A review of computational approaches to predict Enhancer-Promoter Interaction and Pseudouridine Modification

Sourajyoti Datta
Department of Computer Science
Technische Universität Kaiserslautern
Kaiserslautern, 67663, Germany
Email: datta@rhrk.uni-kl.de

Muhammad Nabeel Asim
Deutsches Forschungszentrum für Künstliche Intelligenz
& Technische Universität Kaiserslautern
Kaiserslautern, 67663, Germany
Email: muhammad_nabeel.asim@dfki.de

***Abstract*—In DNA, the regulation of gene expression occurs during the transcription process of creating RNA, by the interaction between Enhancers and Promoters, through the favorable folding of the genome. This activity is essential to all life, providing cell line differentiation and adaptability to the organisms. Once RNA is generated, it undergoes Pseudouridine Modification, one of the abundant forms of RNA editing. It has functional impacts, contributes to structural maintenance of RNA, and has disease related implications. Although there exists numerous wet lab approaches for identifying these mechanisms, they are expensive, skill and labor intensive, and have theoretical and practical limitations. With the advancement of high throughput methods capturing high resolution data, several in-silico computational methods have been proposed which provide economical, fast and accurate alternatives to identify these mechanisms in DNA and RNA sequences. In this study, 10 computational methods for the identification of Enhancer-Promoter Interaction using data from DNA sequences, genomics, epigenomics and 3D chromatin interaction have been summarized. Furthermore, 8 methods for the prediction of Pseudouridine modification utilizing data from RNA sequences and derived physico-chemical features, genomics and mRNA structural contexts have also been summarized.**

1. Introduