learning and capturing such attributes over this matrix via 2D CNN looks quite rational. In general, CNN is composed of more than one layers, where each layer, with its particular function, is useful in transforming its input data into a better representation. CNN utilizes 4 layer types such as max pooling later, convolutional layer, ReLU layer, and fully-connected layer. In our case, BERT2OME focuses on learning spatial knowledge by convolutional operations, which is then used in predicting the modification sites over RNA sequences efficiently. The max pooling process was applied to each hidden layer. In addition, dropout is applied on each hidden layer to prevent overfitting and improve generalization error. We have used Rectified linear activation function (ReLU) as the activation function. The embeddings obtained after the feature extraction stage were given to the fully connected layer after flattening and softmax activation function was used afterwards. We have used categorical cross entropy as the loss function since it is a well-known loss function that is used in multi-class classification tasks. Adam optimizer were used for the optimization process. We have also designed 1D CNN model that is architecturally same as 2D model in terms of number of layers, activation function, loss function to make a fair comparison between these two dimensionally different CNN models. Unlike 1D CNN model, 2D CNN takes 2D matrix as input. One of the CNN dimensions represents each nucleotide with their starting and end tokens ([CLS] token, 41 nucleotide long RNA sequence and [SEP] token), while the other dimension is the embedding vector values generated from the BERT model. Instead of using only the vector values generated from the last layer in the neural network structure obtained from BERT model, we have realized that our prediction values were positively affected when we took the average of last 4 layers [42], so we have designed BERT2OME accordingly. Figure 1 summarizes BERT2OME architecture.

2.5 Dataset Collection for DeepPTM

Proteins consist of amino acids, and lysine is one of the 20 standard amino acids that make up the building blocks of proteins. Post-translational modifications over lysine involve chemical changes or additions that occur on lysine amino acid residues within a protein after its synthesis. Lysine residues are common targets for post-translational modifications, where different chemical groups or molecules can be added to or removed from lysine side chains. We collected protein sequences and corresponding post-translational modifications from the Compendium of Protein Lysine Modifications (CPLM) database [50], specifically from version 4.0. As detailed in the associated paper [50], the CPLM database serves as an integrated resource for a wide range of PTMs, encompasses a vast collection of experimentally identified modification events, including 592,606 PLM events on 463,156 unique lysine residues across 105,673 proteins, spanning 29 types of PLMs and 219 different species. CPLM covers a comprehensive array of PTMs, including well-characterized ones like acetylation and ubiquitination, as well as newer additions like succinylation and crotonylation. The database has been carefully compiled from identified substrates and sites obtained through experiments, serving as a valuable resource for comprehending the complex world of PTMs. In our experiments, we focused on four important post-translational

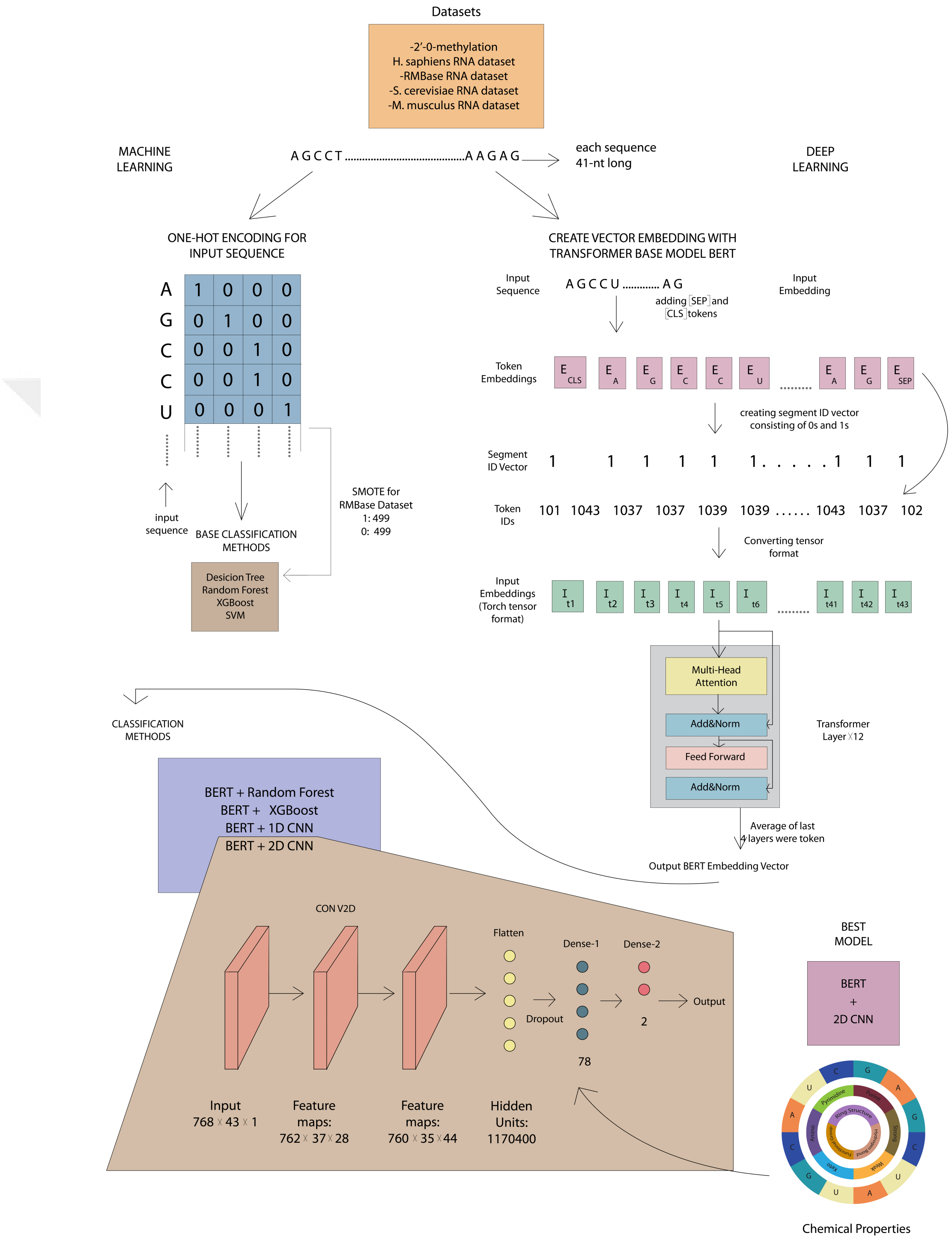


Figure 1: Whole representation for the base methods (left part of the diagram) and the novel part (right part of the diagram). The workflow of the proposed method BERT2OME. 4 different datasets with 3 different species were used. Vector embeddings were created using BERT model, then classification method 2D CNN was implemented to identify 2’-O-methylation sites in the given RNA sequences.

modifications over lysine: Succinylation, Glutarylation, Crotonylation, and Glycation. These PTM datasets vary in size as shown in Table 2. Although our primary focus is on predicting PTMs in Homo sapiens (human), we also tested DeepPTM on two additional species: Mus musculus (mouse) and Saccharomyces cerevisiae (yeast).

As part of the preprocessing step, we addressed redundancy in the datasets using the CD-HIT tool [91], a widely used clustering program known for its ability to remove redundancy in homologous sequences. We decided on a criterion for removal: sequences with over 40% similarity. This choice was made after exploring the impact of various similarity thresholds on the results. Proteins were transformed into peptides with a fixed length of 21. Such fixed length transformation is commonly used in similar research even though the lengths differ. For instance, LMSuccSite [78] and DeepSuccinylSite [79] use 33, whereas [82] uses 31. Our choice of 21 was made after exploring the impact of various peptide lengths on the results. A design was implemented where lysine is positioned at the central residue (indicating whether a modification occurs or not), with 10 residues present on both the upstream and downstream sides. Peptides located near the observed modification sites were considered a positive dataset, while 21-length sequences from other non-modified lysine residues constituted the negative dataset. Before training the method, redundant/duplicated items were eliminated.

Since the imbalance of a training dataset would cause prediction errors, we employed random under-sampling on our datasets, a technique commonly used in the literature for dataset balancing [78, 79]. If more negative samples exist than positive samples, we randomly choose sequences with the same number of positive samples from the non-redundant negative samples. Afterward, all our datasets contain the same number of negative and positive instances as seen in Table 2. For instance, the crotonylation dataset on homo sapiens contains 4122 positive and 4122 negative samples.

Table 2: Summary of datasets for tested protein post-translational modifications and species.

PTM	Species	Positive	Negative
Succinylation	Homo Sapiens	2014	2014
Succinylation	Mus musculus	2510	2510
Succinylation	S. cerevisiae	810	810
Ubiquitination	Homo Sapiens	2000	2000
Crotonylation	Homo Sapiens	4122	4122
Glycation	Homo Sapiens	3342	3342

2.6 Bidirectional Encoder Representations from Transformers (BERT) for DeepPTM

Word embedding methods can be utilized to analyze the strength of the relationship between words and their contexts within a designated window. In contrast to traditional static embedding methods such as Word2vec [43] and Glove [88], newer approaches like BERT [85] have addressed the unidirectional training problem of the current word embedding models. BERT utilizes unlabeled textual data in both directions over all its layers during its pretraining achieved through a mixture of reverse and forward contexts. Consequently, word embeddings generated by BERT are dynamic (contextual), meaning that the same words at different locations result in different real-valued continuous vectors. Pretrained BERT models perform great on some different language processing problems such as language inference from a textual dataset [92], automatic answer generation to a question [93], etc., without making significant modifications for a given task. Even though BERT is not trained specifically on protein sequences, in our baselines, we used BERT in PTM inference to better extract and model protein sequence knowledge.

Various pre-trained BERT models exist with different parameters, layers, heads, and hidden units. In our baseline experiments with BERT, we used the BERT-base-uncased model. After the training step, each amino acid was converted into a contextualized 768-dimensional embedding vector. In this case, every 21 length protein sequence is translated to a matrix consisting of $n = 21$ vectors of 768 dimension each. The pre-trained BERT model consists of twelve layers, with each layer functioning as an encoder. The output from one encoder serves as the input to the next. In Natural Language Processing (NLP), the practice of averaging the embeddings from the last layers of a BERT model has become an important strategy for its effectiveness in various applications [94, 95]. The last layers of BERT are thought to capture more abstract and semantic information about the input sequence, providing a richer representation. Since BERT is pre-trained with unsupervised learning on a large corpus, the features learned in the last layers are task-agnostic and suitable for a wide range of downstream tasks. Averaging over these layers facilitates transfer learning, allowing the model to leverage pre-trained knowledge for diverse applications without extensive fine-tuning.

This approach enhances robustness by reducing the impact of noise or idiosyncrasies in individual layers and contributes to consistent performance across different tasks. Moreover, averaging over a subset of layers strikes a balance between computational efficiency and model effectiveness. This strategy provides fine-tuning flexibility, allowing task-specific fine-tuning while simplifying the model's complexity. We integrated the information from all these encoders by averaging the embedding matrices of the last four layers (of each with $n = 21 \times 768$ dimensions). When used together with the vision transformer as a baseline, these averaged matrices then become input to the vision transformer.

2.7 ProtBERT: Protein BERT

[66] has come up with the protein language model ProtBERT, an adaptation of BERT to protein sequences, which is trained over UniRef100. The number of layers in ProtBERT has increased to 30, instead of 12 as in the original BERT architecture. Firstly, ProtBERT was trained for 300k steps on sequences with 512 maximum length.

Afterward, it was trained for another 100k steps on sequences with a length of a maximum length of 2k. With such training, ProtBERT can infer functional attributes over smaller sequences while running a larger batch size, which leads to training on longer sequences to become more efficient. In our case, we report the results by using only the last layer for ProtBERT, since there is no major performance difference between using the last 4 layers and using only the last layer for ProtBERT.

2.8 DeepPTM: ProtBERT + Vision Transformer (ViT) for Protein Post-translational Modification Prediction

DeepPTM integrates attention-based vision transformer (ViT) [83] while learning according to ProtBERT embeddings, which classify protein sequences with/without a given protein post-translational modification. Previously in computer vision, attention has been proposed together with CNNs, or they have replaced several CNN parts when the CNN structure is still kept intact. Different than these previous structures, ViT does not rely on CNNs anymore. Instead, it directly applies a solid transformer to a sequence of image patches (sub-matrices). Vision transformers have recently outperformed CNNs in multiple vision applications including image classification [96], object detection [97], image segmentation [98], etc. Even though CNNs have also been used in multiple different bioinformatics problems [99], ViTs have not been frequently used in bioinformatics except for a few problems such as X-ray image analysis [100]. Consequently, DeepPTM's performance will increase when ViT analyzes the features of ProtBERT's contextual embeddings.

Initially, ProtBERT generates the embeddings over input protein sequences. For each protein sequence, which is 21 amino acids long, we added CLS and SEP tokens (attached to the sequence's start/end respectively), extending the length of each sequence to 23. We obtained sequence embeddings using ProtBERT, featuring 30 layers with 1024 hidden units. Each sequence is represented by a 23 × 1024 matrix for ViT, following ProtBERT embeddings. Learning from embedding features over that 2D matrix by ViT seems quite promising.

The integration of Vision Transformer in our DeepPTM model represents a different approach aimed at enhancing the overall performance of protein sequence analysis. While ProtBERT embeddings provide a strong architecture by capturing the complex contextual relationships within amino acid sequences, ViT introduces a novel dimension by directly applying a transformer to the sequence of image patches. This change from the usual CNNs is important because ViTs have shown better results in many computer vision tasks like image classification, object detection, and image segmentation. In the context of protein sequence analysis, ViT offers a unique perspective by treating the protein sequence as a visual input, allowing it to effectively capture spatial dependencies and patterns. We decided to combine ViT with ProtBERT embeddings since each of those models brings unique strengths: ProtBERT is great at understanding the protein sequence context, and ViT enhances this by providing spatial awareness, making the representation of protein sequences more comprehensive.

Figure 2 summarizes the architecture of DeepPTM. In contrast to the 2D input scenario, traditional transformers take a one-dimensional series of token embeddings as input. ViT, on the other hand, handles 2D input $x \in R^{H=23\times W=1024}$ by reshaping it into a series of two-dimensional flattened patches $x_p \in R^{N\times P^2}$. Here the original input has a resolution of (H, W), each patch has a resolution of (P, P), and there are $\frac{HW}{P^2}$ patches, equivalent to the effective input series length for the transformer. The ViT transformer utilizes a vector of length D across all layers. In this process, ViT maps the flattened patches to D dimensions through a trainable linear projection (Eq.4). The output of this patch flattening step is referred to as patch embeddings. This process is repeated L times, corresponding to the number of transformer layers. Similar to BERT's [CLS] token, ViT appends a trainable embedding to the series of embedded patches ($z_0^0 = \text{class}$), where the state of the embedded patches at the Transformer encoder's output (z_L^0) works as the representation of y (Eq. 7). A classification head, integrated into z_L^0 during pretraining and finetuning, consists of an MLP with a single latent layer at pretraining and one linear layer during finetuning. The softmax activation function is applied in the end to convert this problem into a binary classification task.

ViT adds position embeddings to patch embeddings to keep positional knowledge. ViT uses traditional one-dimensional position embeddings that are learnable instead of two-dimensional aware position embeddings as no significant performance differences have been observed between them. The combined embedding vector sequence then becomes the input to the transformer encoder [101]. The transformer encoder is composed of consecutive alternating multi-headed self-attention (MSA) and Multilayer Perceptron (MLP) layers. Before each of these layers, ViT applies a normalization layer (LN), and residual connections after each layer [83]. MLP models the nonlinearity by including 2 layers with a Gaussian Error Linear Unit (GELU) [102].

$$z_0 = [x_{\text{class}}; x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{\text{pos}}, \qquad E \in R^{P^2 \times D}, E_{\text{pos}} \in R^{(N+1) \times D} \tag{4}$$

$$z_l^{'} = MSA(LN(z_{l-1})) + z_{l-1}, \qquad l = 1, \ldots, L \tag{5}$$

$$z_l = MLP(LN(z_l^{'})) + z_l^{'}, \qquad l = 1, \ldots, L \tag{6}$$

$$y = LN(z_L^0) \tag{7}$$

In the design of DeepPTM, we conducted experiments with various parameters. Specifically, we explored different numbers of transformer layers (L), considering values of 1, 2, and 3. Additionally, we experimented with patch sizes (P), considering options of 6 and 8. The number of heads in the multi-head attention component was set to 12, while the projection dimension (D) was varied among 32, 64, and 128. The learning rate was adjusted within the range of $1e^{-3}$ to $1e^{-4}$. Throughout these experiments, the batch size remained constant at approximately 200, and the number

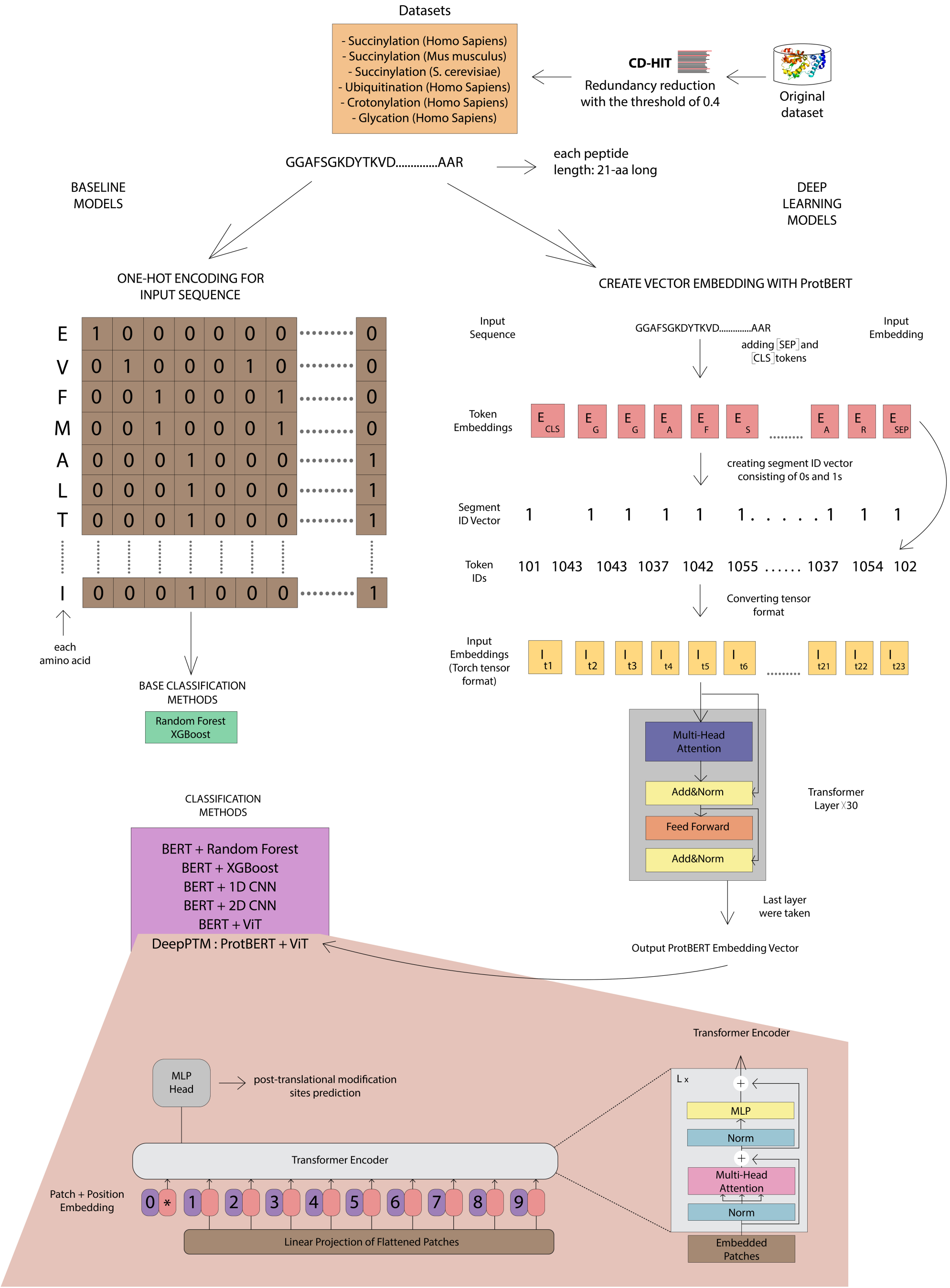

Datasets
- Succinylation (Homo Sapiens)
- Succinylation (Mus musculus)
- Succinylation (S. cerevisiae)
- Ubiquitination (Homo Sapiens)
- Crotonylation (Homo Sapiens)
- Glycation (Homo Sapiens)
CD-HIT
Redundancy reduction with the threshold of 0.4
Original dataset
GGAFSGKDYTKVD.............AAR
each peptide length: 21-aa long
BASELINE MODELS
DEEP LEARNING MODELS
ONE-HOT ENCODING FOR INPUT SEQUENCE
E 1 0 0 0 0 0 0 0
V 0 1 0 0 0 1 0 0
F 0 0 1 0 0 0 1 0
M 0 0 1 0 0 0 1 0
A 0 0 0 1 0 0 0 1
L 0 0 0 1 0 0 0 1
T 0 0 0 1 0 0 0 1
I 0 0 0 1 0 0 0 1
each amino acid
BASE CLASSIFICATION METHODS
Random Forest
XGBoost
CREATE VECTOR EMBEDDING WITH ProtBERT
Input Sequence
GGAFSGKDYTKVD.............AAR
adding SEP and CLS tokens
Input Embedding
Token Embeddings
E CLS
E G
E G
E A
E F
E S
....
E A
E R
E SEP
creating segment ID vector consisting of 0s and 1s
Segment ID Vector
1 1 1 1 1 1 . . . . . . 1 1 1
Token IDs
101 1043 1043 1037 1042 1055 ...... 1037 1054 102
Converting tensor format
Input Embeddings (Torch tensor format)
I t1
I t2
I t3
I t4
I t5
I t6
....
I t21
I t22
I t23
Multi-Head Attention
Add&Norm
Feed Forward
Add&Norm
Transformer Layer X30
Last layer were taken
Output ProtBERT Embedding Vector
CLASSIFICATION METHODS
BERT + Random Forest
BERT + XGBoost
BERT + 1D CNN
BERT + 2D CNN
BERT + ViT
DeepPTM : ProtBERT + ViT
MLP Head
post-translational modification sites prediction
Transformer Encoder
Patch + Position Embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L x
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

Figure 2: Representation of baseline approaches and the workflow of our contribution DeepPTM all together. Vector embeddings were generated via ProtBERT. Afterward, ViT, as a classification method, identifies PTMs over a given protein sequence embeddings.

of epochs was set to 500.

Chapter III

EXPERIMENTAL SETUP

In the upcoming section, I will provide a comprehensive overview of our initial setup before the development of our complex models. Detailed information will be provided about how we designed our baseline machine learning approaches to understand the extent of its contribution before implementing the complex deep learning models. Additionally, I will explain our process of hyperparameter tuning and optimization, for refining the models' performance. Finally, I will present a detailed analysis of our models' predictive capabilities, supported by various metrics.

3.1 Baseline Machine Learning Approaches for 2'-O-Methylation Modification

To train the baseline machine learning approaches, we have converted each nucleotide into numeric values by using one-hot encoding approach. In a given RNA sequence, A, G, C, U are mapped to [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1] respectively. In the end, we have obtained vectors with lengths 164, for each 41 nucleotide long RNA sequences. Similarly, our modification output labels were converted into ones and zeros. Sequences with 2'-O-methylation modifications are represented with 1s, non 2'-O-methylation modifications are represented with 0s.

We have experimented with 4 well-known machine learning (ML) classifiers as our baseline: Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost). Among these baselines, SVM is a well-studied ML algorithm applied to various bioinformatics problems [103]. While determining whether there exists a Nm modification in the given RNA series, we choose to evaluate the RNA series as text and use the linear kernel function with the regularization parameter 1 in sklearn library in our base SVM model. Such linear kernel is highly preferred in text classification problems.

Our baseline Decision Tree approach is also known as CART algorithm [104] which stands for Classification and Regression Tree. In our case, Decision Tree was used for solving the classification problem (Nm modification prediction) for given RNA sequences. We have preferred Gini coefficient after comparing the results, and splitting operation was achieved according to the best split. The remaining parameters were used as their default values in sklearn library. Another approach Random Forest [105] is an ensemble technique where we carry out a hyperparameter tuning to obtain optimal model parameters rather than using the default parameters. Lastly, we also use XGBoost [106] similarly by applying hyperparameter tuning to obtain optimal model parameters

*3.2 Hyperparameter Tuning and Optimization for B*ERT*2O*ME

Deep learning and machine learning approaches require hyperparameter optimization step to achieve the best performance results. In our case, we have focused on tuning the hyperparameters independently for BERT and 2D CNN parts. We have applied 5-fold cross-validation to optimize the hyperparameters. Initially, BERT part (embedding part) hyperparameters such as learning rate, training steps count, maximum sequence length, batch training size were tuned. Several hyperparameter combinations were considered to obtain the optimal one. Secondly, we optimize the hyperparameters for 2D CNN part. For instance, epoch number, neuron number, and batch size count were between 25 to 100 epochs, 50 to 200 neurons, and 10 to 30 batch sizes, respectively. We have also used dropout rate parameters to prevent the overfitting. We use Keras Tuner library to tune CNN hyperparameters.

*3.3 Evaluating the Performance of B*ERT*2O*ME

We have implemented BERT2OME in Python by using Keras deep learning library that uses TensorFlow as the backend. We have implemented the remaining machine learning models in Python. Lastly, we obtain the predictions for existing methods iRNA-2OM [22], NmSEER2.0 [26] from their web servers. We could not obtain the results for DeepOMe [25] from their web servers since they require 120 nucleotide length input. Therefore, we use the reported results in their paper when needed for comparison. We run our experiments on a personal laptop with Intel Core 2.80 GHz CPU and 8 Gb memory. On approximate, BERT2OME takes a minute to train for human datasets, and predicts whether 2'-O-methylation modification exists or not in less than a second. In terms of performance evaluation, we have utilized 5-fold cross-validation method. Commonly-used cross-validation approach statistically evaluates the classification model's performance. Such cross-validation prevents the real performance of the prediction model to be overestimated or underestimated. Additionally, we have performed an independent test with the optimal model to calculate the prediction results on unseen data once hyperparameter tuning via cross-validation is over.

We have evaluated the performance of the modification site prediction via various metrics such as accuracy, precision, recall (sensitivity), F1, Area Under Receiver Operating Characteristic Curve (AUC), and Precision-Recall Curve (PR). Let TP, TN, FP, FN represent true positive, true negative, false positive, and false negative respectively. Then, evaluation metrics are defined as in:

$$
\begin{aligned}
&\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\\
&\text{Precision} = \frac{TP}{TP + FP}\\
&\text{Recall} = \frac{TP}{TP + FN}\\
&\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\\
&\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN}\\
&\text{Specificity} = \frac{TN}{FP + TN}
\end{aligned}
$$

ROC AUC is calculated as the area under the Sensitivity (TPR)-(1-Specificity) (FPR) curve. It takes values between 0 and 1 where random guess obtains an ROC AUC score of 0.5. PR curve plots precision values on the y-axis and recall values on the x-axis.

3.4 Baseline Machine Learning Approaches for Protein Post-Translational Modifications (PTMs)

We trained the baseline methods by converting each amino acid into numerical values using the one-hot encoding technique. Each of the 20 amino acids in a given protein sequence is mapped from 20×1 vectors $[1, 0, \ldots, 0]$ to $[0, 0, \ldots, 1]$ respectively. Consequently, 420-length vectors are formed for each 21 amino acid long protein sequence. Subsequently, we converted the corresponding PTM existence tags into 0s and 1s. Sequences with a given PTM are represented with 1s, whereas sequences without a post-translational modification are tagged as 0s. The baseline machine learning classifiers used for this task include Random Forest (RF) [105] and XGBoost [106]. In this case, we have searched over maximum depth in [3, 5, 10, 20], min samples leaf in [1, 2, 4], min samples split in [2, 5, 10], and the number of estimators in [100, 200, 400] via grid search for RF. For XGBoost, we have carried out a grid search to select the best hyperparameters when the maximum depth can be [1, 2, 4, 6], the minimum child weight can be [1, 3, 5], and gamma can be [0, 0.1, 0.2, 0.3, 0.4, 0.5].

In terms of deep learning methods, we have compared DeepPTM with BERT + 1D CNN and BERT + 2D CNN. Both models are trained on a dataset split into training, validation, and test sets with respective sizes of 80%, 10%, 10%. The 1D CNN architecture consists of two convolutional layers, the first with 8 filters and a kernel size of 3, and the second with 16 filters and a kernel size of 3. Both convolutional layers are followed by max-pooling layers with a pool size of 2 as well as dropout layers with dropout rates of 0.1 and 0.2 which are effective in preventing overfitting. The flattened output is then connected to a dense layer with 64 neurons. The final layer is a dense layer with 2 neurons and a softmax activation function. The model is trained for 50 epochs with a batch size of 20, using the Adam optimizer with a learning rate of $5e^{-5}$.

The BERT + 2D CNN model is tuned using a random search with a maximum of 20 trials. The convolutional layers have varying configurations: the first layer has filters in the range of 8 to 32 with a kernel size of either 2 or 6, and the second layer has filters ranging from 10 to 32 with a kernel size of either 2 or 8. The dense layers consist of units in the range of 64 to 128 for the first dense layer and 16 to 96 for the second dense layer. The learning rate is chosen from the set $[1e^{-5}, 3e^{-5}, 6e^{-5}, 8e^{-5}]$. The Adam optimizer is implemented. Early stopping is applied with a minimum delta of $8e^{-4}$ and patience of 40 epochs. The best model obtained from the hyperparameter search is trained on the test set for 200 epochs, with a batch size of 10. Performance metrics, including accuracy, precision, and recall are used throughout the training and validation parts of all the models.

3.5 Hyperparameter Tuning and Performance Evaluation of DeepPTM

We independently performed hyperparameter tuning for ViT part by using Keras-Tuner [107], considering epoch numbers between 50 to 300, transformer layer counts ranging from 1 to 3, and learning rate in either $1e^{-4}$ or $5e^{-4}$.

We implemented DeepPTM in Python by using Keras [107] and Hugging Face [108]. Predictions for a specific modification type were obtained by comparing them with existing state-of-the-art techniques. For succinylation post-translational site prediction, we compared with LMSuccSite [78] and DeepSuccinylSite [79]. For ubiquitination site prediction, comparisons were made with DeepTL-Ubi [77], and Ubisite [76]. Similarly, for glycation site prediction, we compared with GlyStruct [52], and BERT-Kgly [82]. Finally, for crotonylation prediction, comparisons were made with Deep-Kcr [80] and BERT-Kcr [81]. Our experiments were conducted on Google Colab Pro+ with A100 GPUs which have P100, T4, V100 GPUs with 52 GB of memory. On average, training DeepPTM takes approximately 5 minutes over the human dataset, and DeepPTM can predict the existence of a given PTM for each protein sequence in less than a second.

*3.6 Evaluating the Performance of D*eep*P*TM

The evaluation of prediction performance involves several metrics, each providing unique insights into the model's effectiveness. We evaluated prediction performance using multiple metrics including accuracy, precision, recall, F1, Area Under Receiver Operating Characteristic Curve (AUC), and Precision-Recall Curve (PR) [109]. Accuracy measures the overall correctness of predictions, while precision quantifies the proportion of true positive predictions among all positive predictions. Recall, also known as sensitivity, assesses the ability of the model to capture all relevant positive instances. F1 score combines precision and recall, offering a balanced metric that considers false positives and false negatives. ROC AUC assesses the model's ability to discriminate between classes and is calculated as the area under the Sensitivity (TPR)-(1-Specificity) (FPR) curve. The range of ROC AUC is between 0 and 1 where random guess' ROC AUC is 0.5. PR curve plots recall values on the x-axis and precision values on the y-axis. The Precision-Recall (PR) curve plots precision against recall, highlighting the trade-off between these two metrics.

Chapter IV

RESULTS AND DISCUSSION

4.1 2D CNN Performs the Best Among Multiple Classifiers for 2'-O-Methylation Modification Site Prediction

Before constructing our deep learning-based prediction model namely BERT2OME, we have started by applying our baseline models (Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), and XGBoost) for each species (H sapiens, S. cerevisiae and M. musculus) in order to classify whether given input RNA sequence contains 2'-O-methylation modification or not. For feeding our basic machine learning baseline models, each RNA sequence has been converted into the one-hot encoding format which is followed by the implementation of training and testing parts. We have also developed new enhanced baseline models that combine the best-performing 2 baseline methods such as Random Forest and XGBoost with BERT embeddings: BERT + Random Forest, BERT + XGBoost. In this case, instead of giving RNA sequences as in one-hot encoding format to these classification models, we have tried to improve the prediction performance by converting these 41 nucleotide long RNA series into 768 long embedding vector formats by using transformer-based BERT. Then, in order to test the impact of deep learning models with a simpler convolutional neural network structure, we have designed another baseline BERT + 1D CNN. This BERT + 1D CNN approach would give us an idea about how the CNN model performs when we set specific number of nodes, number of layers, batch size or epoch numbers, respectively in the architecture. Also, dropout was used in some layers in order to prevent the overfitting. Lastly, BERT2OME results were added to bottom row which has performed the best over all tested species. The last row in the tables show

BERT2OME performance which learns BERT embeddings combined with chemical properties via 2D CNN. The proposed BERT2OME method has been implemented on Homo sapiens, S. cerevisiae and M. musculus species and we have in general achieved very good results across all these species by using both 5-fold cross-validation and independent testing.

According to Table 3 and Table 4 for human datasets, random forest and XGBoost algorithms have generated better results compared to SVM and decision tree models. These results are consistent across both Human^1 and Human^2 datasets. Once hyper-parameter optimization is carried out, BERT2OME has achieved 0.99 ROC AUC and 99% accuracy scores over Human^1 dataset, 99% accuracy and approximately 0.99 ROC AUC scores over Human^2 dataset. Such results are obtained after integrating the chemical properties approach to our model (We have added three features: ring structure, hydrogen bond and nucleotide frequency while designing the BERT2OME model). In both datasets, combining deep convolutional 2D CNN with BERT embeddings and applying chemical property approach have clearly outperformed all the enhanced baselines. Corresponding ROC AUC curves with respect to the baseline models and our proposed method BERT2OME for Human^1 and Human^2 datasets are given in Figure 3 respectively. Furthermore, our PR AUC curves are represented in Figure 4 for Human^1 and Human^2 datasets. BERT2OME gave us 0.983 PR AUC score for Human^1 dataset. Afterwards, applying chemical properties approach increased this score to 0.998. For Human^2 dataset, BERT2OME achieved 0.976 PR AUC score and it reached 0.999 after applying chemical properties approach. These results show the quality and significance of BERT2OME results for detecting 2'-O-methylation modification sites. When followed by a proper deep learning model such as 2D CNN, BERT embeddings have enhanced the quality of 2'-O-methylation modification site prediction.

We have also evaluated the performance of BERT2OME on two more species:

Table 3: Performance evaluation of BERT2OME with respect to machine learning and enhanced baseline models over RNA Human^1 2'-O-methylation modification dataset in terms of various metrics: Accuracy, Precision, Recall, and F1. BERT2OME (combining 2D CNN with BERT and chemical feature embeddings) achieves the best performance. (Ch. Prop. stands for Chemical Properties)

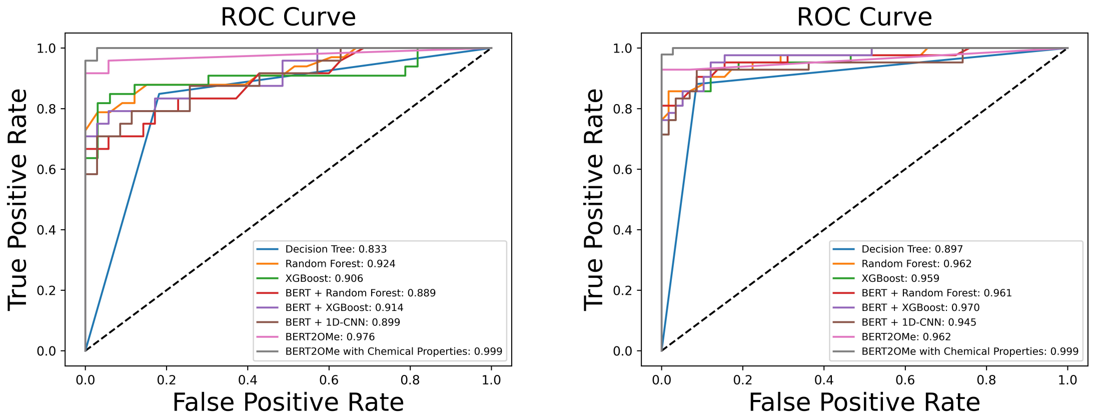
Methods	Accuracy	Precision	Recall	F1
SVM	0.81	0.71	0.81	0.76
Decision Tree	0.80	0.75	0.75	0.75
Random Forest	0.86	0.75	0.9	0.82
XGBoost	0.83	0.71	0.85	0.77
BERT + Random Forest	0.86	0.67	**1.0**	0.80
BERT + XGBoost	0.88	0.71	**1.0**	0.83
BERT + 1D CNN	0.81	0.81	0.81	0.81
BERT2OME (Ch. Prop.)	**0.99**	**0.98**	**1.0**	**0.99**

Table 4: Performance evaluation of BERT2OME with respect to machine learning and enhanced baseline models over RNA Human^2 2'-O-methylation modification dataset in terms of various metrics: Accuracy, Precision, Recall, and F1. BERT2OME (combining 2D CNN with BERT and chemical feature embeddings) achieves the best performance. (Ch. Prop. stands for Chemical Properties)

Methods	Accuracy	Precision	Recall	F1
SVM	0.88	0.83	0.88	0.85
Decision Tree	0.90	0.88	0.88	0.88
Random Forest	0.92	0.86	0.95	0.90
XGBoost	0.90	0.86	0.90	0.88
BERT + Random Forest	0.92	0.88	0.92	0.90
BERT + XGBoost	0.92	0.88	0.93	0.90
BERT + 1D CNN	0.92	0.93	0.89	0.91
BERT2OME (Ch. Prop.)	**0.99**	**1.0**	**0.98**	**0.99**

S. cerevisiae and M. musculus. According to Table 5 and Table 6, we have still obtained significantly better results than the baseline methods according to all considered metrics. For instance, we have obtained higher accuracy and ROC AUC scores for S. cerevisiae as 97% and 0.98 respectively, and for M. musculus as 94% and 0.94 respectively. Even though the datasets are relatively smaller for these species compared to human, these species performance is still quite accurate.

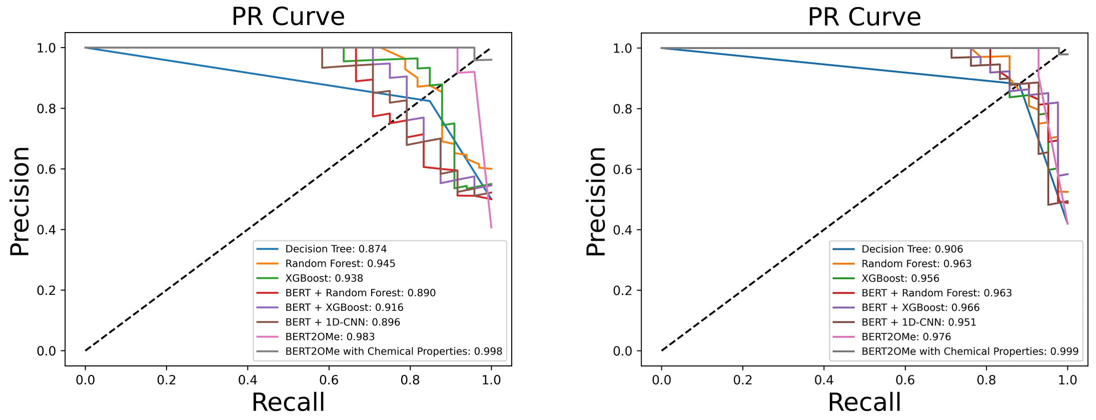
In addition to Table 3 and Table 4, we removed sequences with more than 30% similarity for our Human^1 and Human^2 datasets to decrease the similarity of the



(a) ROC Curve for Human^1

(b) ROC Curve for Human^2

Figure 3: Performance evaluation of BERT2OME with respect to baseline models over RNA Human^1 and Human^2 2'-O-methylation modification datasets in terms of ROC AUC curve. 2D CNN obtains better results than the remaining classifiers, and BERT2OME (combining 2D CNN with BERT and possibly chemical feature embeddings) achieves the best performance. ROC AUC scores are also reported next to the model name.



(a) PR Curve for Human^1

(b) PR Curve for Human^2

Figure 4: Performance evaluation of BERT2OME with respect to baseline models over RNA Human^1 and Human^2 2'-O-methylation modification datasets in terms of PR AUC curve. 2D CNN obtains better results than the remaining classifiers, and BERT2OME (combining 2D CNN with BERT and possibly chemical feature embeddings) achieves the best performance. PR AUC scores are also reported next to the model name.

input sequences. As seen in Table 7, our ROC AUC scores are similar with and without removing the similar sequences for Human^1 and Human^2 datasets. Besides,

Table 5: Performance evaluation of BERT2OME with respect to machine learning and enhanced baseline models over RNA S. cerevisiae 2'-O-methylation modification datasets in terms of various metrics: Accuracy, Precision, Recall, F1 and ROC AUC. BERT2OME (combining 2D CNN with BERT and chemical feature embeddings) achieves the best performance.

Methods	Accuracy	Precision	Recall	F1	ROC AUC
SVM	0.93	0.91	0.96	0.94	0.94
Decision Tree	0.84	0.79	0.92	0.85	0.85
Random Forest	0.86	0.77	0.97	0.86	0.87
XGBoost	0.88	0.81	0.97	0.89	0.89
BERT + Random Forest	0.95	0.96	0.96	**0.96**	0.94
BERT + XGBoost	**0.97**	0.96	**1.0**	0.88	**0.98**
BERT + 1D-CNN	0.95	0.95	0.95	0.95	0.97
BERT2OME (Ch. Prop.)	**0.97**	**0.97**	0.95	**0.96**	**0.98**

Table 6: Performance evaluation of BERT2OME with respect to machine learning and enhanced baseline models over RNA M. musculus 2'-O-methylation modification datasets in terms of various metrics: Accuracy, Precision, Recall, F1 and ROC AUC. BERT2OME (combining 2D CNN with BERT and chemical feature embeddings) achieves the best performance.

Methods	Accuracy	Precision	Recall	F1	ROC AUC
SVM	0.90	0.83	1.0	0.91	0.92
Decision Tree	0.79	0.75	0.86	0.80	0.80
Random Forest	0.86	0.75	1.0	0.86	0.88
XGBoost	0.86	0.75	1.0	0.86	0.88
BERT + Random Forest	0.93	0.88	1.0	0.93	0.94
BERT + XGBoost	0.93	1.0	0.89	0.94	0.92
BERT + 1D-CNN	0.93	0.93	0.93	0.93	0.94
BERT2OME (Ch. Prop.)	0.94	0.94	0.90	0.93	0.94

we achieved 0.995 and 0.999 PR AUC scores for our human datasets as seen in Table 8. In general, results have not changed even after removing the similar sequences. The results show the robustness of BERT2OME across multiple datasets with varying degrees of similarity.

Table 7: ROC AUC scores of BERT2OME with respect to machine learning and enhanced baseline models over RNA Human^1 and Human^2 2'-O-methylation modification datasets after removing the sequences with more than 30% similarity. (Ch. Prop. stands for Chemical Properties)

	Human^1		Human^2	
Methods	w/o removal	with removal	w/o removal	with removal
Decision Tree	0.833	0.817	0.897	0.836
Random Forest	0.924	0.910	0.962	0.976
XGBoost	0.906	0.889	0.959	0.983
BERT + Random Forest	0.889	0.947	0.961	0.965
BERT + XGBoost	0.914	0.957	0.970	0.966
BERT + 1D-CNN	0.899	0.905	0.945	0.934
BERT2OME	0.976	0.989	0.962	**0.999**
BERT2OME (Ch. Prop.)	**0.999**	**0.996**	**0.999**	**0.999**

Table 8: PR AUC scores of BERT2OME with respect to machine learning and enhanced baseline models over RNA Human^1 and Human^2 2'-O-methylation modification datasets after removing the sequences with more than 30% similarity. (Ch. Prop. stands for Chemical Properties)

	Human^1		Human^2	
Methods	w/o removal	with removal	w/o removal	with removal
Decision Tree	0.874	0.909	0.906	0.882
Random Forest	0.945	0.957	0.963	0.981
XGBoost	0.938	0.951	0.956	0.989
BERT + Random Forest	0.890	0.953	0.963	0.978
BERT + XGBoost	0.916	0.958	0.966	0.977
BERT + 1D-CNN	0.896	0.910	0.951	0.963
BERT2OME	0.983	0.989	0.976	**0.999**
BERT2OME (Ch. Prop.)	**0.998**	**0.995**	**0.999**	**0.999**

4.2 Comparison with the State-of-the-art Approaches and BERT2OME

The existing approaches to extract 2'-O-methylation modification sites mostly utilize machine learning algorithms, and they also focus on making predictions only on human dataset. While comparing with the current state-of-the-art approaches, we run these existing methods (NmSEER2.0 [26], iRNA-2OM [22]) to obtain results on multiple human modification site datasets: Human^1 and Human^2. On the other hand, we could not obtain DeepOMe [25] results from their web servers since they always require 120 nucleotide length input. We use the reported results in their paper when needed for comparison with our approach BERT2OME.

Table 9 and Table 10 compares the performance of BERT2OME with the state-of-the-art approaches for Human^1 dataset. According to this table, BERT2OME outperforms all the compared methods. BERT2OME with chemical properties (Obtaining vector embeddings from BERT model, combining with chemical properties vectors, and then applying 2D CNN) with 5-fold cross-validation has given us the best prediction results, 0.999 ROC AUC. Among the rest of the methods, NmSEER2.0 follows us with about 0.578 ROC AUC score, and then iRNA-2OM with about 0.568 ROC AUC score. As stated above, we could not report DeepOMe results for human^1 dataset in Table 9 since DeepOMe always requires 120 nucleotide length input whereas our input nucleotide length is 41 for human^1 dataset. We have also verified the model performance on independent testing, and results are similar when we apply independent testing instead of 5-fold cross-validation. The main reason why our proposed BERT2OME method has obtained significantly better prediction performance than all compared methods is that we have preferred BERT model to generate vector embeddings for a given RNA sequence instead of following up the one-hot encoding strategy. This idea helped us to represent RNA sequences with more numeric values (768 numeric values for each nucleotide), instead of 4 values for each nucleotide. In addition to that strategy, unlike the other studies, we did not just focus on more traditional machine learning models for the training the model, we have instead used 2D CNN with different dimensions.

Besides Human^1 dataset, we have also tested the performance of BERT2OME on a different RNA 2'-O-methylation modification sites dataset over human, Human^2 dataset. To investigate the impact of appending the chemical properties vectors to the BERT embedding vectors, we have also reported BERT2OME results with

Table 9: Performance evaluation of BERT2OME with respect to existing methods for Human^1 dataset in terms of ROC AUC and PR AUC scores for 2'-O-methylation modification. (Ch. Prop. stands for Chemical Properties)

Methods	Classifier	ROC AUC	PR AUC
BERT2OME	2D-CNN	0.976	0.983
BERT2OME (Ch. Prop.)	2D-CNN	**0.999**	**0.998**
DeepOMe	CNN-BiLSTM	N/A	N/A
NmSEER V2.0	RF	0.578	0.254
iRNA-2OM	SVM	0.568	N/A

Table 10: Performance evaluation of BERT2OME with respect to existing methods for Human^2 dataset in terms of ROC AUC and PR AUC scores for 2'-O-methylation modification. (Ch. Prop. stands for Chemical Properties)

Methods	Classifier	ROC AUC	PR AUC
BERT2OME	2D-CNN	0.962	0.976
BERT2OME (Ch. Prop.)	2D-CNN	**0.999**	**0.999**
DeepOMe	CNN-BiLSTM	0.993	0.843
NmSEER V2.0	RF	0.597	0.001
iRNA-2OM	SVM	0.607	0.065

and without the chemical properties vectors for Human^2 dataset. By appending the chemical properties vectors to BERT2OME vectors, we have obtained much longer vector embeddings which help BERT2OME to make more accurate predictions compared to the baseline machine learning approaches. 2D CNN, which is part of our BERT2OME model, again performs the best among different set of classifiers. Similar to the Human^1 dataset results above, we have compared our prediction performance with the other state-of-the-art approaches over Human^2 dataset as seen in Table 10 as well. We have again obtained much better results compared to NmSeer2.0 and iRNA-2OM. Even though DeepOMe seems to perform slightly better than BERT2OME without chemical properties approach, it is still outperformed by BERT2OME with chemical properties approach. Additionally, we used the reported scores in their paper since we could not obtain predictions from their web server. Lastly, we have obtained our results by training only 10% of the Human^2 dataset.

4.3 Performance of BERT2OME Across Multiple Species

To further analyze the performance of BERT2OME in detail, we have trained it on one species and tested on another species to extract 2'-O-methylation sites. We have implemented cross-species prediction between 4 different datasets for 3 different species to understand the 2'-O-methylation sites relationship among the species. We took RNA sequence of each species, train BERT2OME according to the selected species, and evaluated the performance on other species by applying independent testing. Both ROC AUC, accuracy and PR AUC results after the cross-species prediction are given in Table 11 respectively. These 4×4 result matrices can be explained as the following: After training BERT2OME with Human^1 dataset, we obtain 0.96 ROC AUC score and 89% accuracy value if we test the prediction performance on M. musculus. Results in the tables provide us knowledge about cross-species predictability. In general, training BERT2OME with human datasets and predicting on the remaining mouse and yeast species has reasonable prediction performance both in terms of ROC AUC and accuracy. However, training on mouse or yeast and test on human datasets has relatively lower performance. Such non-ideal asymmetric prediction performance can be explained mainly via mouse and yeast datasets being significantly smaller than human datasets which decrease the training quality of BERT2OME. Such smaller datasets may incorporate significantly lower sequence knowledge. Moreover, cross-species prediction results between human and mouse is significantly better than the results between human and yeast. Such result difference can be due to evolutionary similarity between human and mouse, relative to yeast. Species similarity in general also bring about similarity between methylation sequences in the RNAs and the corresponding 2'-O-methylation sites. The significance of benchmark datasets for BERT2OME construction become more apparent according to the cross-species analysis above. Additionally, these results show that 2'-O-methylation site prediction via BERT2OME is robust and stable.

Table 11: Cross-species prediction performance for 4 datasets over 3 species in terms of ROC AUC, Accuracy and PR AUC scores are shown as a matrix. The x-axis defines the species over which BERT2OME is tested, and the y-axis defines the training species.

		Test Species			
	Train Species	Human[1]	S. cerevisiae	M. musculus	Human[2]
ROC AUC	Human[1]	0.999	0.630	0.955	0.982
	S. Cerevisiae	0.597	0.978	0.433	0.632
	M. Musculus	0.625	0.470	0.938	0.478
	Human[2]	0.917	0.529	0.809	0.999
Accuracy	Human[1]	0.992	0.537	0.886	0.941
	S. Cerevisiae	0.566	0.974	0.543	0.602
	M. Musculus	0.607	0.491	0.946	0.480
	Human[2]	0.856	0.512	0.829	0.988
PR AUC	Human[1]	0.998	0.658	0.963	0.986
	S. Cerevisiae	0.534	0.976	0.419	0.617
	M. Musculus	0.575	0.570	0.926	0.468
	Human[2]	0.926	0.548	0.879	0.999

4.4 RNA Sequence Analysis and Interpretation of Results

Consensus motifs across species are quite important to understand why RNA sequences have 2'-O-methylation modification. They are also important to gain biological understanding and insights on these modifications. In this case, we use state-of-the-art method STREME [110] to discover consensus sequence motifs related to 2'-O-methylation modifications. STREME identifies consecutive motifs which are relatively enriched among the input sequences compared to the control sequences. Table 12 shows the topmost 5 enriched statistically significant consensus motifs of length 3 for all datasets. In general, we found conservation between different datasets and species. According to results, all 4 datasets are biased towards consensus motifs with rich G, showing that 2'-O-methylation modification sites have a major degree of conservation at the single nucleotide and fragment levels.

We identified motif UGA to exist across all species showing its importance in observing 2'-O-methylation modification in a given RNA sequence. Similarly, motif

UGG are enriched in 3 datasets among 4 of them. Thus, we observe some consensus patterns across species which is important for cross-species prediction. However, there are still some motif differences between species showing existing shallow sequence-based methods may not be enough for best prediction performance, whereas deep models have a potential in increasing prediction performance by extracting the consensus patterns more efficiently.

Additionally, the extracted consensus motifs might indicate a number of potential relations between different types of sequence modifications. For example, GAA and AGA motifs are close to the sequence motif reported by m^{1}A RNA modification as discussed in [111]. The identified GGACU/GAACU consensus motifs are the most powerful sequence motifs of m^{6}A. These observations may provide some evidence for potential interplay between 2'-O-methylation and m^{1}A as well as m^{6}A which is compatible with results in [28]. When analyzed contextually from the perspective of sequences, there exists a remarkable correlation between various RNA modification types.

Table 12: The top 5 consensus motifs of length 3 for each dataset.

Dataset	Top 1	Top 2	Top 3	Top 4	Top 5
Human^1	GAA	UGA	CUG	UGG	UUG
Human^2	AGA	GGA	UGA	GCA	GAA
S. cerevisiae	AAG	UGG	GUA	UGA	GGA
Mus musculus	AUG	UGA	CCC	UCC	UGG

4.5 Sequence Parameter Selection for DeepPTM

Before comparing DeepPTM with the baseline approaches, we have decided on protein sequence length and elimination thresholds of similar sequences for DeepPTM. As seen in Table 13 and Table 14 for the Homo Sapiens Succinylation dataset, we decided to proceed with 40% CD-HIT threshold to remove redundancy in homologous sequences since it DeepPTM performs the best with 40% CD-HIT threshold compared to applying 30% threshold and without removing any sequences.

Similarly, we also decided to transform proteins into peptides with a fixed 21 length. The performance of DeepPTM with sequence length 21 outperforms other commonly-used sequence lengths such as 15, 27, and 33 [112, 82] even though it is shorter than most of them, as seen in Table 15 and Table 16 over the Homo Sapiens Succinylation dataset.

Table 13: Performance comparison of DeepPTM based on accuracy, precision, recall and f1 metrics with different CD-HIT threshold values for Homo Sapiens Succinylation dataset.

Methods	Accuracy	Precision	Recall	F1
DeepPTM (CD-HIT threshold 40%)	0.940	0.940	0.940	0.940
DeepPTM (CD-HIT threshold 30%)	0.886	0.885	0.895	0.890
DeepPTM (Without removal)	0.876	0.875	0.870	0.872

Table 14: Performance comparison of DeepPTM based on ROC AUC and PR AUC metrics with different CD-HIT threshold values for Homo Sapiens Succinylation dataset.

Methods	ROC AUC	PR AUC
DeepPTM (CD-HIT threshold 40%)	0.988	0.990
DeepPTM (CD-HIT threshold 30%)	0.959	0.960
DeepPTM (Without removal)	0.940	0.937

Table 15: Performance comparison of DeepPTM based on accuracy, precision, recall and f1 metrics with different sequence lengths for Homo Sapiens Succinylation dataset.

Methods	Accuracy	Precision	Recall	F1
DeepPTM (sequence length = 15)	0.933	0.928	0.930	0.929
DeepPTM (sequence length = 21)	0.940	0.940	0.940	0.940
DeepPTM (sequence length = 27)	0.928	0.913	0.923	0.918
DeepPTM (sequence length = 33)	0.903	0.904	0.903	0.903

Table 16: Performance comparison of DeepPTM based on ROC AUC and PR AUC metrics with different sequence lengths for Homo Sapiens Succinylation dataset.

Methods	ROC AUC	PR AUC
DeepPTM (sequence length = 15)	0.988	0.987
DeepPTM (sequence length = 21)	0.988	0.990
DeepPTM (sequence length = 27)	0.986	0.986
DeepPTM (sequence length = 33)	0.975	0.976

4.6 ViT Performs the Best Among Multiple Classifiers for Protein Post-Translational Modification Prediction

Before the construction of attention-based DeepPTM, we have applied our baseline learning techniques (Random Forest, XGBoost, 1D CNN, 2D CNN) to considered species (H. sapiens (human), S. cerevisiae (yeast), M. musculus (mouse)) in order to categorize whether input protein sequence includes a given post-translational modification or not. We applied one hot encoding to each protein sequence for our baseline models without BERT. We have also designed stronger baseline approaches BERT + Random Forest, BERT + XGBoost, BERT + 1D CNN, and BERT + 2D CNN which integrate machine/deep learning classifiers with BERT. In such an enhanced baseline setting, we do not encode protein sequences by one hot encoding but instead, we convert protein series into 768 long embedding vectors via BERT. The performance of 1D CNN and 2D CNN gives us the impact of deep learning methods relative to machine learning methods. The proposed DeepPTM has been implemented across Homo sapiens, M. musculus, S. cerevisiae species as well as across succinylation, glycation, crotonylation, and ubiquitination modifications. We have generally obtained remarkable results over all those species and modification types via 10-fold cross-validation.

DeepPTM has achieved 0.988 ROC AUC, 0.990 PR AUC, and 94.0% accuracy while predicting human succinylation sites as seen in Table 17 and Table 18, outperforming all the enhanced baselines. ProtBERT combined with 2D CNN outperforms the machine learning models. Table 19, Table 20 and Table 21 evaluate DeepPTM's performance for inferring 3 other PTMs on human: glycation, crotonylation, and ubiquitination. For instance, DeepPTM significantly outperforms BERT + 1D CNN, BERT + 2D CNN, and BERT + ViT on crotonylation prediction: It can achieve 0.955 ROC AUC, 0.955 PR AUC, and 87.8% accuracy. Overall, those tables and results indicate the significance of DeepPTM predictions. When followed by an attention-based deep learning model ViT, BERT embeddings have topped the quality of all PTM predictions. Additionally, the performance of DeepPTM is quite robust across multiple runs, which can be seen from smaller standard deviations in Table 17 and Table 18.

Table 17: Performance comparison of DeepPTM based on accuracy, precision, recall, f1 metrics with standard machine learning and enhanced baseline models over human Succinylation modification dataset. DeepPTM attains the topmost performance.

Methods	Accuracy	Precision	Recall	F1
Random Forest	0.687	0.724	0.649	0.685
XGBoost	0.692	0.718	0.677	0.697
BERT + Random Forest	0.660	0.681	0.658	0.669
BERT + XGBoost	0.652	0.671	0.658	0.665
BERT + 1D CNN	0.670	0.652	0.768	0.706
BERT + 2D CNN	0.814	0.780	0.849	0.813
BERT + ViT	0.911	0.896	0.931	0.914
DeepPTM	**0.940** ±0.01	**0.940** ±0.01	**0.940** ±0.01	**0.940** ±0.01

We have also evaluated the performance of DeepPTM for predicting succinylation modification over 2 additional species: M. musculus (mouse) and S. cerevisiae (yeast) as seen in Table 22 and Table 23. According to the tables, DeepPTM has significantly outperformed all the baseline techniques except BERT + ViT in terms of all studied performance metrics. As an example, DeepPTM has attained better accuracy and ROC AUC over yeast as 90.7% and 0.970 respectively, and over the mouse as 89.0% and 0.965 respectively. Although the datasets of those two species are slightly smaller than the human dataset, their performance is still as accurate as the human dataset.

Table 18: Performance comparison of DeepPTM based on ROC AUC, PR AUC metrics with standard machine learning and enhanced baseline models over human Succinylation modification dataset. DeepPTM attains the topmost performance.

Methods	ROC AUC	PR AUC
Random Forest	0.756	0.746
XGBoost 0.755	0.735	
BERT + Random Forest	0.706	0.673
BERT + XGBoost	0.703	0.689
BERT + 1D CNN	0.737	0.748
BERT + 2D CNN	0.893	0.890
BERT + ViT	0.974	0.975
DeepPTM: ProtBERT + ViT	**0.988** ±0.01	**0.990** ±0.01

Table 19: Performance comparison of DeepPTM with standard machine learning and enhanced baseline models over human on Glycation dataset. ROC stands for ROC AUC and PR stands for precision-recall curve AUC.

Methods	Accuracy	PR	ROC
Random Forest	0.542	0.574	0.564
XGBoost	0.571	0.606	0.610
BERT + Random Forest	0.563	0.571	0.589
BERT + XGBoost	0.568	0.563	0.588
BERT + 1D CNN	0.575	0.663	0.625
BERT + 2D CNN	0.613	0.671	0.637
BERT + ViT	**0.881**	0.953	0.951
DeepPTM: ProtBERT + ViT	0.880	**0.956**	**0.953**

Table 20: Performance comparison of DeepPTM with standard machine learning and enhanced baseline models over human on Crotonylation dataset. ROC stands for ROC AUC and PR stands for precision-recall curve AUC.

Methods	Accuracy	PR	ROC
Random Forest	0.667	0.731	0.746
XGBoost	0.689	0.739	0.755
BERT + Random Forest	0.636	0.660	0.687
BERT + XGBoost	0.677	0.705	0.716
BERT + 1D CNN	0.692	0.704	0.743
BERT + 2D CNN	0.650	0.701	0.715
BERT + ViT	0.863	0.946	0.945
DeepPTM: ProtBERT + ViT	**0.878**	**0.955**	**0.955**

The effectiveness of ViT can also be seen in BERT + ViT results, which outperforms all other baselines without ViT.

Table 21: Performance comparison of DeepPTM with standard machine learning and enhanced baseline models over human on Ubiquitination dataset. ROC stands for ROC AUC and PR stands for precision-recall curve AUC.

Methods	Accuracy	PR	ROC
Random Forest	0.545	0.598	0.605
XGBoost	0.564	0.587	0.589
BERT + Random Forest	0.590	0.624	0.622
BERT + XGBoost	0.577	0.615	0.599
BERT + 1D CNN	0.608	0.605	0.612
BERT + 2D CNN	0.635	0.638	0.667
BERT + ViT	0.910	0.972	0.971
DeepPTM: ProtBERT + ViT	**0.940**	**0.980**	**0.982**

Table 22: Performance comparison of DeepPTM with standard machine learning and enhanced baseline models over S. cerevisiae Succinylation modifications. ROC and PR stand for ROC AUC and PR AUC respectively. DeepPTM and BERT + ViT attain the topmost performance for S. cerevisiae.

Methods	Accuracy	F1	ROC	PR
Random Forest	0.650	0.678	0.712	0.709
XGBoost	0.687	0.705	0.750	0.761
BERT + Random Forest	0.580	0.600	0.597	0.645
BERT + XGBoost	0.623	0.655	0.691	0.757
BERT + 1D CNN	0.660	0.710	0.695	0.734
BERT + 2D CNN	0.716	0.717	0.772	0.762
BERT + ViT	**0.926**	**0.936**	**0.974**	**0.976**
DeepPTM: ProtBERT + ViT	0.907	0.910	0.970	0.970

Table 23: Performance comparison of DeepPTM with standard machine learning and enhanced baseline models over M. musculus Succinylation modifications. ROC and PR stand for ROC AUC and PR AUC respectively. DeepPTM and BERT + ViT attain the topmost performance for M. musculus.

Methods	Accuracy	F1	ROC	PR
Random Forest	0.564	0.552	0.605	0.618
XGBoost	0.590	0.589	0.592	0.601
BERT + Random Forest	0.541	0.532	0.600	0.592
BERT + XGBoost	0.561	0.549	0.596	0.587
BERT + 1D CNN	0.616	0.660	0.648	0.655
BERT + 2D CNN	0.618	0.610	0.662	0.624
BERT + ViT	**0.896**	**0.922**	**0.967**	**0.967**
DeepPTM: ProtBERT + ViT	0.890	0.890	0.965	0.965

4.7 Comparison with the State-of-the-art Approaches and DeepPTM

The current methods to infer these PTMs may range from machine learning-based to deep learning-based methods. In Table 24, we compare DeepPTM with the two best-performing methods for each of these four modifications: succinylation, ubiquitination, glycation, and crotonylation. DeepPTM significantly outperforms all the compared methods as seen in the table in terms of ROC AUC. The most of the benchmark methods such as LMSuccSite [78], DeepSuccinylSite [79], DeepTL-Ubi [77], Ubisite [76], GlyStruct [52], and Deep-Kcr [80] do not take into account BERT embeddings. Instead, they model input by different types of feature encoding schemes such as local backbone angles, accessible surface area, secondary structure, sequence-based features, etc. In this case, the outperformance of DeepPTM over these methods is mainly due to its ability to use the generated BERT embeddings for a given protein sequence. Even though BERT-Kcr and BERT-Kgly also focus on benefiting from BERT embeddings, they do not incorporate attention-based transformer structures such as ViT which explains their lower performance with respect to DeepPTM. However, among these benchmark methods, BERT-Kcr and Kert-Kgly still perform better than the rest of the methods. Additionally, the predictions made by all of these methods are limited to a certain type of PTM whereas DeepPTM is capable of generating different types of PTMs for a given protein sequence. In Table 24, we put - when a method cannot generate predictions for a given modification type.

4.8 Performance of DeepPTM Across Different Species and Modifications

We have analyzed DeepPTM's performance in inferring succinylation PTM further by training it on a single species and testing it on a different species. By carrying out

Table 24: The performance of DeepPTM in terms of ROC AUC relative to existing methods (2 top-performing methods for each modification type): LMSuccSite, DeepSuccinylSite, DeepTL-Ubi, UbiSite, GlyStruct, BERT-Kgly, Deep-Kcr, and BERT-Kcr for 4 human modification datasets. We put - for methods that cannot generate predictions for a given modification type.

Methods	Succinylation	Ubiquitination	Glycation	Crotonylation
DeepPTM	**0.988**	**0.982**	**0.953**	**0.955**
LMSuccSite [78]	0.756	-	-	-
DeepSuccinylSite [79]	0.789	-	-	-
DeepTL-Ubi [77]	-	0.772	-	-
Ubisite [76]	-	0.748	-	-
GlyStruct [52]	-	-	0.721	-
BERT-Kgly [82]	-	-	0.840	-
Deep-Kcr [80]	-	-	-	0.859
BERT-Kcr [81]	-	-	-	0.905

cross-species prediction among 3 species, we can understand the succinylation modification sites relationships between these different species. Once we train DeepPTM in terms of the chosen species protein sequences, we have applied independent testing to evaluate the performance of the remaining species as seen in Table 25. For instance, after training DeepPTM with human succinylation modifications, we attain 0.955 ROC AUC if we test the prediction performance on the mouse. According to the table, all cross-species prediction performance is reasonably high. Results show that succinylation modification prediction across species via DeepPTM is robust and stable.

Table 25: Cross-species prediction performance for Succinylation dataset over 3 species in terms of ROC AUC. The x-axis defines the species over which DeepPTM is tested, and the y-axis defines the training species.

	Test		
Train	Homo Sapiens	S. Cerevisiae	M. Musculus
Homo Sapiens	0.988	0.954	0.955
S. Cerevisiae	0.969	0.970	0.938
M. Musculus	0.964	0.956	0.965

We have also tested the predictability of a modification by training DeepPTM on a different PTM site. As shown in Table 26, we have applied such cross-training and testing on 4 modifications for human. In terms of ROC AUC, we still obtain a high score when we train DeepPTM on crotonylation sites and test the prediction performance on succinylation sites. Similar to the cross-species prediction performance above, all cross-modification prediction performance is again reasonably high as well as quite robust.

Table 26: Cross-modification prediction performance for 4 modifications over human in terms of ROC AUC. The x-axis defines the modifications over which DeepPTM is tested, and the y-axis defines the training modifications.

	Test			
Train	Succinylation	Ubiquitination	Glycation	Crotonylation
Succinylation	0.988	0.975	0.960	0.939
Ubiquitination	0.961	0.982	0.945	0.943
Glycation	0.978	0.973	0.951	0.958
Crotonylation	0.983	0.952	0.949	0.955

4.9 Protein Sequence Analysis and Interpretation of Results

Consensus sequence motifs across species are crucial in understanding the reason proteins possess various types of lysine modifications. Additionally, these motifs are crucial in gaining a biological understanding of PTMs. Here, we utilize frequently-used STREME [110] in discovering consensus motifs associated with succinylation, ubiquitination, glycation, and crotonylation modifications. STREME identifies consensus motifs among input sequences that are significantly enriched relative to the control sequences. The topmost statistically significant five motifs with length 3 are shown in Table 27 for different PMTs and species. Generally, consensus motifs among various species and PTMs are conserved robustly. Apart from lysine (K), we observe bias towards motifs with rich Alanine/A or Valine/V, showing that all tested PTMs are significantly conserved at the single amino acid level.

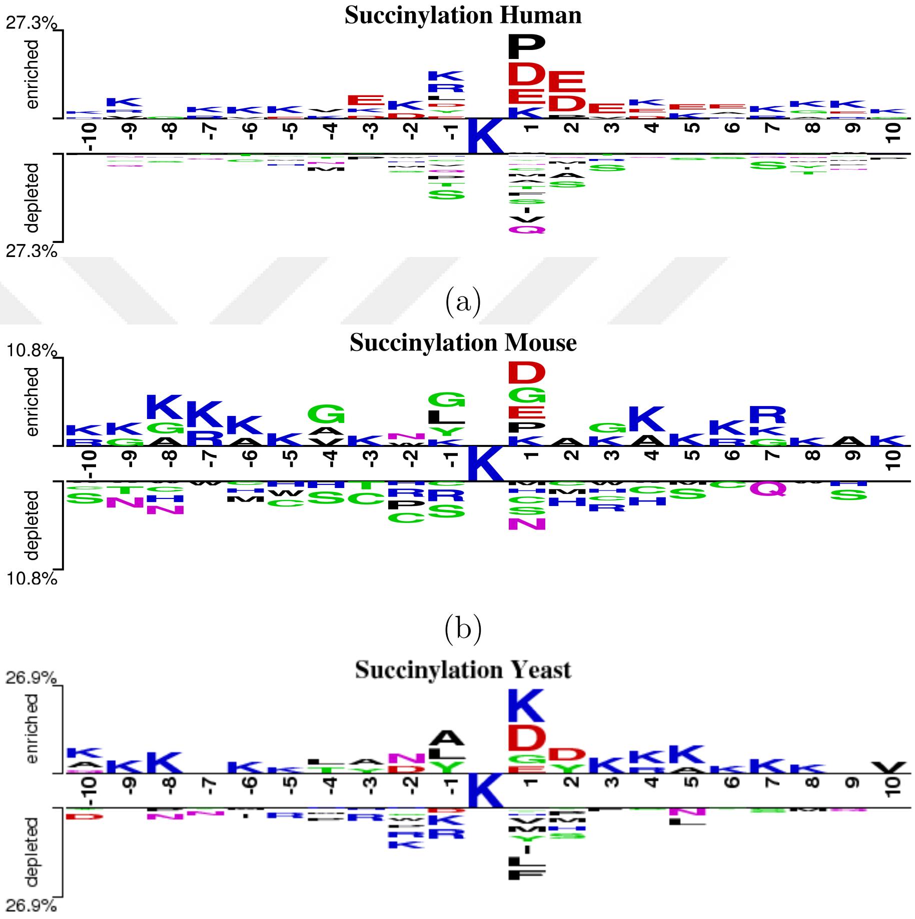
We identified motif AK (Alanine-Lysine) to occur over all tested species indicating its prominence in all considered lysine modifications for a given protein sequence.

Similarly, motif RVL (Arginine-Valine-Leucine) is enriched in most of the datasets; for both succinylation, ubiquitination, and crotonylation. Those enrichment results might potentially indicate the interaction between these different types of modifications that is consistent with the findings discussed in [46]. A significant correlation exists between different protein PTM types when examined from the sequences perspective. Similarly, motif AKK (Alanine-Lysine-Lysine) is enriched in all species except human for succinylation. Thus, we detect several consensus patterns between species and PTMs that are important for cross-species and cross-modification predictions. Nevertheless, the existence of several motif differences among species shows that shallower sequence-based approaches might not be sufficient for optimal performance. However, deeper methods can potentially increase the prediction performance via better utilizing the consensus sequence patterns.

Table 27: 3-length topmost four consensus motifs across different species and modifications

Modification	Species	Top 1	Top 2	Top 3	Top 4
Succinylation	Homo Sapiens	KPE	EEE	RVL	DDE
Succinylation	M. Musculus	AKK	AVA	KDF	PRL
Succinylation	S. Cerevisiae	EKD	AKK	GVT	NLK
Ubiquitination	Homo Sapiens	EKQ	RVL	KLD	AKK
Glycation	Homo Sapiens	GLK	KLS	VAL	EKL
Crotonylation	Homo Sapiens	RVL	AKV	KAA	VKD

Two-sample logo software [1] is run to discover the patterns of succinylation modification site sequences on human, mouse, and yeast as seen in Figure 5 and the patterns of crotonylation, ubiquitination, and glycation sites modification site sequences on human as seen in Figure 6. In general, the sequence preference profile of human succinylation is similar to the profiles of the remaining species. For instance, over-represented Aspartic acid/D and Lysine/K residues occur at position +1 in succinylation sites of different species. Lysine/K residue is enriched at position -1 of the succinylation site in all the tested species. Besides, Arginine/R residues are depleted at multiple positions (-1, +1, +2, +3) across many species for succinylation. Collectively, those results indicate the existence of certain similarities across patterns of sequence between human and the remaining species. Hence, transfer learning emerges as a tempting method that can be used to efficiently infer sites specific to species.

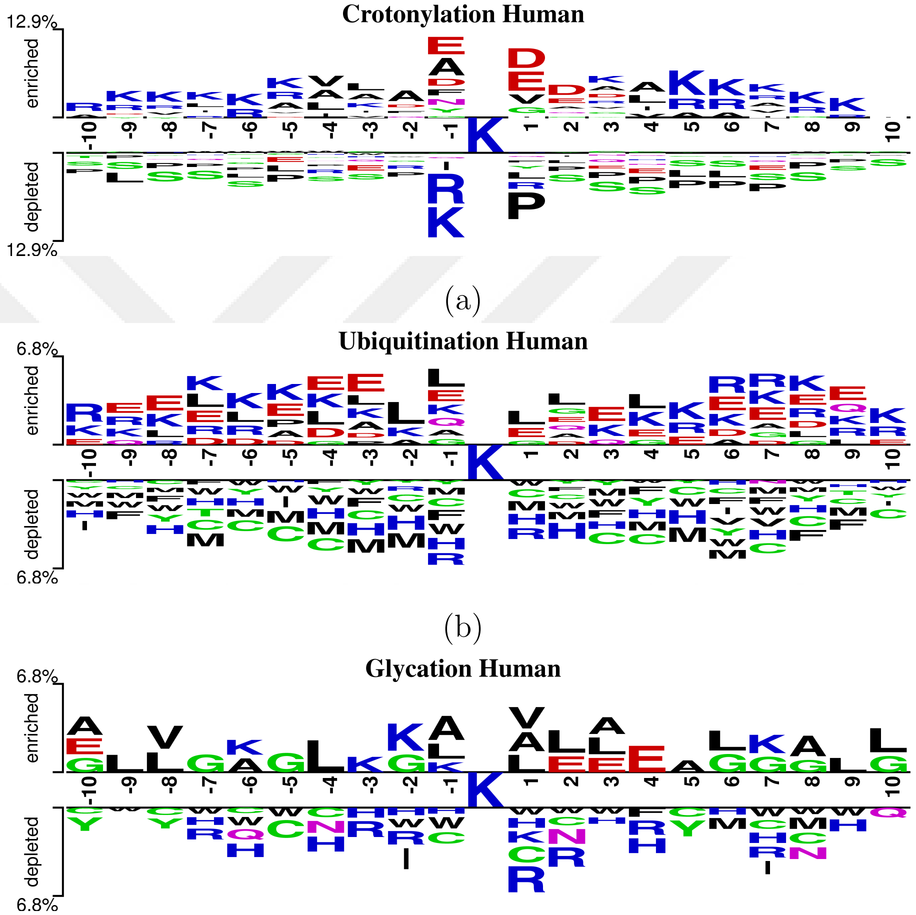


(a)

(b)

(c)

Figure 5: Location-specific residue composition near the Succinylation, site and non-modification sites over human, mouse, and yeast datasets, represented by two-sample logo software [1]. For all these modifications, we only show the residues that are statistically significantly enriched or depleted (t-test, $p < 0.05$) surrounding the centered modifications sites (downstream 10 residues and upstream 10 residues).



(a)

(b)

(c)

Figure 6: Location-specific residue composition near the Crotonylation, Ubiquitination, and Glycation sites and non-modification sites over human dataset, represented by two-sample logo software [1]. For all these modifications, we only show the residues that are statistically significantly enriched or depleted (t-test, $p < 0.05$) surrounding the centered modifications sites (downstream 10 residues and upstream 10 residues).

Chapter V

CONCLUSIONS

In this research, we come up with a novel approach BERT2OME to extract RNA 2'-O-methylation modification sites across multiple species from RNA sequences. BERT2OME is based on analyzing the latent knowledge in RNA sequences by BERT model. We have assumed each RNA sequence as text and implemented BERT to convert RNA sequences into vector embeddings. By utilizing BERT pretrained models, BERT2OME can transfer the syntactic and semantic information from massive human language corpora to RNA datasets. In addition to BERT embeddings of RNA sequences, we also encode for nucleotide chemical attributes to obtain a better performance. In our experiments, we have verified BERT2OME performance by comparing with existing techniques as well as commonly-used machine learning approaches. We have identified deeper neural network techniques such as 2D CNN to outperform shallower neural networks and more traditional machine learning algorithms in terms of accuracy when learning BERT features. As a result, BERT2OME, by incorporating BERT-base uncased model and 2D CNN, increases RNA 2'-O-methylation modification site prediction performance compared to the existing methods via transferring the RNA language. We have also found BERT2OME to be robust according to various detailed experiments. Additionally, BERT2OME has the potential to detect novel modification sites without even running biological experiments.

The limitations of this study can be summarized as follows: 1- Even though we have significantly achieved a better performance, added biological insights, and biological inferences, our deep-learning method is still not fully interpretable like all existing deep-learning based methods, 2- We have focused on datasets with relatively smaller sequence fragment lengths, so our running time may increase once we adapt our method to longer sequences. As a future work, one can use cross-species training capability of BERT2OME to predict modification sites across species without any experimental RNA 2'-O-methylation dataset. Such cross-species training will be important in expanding RNA 2'-O-methylation dataset of multiple species. BERT2OME can be applied to other RNA modification's prediction problem. Lastly, graph neural networks can be integrated to feature encoding to enhance the model performance.

Besides, we present a novel method DeepPTM to infer various types of PTMs from protein sequences over multiple species. DeepPTM extracts the hidden information over protein sequences via ProtBERT where each sequence is modeled as a text and ProtBERT converts the protein sequences into numerical embeddings. As a result, DeepPTM could transfer the semantic and syntactic knowledge to protein datasets. In our experiments, DeepPTM has outperformed the competing methods and frequently-used machine/deep learning techniques across different types of post-translational modifications. We have identified deeper and more recently introduced neural network ViT to perform better than shallower methods when they learn from ProtBERT and BERT attributes. Consequently, DeepPTM improves PTM detection performance by transferring the protein language and better utilization of ProtBERT embeddings by ViT. Additionally, DeepPTM could potentially infer novel PTM sites without even carrying out laboratory experiments. Since our cross-species and cross-modification prediction results are reasonably accurate, transferring the knowledge from extensively-studied species such as humans to other less-studied species will speed up biological and medical research.

Although we have obtained better performance with derived biological insights, DeepPTM results cannot be perfectly interpreted, which can be considered as one possible enhancement for the future. In the future, DeepPTM's cross-species training property can be used to infer PTMs over species without any experimental protein PTM dataset, which will be crucial to expanding PTM knowledge to numerous species. Moreover, graph neural network-based feature encoding can increase the prediction performance. Using a predetermined fixed sequence length to analyze the protein sequences embeddings can be seen as one main limitation of DeepPTM. If we prefer to transfer our learning between different sequence-length proteins, DeepPTM needs to be enhanced.

Recently, newer language models have been trained and made publicly available to learn representations over a broad set of protein sequences. In future work, in line with such enhanced protein language models, we plan to evaluate the impact of multiple different protein language models on PTM prediction performance.

REFERENCES

[1] V. Vacic, L. M. Iakoucheva, and P. Radivojac, “Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments,” *Bioinformatics*, vol. 22, pp. 1536–1537, 04 2006.

[2] J.-J. Xuan, W.-J. Sun, P.-H. Lin, K.-R. Zhou, S. Liu, L.-L. Zheng, L.-H. Qu, and J.-H. Yang, “RMBase v2.0: deciphering the map of RNA modifications from epitranscriptome sequencing data,” *Nucleic Acids Research*, vol. 46, pp. D327–D334, 10 2017.

[3] X. Zeng, X. Zhang, and Q. Zou, “Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks,” *Briefings in Bioinformatics*, vol. 17, pp. 193–203, 06 2015.

[4] Y. Zhou, Y. Zhang, X. Lian, F. Li, C. Wang, F. Zhu, Y. Qiu, and Y. Chen, “Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents,” *Nucleic Acids Research*, vol. 50, pp. D1398–D1407, 10 2021.

[5] X. Luo, F. Wang, G. Wang, and Y. Zhao, “Identification of methylation states of dna regions for illumina methylation beadchip,” *BMC Genomics*, vol. 21, no. 1, p. 672, 2020.

[6] X. Luo, T. Zhang, Y. Zhai, F. Wang, S. Zhang, and G. Wang, “Effects of dna methylation on tfs in human embryonic stem cells,” *Frontiers in Genetics*, vol. 12, pp. 1–10, 2021.

[7] L. Liu, B. Song, J. Ma, Y. Song, S.-Y. Zhang, Y. Tang, X. Wu, Z. Wei, K. Chen, J. Su, R. Rong, Z. Lu, J. P. de Magalhães, D. J. Rigden, L. Zhang, S.-W. Zhang, Y. Huang, X. Lei, H. Liu, and J. Meng, “Bioinformatics approaches for deciphering the epitranscriptome: Recent progress and emerging topics,” *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1587–1604, 2020.

[8] T. Kiss, “Small nucleolar rnas: An abundant group of noncoding rnas with diverse cellular functions,” *Cell*, vol. 109, no. 2, pp. 145–148, 2002.

[9] D. Incarnato, F. Anselmi, E. Morandi, F. Neri, M. Maldotti, S. Rapelli, C. Parlato, G. Basile, and S. Oliviero, “High-throughput single-base resolution mapping of RNA 2’-O-methylated residues,” *Nucleic Acids Research*, vol. 45, pp. 1433–1441, 09 2016.

[10] C. Y. Choi, Y. Zhao, F. Wang, and L. Juan, “Microrna promoter identification in arabidopsis using multiple histone markers,” *BioMed Research International*, vol. 2015, p. 861402, 2015.

[11] W. Chen, Y. Zhao, F. Wang, S. Chen, J. Wan, and G. Wang, “Methods of microrna promoter prediction and transcription factor mediated regulatory network,” *BioMed Research International*, vol. 2017, p. 7049406, 2017.

[12] J.-P. Bachellerie, J. Cavaillé, and A. Hüttenhofer, “The expanding snorna world,” *Biochimie*, vol. 84, no. 8, pp. 775–790, 2002.

[13] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, and F. Zhu, “Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics,” *Nucleic Acids Research*, vol. 48, pp. D1031–D1041, 11 2019.

[14] X. Zeng, L. Liu, L. Lü, and Q. Zou, “Prediction of potential disease-associated microRNAs using structural perturbation method,” *Bioinformatics*, vol. 34, pp. 2425–2432, 02 2018.

[15] J. Yin, W. Sun, F. Li, J. Hong, X. Li, Y. Zhou, Y. Lu, M. Liu, X. Zhang, N. Chen, X. Jin, J. Xue, S. Zeng, L. Yu, and F. Zhu, “VARIDT 1.0: variability of drug transporter database,” *Nucleic Acids Research*, vol. 48, pp. D1042–D1050, 09 2019.

[16] Z.-W. Dong, P. Shao, L.-T. Diao, H. Zhou, C.-H. Yu, and L.-H. Qu, “RTL-P: a sensitive approach for detecting sites of 2’-O-methylation in RNA molecules,” *Nucleic Acids Research*, vol. 40, pp. e157–e157, 07 2012.

[17] R. Züst, L. Cervantes-Barragan, M. Habjan, R. Maier, B. W. Neuman, J. Ziebuhr, K. J. Szretter, S. C. Baker, W. Barchet, M. S. Diamond, S. G. Siddell, B. Ludewig, and V. Thiel, “Ribose 2’-o-methylation provides a molecular signature for the distinction of self and non-self mrna dependent on the rna sensor mda5,” *Nature Immunology*, vol. 12, no. 2, pp. 137–143, 2011.

[18] K. Chen, B. Song, Y. Tang, Z. Wei, Q. Xu, J. Su, J. P. de Magalhães, D. J. Rigden, and J. Meng, “RMDisease: a database of genetic variants that affect RNA modifications, with implications for epitranscriptome pathogenesis,” *Nucleic Acids Research*, vol. 49, pp. D1396–D1404, 10 2020.

[19] Z.-W. Dong, P. Shao, L.-T. Diao, H. Zhou, C.-H. Yu, and L.-H. Qu, “RTL-P: a sensitive approach for detecting sites of 2’-O-methylation in RNA molecules,” *Nucleic Acids Research*, vol. 40, pp. e157–e157, 07 2012.

[20] Y.-T. Yu, M. Shu, and J. A. Steitz, “A new method for detecting sites of 2’-o-methylation in rna molecules,” *RNA (New York, N.Y.)*, vol. 3, pp. 324–31, 04 1997.

[21] W. Chen, P. Feng, H. Tang, H. Ding, and H. Lin, “Identifying 2’-o-methylationation sites by integrating nucleotide chemical properties and nucleotide compositions,” *Genomics*, vol. 107, no. 6, pp. 255–258, 2016.

[22] H. Yang, H. Lv, H. Ding, W. Chen, and H. Lin, "irna-2om: A sequence-based predictor for identifying 2'-o-methylation sites in homo sapiens," *Journal of Computational Biology*, vol. 25, no. 11, pp. 1266–1277, 2018. PMID: 30113871.

[23] C. Ao, Q. Zou, and L. Yu, "NmRF: identification of multispecies RNA 2'-O-methylation modification sites from RNA sequences," *Briefings in Bioinformatics*, vol. 23, pp. 1–13, 11 2021. bbab480.

[24] M. Tahir, H. Tayara, and K. T. Chong, "irna-pseknc(2methyl): Identify rna 2'-o-methylation sites by convolution neural network and chou's pseudo components," *Journal of Theoretical Biology*, vol. 465, pp. 1–6, 2019.

[25] H. Li, L. Chen, Z. Huang, X. Luo, H. Li, J. Ren, and Y. Xie, "Deepome: A web server for the prediction of 2'-o-me sites based on the hybrid cnn and blstm architecture," *Frontiers in Cell and Developmental Biology*, vol. 9, pp. 1–9, 2021.

[26] Y. Zhou, Q. Cui, and Y. Zhou, "Nmseer v2.0: a prediction tool for 2'-o-methylation sites based on random forest and multi-encoding combination," *BMC Bioinformatics*, vol. 20, no. 25, p. 690, 2019.

[27] L. Zhang, G. Li, X. Li, H. Wang, S. Chen, and H. Liu, "Edlm6apred: ensemble deep learning approach for mrna m6a site prediction," *BMC Bioinformatics*, vol. 22, no. 1, p. 288, 2021.

[28] Z. Song, D. Huang, B. Song, K. Chen, Y. Song, G. Liu, J. Su, J. P. d. Magalhães, D. J. Rigden, and J. Meng, "Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring rna modifications," *Nature Communications*, vol. 12, no. 1, p. 4011, 2021.

[29] K. Liu and W. Chen, "iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications," *Bioinformatics*, vol. 36, pp. 3336–3342, 03 2020.

[30] H. Wang, H. Liu, T. Huang, G. Li, L. Zhang, and Y. Sun, "Emdlp: Ensemble multiscale deep learning model for rna methylation site prediction," *BMC Bioinformatics*, vol. 23, no. 1, p. 221, 2022.

[31] Y. Fan, G. Sun, and X. Pan, "Elmo4m6a: a contextual language embedding-based predictor for detecting rna n6-methyladenosine sites," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2022.

[32] J. Li, Y. Pu, J. Tang, Q. Zou, and F. Guo, "Deepavp: A dual-channel deep neural network for identifying variable-length antiviral peptides," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 3012–3019, 2020.

[33] N. Q. K. Le, Q.-T. Ho, E. K. Y. Yapp, Y.-Y. Ou, and H.-Y. Yeh, "Deepetc: A deep convolutional neural network architecture for investigating and classifying electron transport chain's complexes," *Neurocomputing*, vol. 375, pp. 71–79, 2020.

[34] E. Sefer and C. Kingsford, “Metric labeling and semimetric embedding for protein annotation prediction,” *Journal of Computational Biology*, vol. 28, no. 5, pp. 514–525, 2021. PMID: 33370163.

[35] E. Sefer, “Probc: joint modeling of epigenome and transcriptome effects in 3d genome,” *BMC Genomics*, vol. 23, no. 1, p. 287, 2022.

[36] E. Sefer, G. Duggal, and C. Kingsford, “Deconvolution of ensemble chromatin interaction data reveals the latent mixing structures in cell subpopulations,” *Journal of Computational Biology*, vol. 23, no. 6, pp. 425–438, 2016. PMID: 27267775.

[37] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, “DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome,” *Bioinformatics*, vol. 37, pp. 2112–2120, 02 2021.

[38] E. Asgari and M. R. K. Mofrad, “Continuous distributed representation of biological sequences for deep proteomics and genomics,” *PLOS ONE*, vol. 10, pp. 1–15, 11 2015.

[39] N. Q. K. Le, E. K. Y. Yapp, Q.-T. Ho, N. Nagasundaram, Y.-Y. Ou, and H.-Y. Yeh, “ienhancer-5step: Identifying enhancers using hidden information of dna sequences via chou’s 5-step rule and word embedding,” *Analytical Biochemistry*, vol. 571, pp. 53–61, 2019.

[40] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, “Modeling aspects of the language of life through transfer-learning protein sequences,” *BMC Bioinformatics*, vol. 20, no. 1, p. 723, 2019.

[41] S. Hu, R. Ma, and H. Wang, “An improved deep learning method for predicting dna-binding proteins based on contextual features in amino acid sequences,” *PLOS ONE*, vol. 14, pp. 1–21, 11 2019.

[42] N. Q. K. Le, Q.-T. Ho, T.-T.-D. Nguyen, and Y.-Y. Ou, “A transformer architecture based on BERT and 2D convolutional neural network to identify DNA enhancers from sequence information,” *Briefings in Bioinformatics*, vol. 22, pp. 1–7, 02 2021. bbab005.

[43] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS’13, (Red Hook, NY, USA), p. 3111–3119, Curran Associates Inc., 2013.

[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

and Short Papers), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[45] A. C. Conibear, "Deciphering protein post-translational modifications using chemical biology tools," *Nature Reviews Chemistry*, vol. 4, no. 12, pp. 674–695, 2020.

[46] S. Ramazi and J. Zahiri, "Post-translational modifications in proteins: resources, tools and prediction methods," *Database*, vol. 2021, pp. 1–20, 04 2021.

[47] T. Kouzarides, "Chromatin modifications and their function," *Cell*, vol. 128, no. 4, pp. 693–705, 2007.

[48] G. Huang and J. Li, "Feature extractions for computationally predicting protein post-translational modifications," *Current Bioinformatics*, vol. 13, no. 4, pp. 387–395, 2018.

[49] H. Yu, C. Bu, Y. Liu, T. Gong, X. Liu, S. Liu, X. Peng, W. Zhang, Y. Peng, J. Yang, L. He, Y. Zhang, X. Yi, X. Yang, L. Sun, Y. Shang, Z. Cheng, and J. Liang, "Global crotonylome reveals cdyl-regulated rpa1 crotonylation in homologous recombination&#x2013;mediated dna repair," *Science Advances*, vol. 6, no. 11, p. eaay4697, 2020.

[50] W. Zhang, X. Tan, S. Lin, Y. Gou, C. Han, C. Zhang, W. Ning, C. Wang, and Y. Xue, "CPLM 4.0: an updated database with rich annotations for protein lysine modifications," *Nucleic Acids Research*, vol. 50, pp. D451–D459, 09 2021.

[51] Z. Li, S. Li, M. Luo, J.-H. Jhong, W. Li, L. Yao, Y. Pang, Z. Wang, R. Wang, R. Ma, J. Yu, Y. Huang, X. Zhu, Q. Cheng, H. Feng, J. Zhang, C. Wang, J.-K. Hsu, W.-C. Chang, F.-X. Wei, H.-D. Huang, and T.-Y. Lee, "dbPTM in 2022: an updated database for exploring regulatory networks and functional associations of protein post-translational modifications," *Nucleic Acids Research*, vol. 50, pp. D471–D479, 11 2021.

[52] H. M. Reddy, A. Sharma, A. Dehzangi, D. Shigemizu, A. A. Chandra, and T. Tsunoda, "Glystruct: glycation prediction using structural properties of amino acid residues," *BMC Bioinformatics*, vol. 19, no. 13, p. 547, 2019.

[53] X. Zhang, A. H. Smits, G. B. van Tilburg, H. Ovaa, W. Huber, and M. Vermeulen, "Proteome-wide identification of ubiquitin interactions using ubia-ms," *Nature Protocols*, vol. 13, no. 3, pp. 530–550, 2018.

[54] I. A. Hendriks and A. C. O. Vertegaal, "A comprehensive compilation of sumo proteomics," *Nature Reviews Molecular Cell Biology*, vol. 17, no. 9, pp. 581–595, 2016.

[55] Y. Kori, S. Sidoli, Z.-F. Yuan, P. J. Lund, X. Zhao, and B. A. Garcia, "Proteome-wide acetylation dynamics in human cells," *Scientific Reports*, vol. 7, no. 1, p. 10296, 2017.

[56] S. Sadhukhan, X. Liu, D. Ryu, O. D. Nelson, J. A. Stupinski, Z. Li, W. Chen, S. Zhang, R. S. Weiss, J. W. Locasale, J. Auwerx, and H. Lin, "Metabolomics-assisted proteomics identifies succinylation and sirt5 as important regulators of cardiac function," *Proceedings of the National Academy of Sciences*, vol. 113, no. 16, pp. 4320–4325, 2016.

[57] D. J. Welsch and G. L. Nelsestuen, "Amino-terminal alanine functions in a calcium-specific process essential for membrane binding by prothrombin fragment 1," *Biochemistry*, vol. 27, pp. 4939–4945, 06 1988.

[58] D. J. Slade, V. Subramanian, J. Fuhrmann, and P. R. Thompson, "Chemical and biological methods to detect post-translational modifications of arginine," *Biopolymers*, vol. 101, no. 2, pp. 133–143, 2014.

[59] D. Umlauf, Y. Goto, and R. Feil, *Site-Specific Analysis of Histone Methylation and Acetylation*, pp. 99–120. Totowa, NJ: Humana Press, 2004.

[60] S. R. Jaffrey, H. Erdjument-Bromage, C. D. Ferris, P. Tempst, and S. H. Snyder, "Protein s-nitrosylation: a physiological signal for neuronal nitric oxide," *Nature Cell Biology*, vol. 3, no. 2, pp. 193–197, 2001.

[61] K. F. Medzihradszky, "Peptide sequence analysis," in *Biological Mass Spectrometry*, vol. 402 of *Methods in Enzymology*, pp. 209–244, Academic Press, 2005.

[62] Y. Ji, Z. Zhou, H. Liu, and R. V. Davuluri, "DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome," *Bioinformatics*, vol. 37, pp. 2112–2120, 02 2021.

[63] E. Asgari and M. R. K. Mofrad, "Continuous distributed representation of biological sequences for deep proteomics and genomics," *PLOS ONE*, vol. 10, pp. 1–15, 11 2015.

[64] M. Heinzinger, A. Elnaggar, Y. Wang, C. Dallago, D. Nechaev, F. Matthes, and B. Rost, "Modeling aspects of the language of life through transfer-learning protein sequences," *BMC Bioinformatics*, vol. 20, no. 1, p. 723, 2019.

[65] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: a universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, pp. 2102–2110, 02 2022.

[66] A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost, "Prottrans: Toward understanding the language of life through self-supervised learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7112–7127, 2022.

[67] N. N. Soylu and E. Sefer, “Bert2ome: Prediction of 2’-o-methylation modifications from rna sequence by transformer architecture based on bert,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 20, no. 03, pp. 2177–2189, 2023.

[68] G. B. Oliveira, H. Pedrini, and Z. Dias, “Temprot: protein function annotation using transformers embeddings and homology search,” *BMC Bioinformatics*, vol. 24, no. 1, p. 242, 2023.

[69] A. Chandra, L. Tünnermann, T. Löfstedt, and R. Gratz, “Transformer-based deep learning for predicting protein properties in the life sciences,” *eLife*, vol. 12, p. e82819, jan 2023.

[70] A. Behjati, F. Zare-Mirakabad, S. S. Arab, and A. Nowzari-Dalini, “Protein sequence profile prediction using protalbert transformer,” *Computational Biology and Chemistry*, vol. 99, p. 107717, 2022.

[71] J. Raad, L. A. Bugnon, D. H. Milone, and G. Stegmayer, “miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs,” *Bioinformatics*, vol. 38, pp. 1191–1197, 12 2021.

[72] N. Q. K. Le, “Leveraging transformers-based language models in proteome bioinformatics,” *PROTEOMICS*, vol. n/a, no. n/a, p. 2300011.

[73] N. Q. K. Le, “Potential of deep representative learning features to interpret the sequence information in proteomics,” *PROTEOMICS*, vol. 22, no. 1-2, p. 2100232, 2021.

[74] Z. Chen, X. Liu, F. Li, C. Li, T. Marquez-Lago, A. Leier, T. Akutsu, G. I. Webb, D. Xu, A. I. Smith, L. Li, K.-C. Chou, and J. Song, “Large-scale comparative assessment of computational predictors for lysine post-translational modification sites,” *Briefings in Bioinformatics*, vol. 20, pp. 2267–2290, 10 2018.

[75] Q. Zou, P. Xing, L. Wei, and B. Liu, “Gene2vec: gene subsequence embedding for prediction of mammalian n6-methyladenosine sites from mrna,” *RNA*, vol. 25, no. 2, pp. 205–218, 2019.

[76] V. Akimov, I. Barrio-Hernandez, S. V. F. Hansen, P. Hallenborg, A.-K. Pedersen, D. B. Bekker-Jensen, M. Puglia, S. D. K. Christensen, J. T. Vanselow, M. M. Nielsen, I. Kratchmarova, C. D. Kelstrup, J. V. Olsen, and B. Blagoev, “Ubisite approach for comprehensive mapping of lysine and n-terminal ubiquitination sites,” *Nature Structural & Molecular Biology*, vol. 25, no. 7, pp. 631–640, 2018.

[77] Y. Liu, A. Li, X.-M. Zhao, and M. Wang, “Deeptl-ubi: A novel deep transfer learning method for effectively predicting ubiquitination sites of multiple species,” *Methods*, vol. 192, pp. 103–111, 2021. Deep networks and network representation in bioinformatics.

[78] S. Pokharel, P. Pratyush, M. Heinzinger, R. H. Newman, and D. B. KC, “Improving protein succinylation sites prediction using embeddings from protein language model,” *Scientific Reports*, vol. 12, no. 1, p. 16933, 2022.

[79] N. Thapa, M. Chaudhari, S. McManus, K. Roy, R. H. Newman, H. Saigo, and D. B. KC, “Deepsuccinylsite: a deep learning based approach for protein succinylation site prediction,” *BMC Bioinformatics*, vol. 21, no. 3, p. 63, 2020.

[80] H. Lv, F.-Y. Dao, Z.-X. Guan, H. Yang, Y.-W. Li, and H. Lin, “Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method,” *Briefings in Bioinformatics*, vol. 22, pp. 1–10, 10 2020. bbaa255.

[81] Y. Qiao, X. Zhu, and H. Gong, “BERT-Kcr: prediction of lysine crotonylation sites by a transfer learning method with pre-trained BERT models,” *Bioinformatics*, vol. 38, pp. 648–654, 10 2021.

[82] Y. Liu, Y. Liu, G. Wang, Y. Cheng, S. Bi, and X. Zhu, “Bert-kgly: A bidirectional encoder representations from transformers (bert)-based model for predicting lysine glycation site for homo sapiens,” *Frontiers Bioinform.*, vol. 2, p. 834153, 2022.

[83] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” pp. 1–22, 2021.

[84] Y. Yang, H. Wang, W. Li, X. Wang, S. Wei, Y. Liu, and Y. Xu, “Prediction and analysis of multiple protein lysine modified sites based on conditional wasserstein generative adversarial networks,” *BMC Bioinformatics*, vol. 22, no. 1, p. 171, 2021.

[85] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[86] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: Synthetic minority over-sampling technique,” *J. Artif. Int. Res.*, vol. 16, p. 321–357, jun 2002.

[87] Z. Zhao, H. Peng, C. Lan, Y. Zheng, L. Fang, and J. Li, “Imbalance learning for the prediction of n6-methylation sites in mrnas,” *BMC Genomics*, vol. 19, no. 1, p. 574, 2018.

[88] J. Pennington, R. Socher, and C. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods*

in Natural Language Processing (EMNLP), (Doha, Qatar), pp. 1532–1543, Association for Computational Linguistics, Oct. 2014.

[89] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," vol. 32, pp. 1–11, 2019.

[90] Z. Lv, H. Ding, L. Wang, and Q. Zou, "A convolutional neural network using dinucleotide one-hot encoder for identifying dna n6-methyladenine sites in the rice genome," *Neurocomputing*, vol. 422, pp. 214–221, 2021.

[91] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, pp. 3150–3152, 10 2012.

[92] C. Sun, L. Huang, and X. Qiu, "Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 380–385, Association for Computational Linguistics, June 2019.

[93] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, (Hong Kong, China), pp. 154–162, Association for Computational Linguistics, Nov. 2019.

[94] T. Jiang, J. Jiao, S. Huang, Z. Zhang, D. Wang, F. Zhuang, F. Wei, H. Huang, D. Deng, and Q. Zhang, "PromptBERT: Improving BERT sentence embeddings with prompts," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 8826–8837, Association for Computational Linguistics, Dec. 2022.

[95] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *CoRR*, vol. abs/1908.10063, pp. 1–11, 2019.

[96] J. Maurício, I. Domingues, and J. Bernardino, "Comparing vision transformers and convolutional neural networks for image classification: A literature review," *Applied Sciences*, vol. 13, no. 9, pp. 1–13, 2023.

[97] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, (Berlin, Heidelberg), p. 280–296, Springer-Verlag, 2022.

[98] M. K. H. Thisanke, L. A. C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *CoRR*, vol. abs/2305.03273, p. 106669, 2023.

[99] Z. Lv, H. Ding, L. Wang, and Q. Zou, "A convolutional neural network using dinucleotide one-hot encoder for identifying dna n6-methyladenine sites in the rice genome," *Neurocomputing*, vol. 422, pp. 214–221, 2021.

[100] S. H. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10s, pp. 200:1–200:41, 2022.

[101] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," vol. 30, pp. 1–11, 2017.

[102] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, pp. 1–10, 2016.

[103] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[104] L. Rutkowski, M. Jaworski, L. Pietruczuk, and P. Duda, "The cart decision tree for mining data streams," *Information Sciences*, vol. 266, pp. 1–15, 2014.

[105] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[106] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 785–794, Association for Computing Machinery, 2016.

[107] F. Chollet *et al.*, "Keras." `https://keras.io`, 2015.

[108] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, (Online), pp. 38–45, Association for Computational Linguistics, Oct. 2020.

[109] C. D. Brown and H. T. Davis, "Receiver operating characteristics curves and related decision measures: A tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 80, no. 1, pp. 24–38, 2006.

[110] T. L. Bailey, "STREME: accurate and versatile sequence motif discovery," *Bioinformatics*, vol. 37, pp. 2834–2840, 03 2021.

[111] X. Li, X. Xiong, K. Wang, L. Wang, X. Shu, S. Ma, and C. Yi, "Transcriptome-wide mapping reveals reversible and dynamic n1-methyladenosine methylome," *Nature Chemical Biology*, vol. 12, no. 5, pp. 311–316, 2016.

[112] D. Zhang and S. Wang, “A protein succinylation sites prediction method based on the hybrid architecture of lstm network and cnn,” *Journal of Bioinformatics and Computational Biology*, vol. 20, no. 02, p. 2250003, 2022. PMID: 35191361.

VITA

Necla Nisa Soylu graduated from Özyeğin University in Computer Science with an honors degree in 2021. Then she started her master's program at the same university in the field of Artificial Intelligence and graduated as the top-ranked student with high honors. Her research interest is in developing transformer-based architectures in Bioinformatics, particularly in RNA and protein modification prediction from biological sequences. One of her research papers was published in the IEEE/ACM Transactions on Computational Biology and Bioinformatics journal, and the other was already published in the Current Bioinformatics journal, while the rest are still under review. She also worked on the TÜBİTAK 1002 and TÜBİTAK 3501 projects as a researcher under the supervision of Assistant Professor Emre Sefer. Besides, she was accepted as a research student at the Korea Institute of Science and Technology – Center for Artificial Intelligence Lab, worked on different projects, and was accepted KSPE 2023 Spring Conference in Jeju Island, Korea, and The 12th International Conference on Biomedical Engineering and Biotechnology (ICBEB 2023) in Macau. She has been awarded Outstanding Teaching Assistant at Özyeğin University for the 2022–2023 academic year by the Graduate School of Engineering and Science. Related publication list:

Journals

1. DeepPTM: DeepPTM Protein Post-translational Modification Prediction from Protein Sequences by Combining Deep Protein Language Model with Vision Transformers- Current Bioinformatics (2023)

2. BERT2OME: Prediction of 2'-O-methylation Modifications from RNA Sequence by Transformer Architecture Based on BERT - IEEE/ACM Transactions on

Computational Biology and Bioinformatics (17 May 2023)

Conferences

1. Comparison of Muscle Characteristics between Healthy Control Group and Sarcopenia Risk Group using Surface Electromyography during Wide-SquatComparison of Muscle Characteristics between Healthy Control Group and Sarcopenia Risk Group using Surface Electromyography during Wide-Squat - KSPE 2023 Spring Conference (South Korea) (11 May 2023)

2. sEMG-based Sarcopenia Risk Classification using Empirical Mode Decomposition and Machine Learning Algorithms - The 12th International Conference on Biomedical Engineering and Biotechnology - ICBEB 2023 (Macau) (17 Nov 2023)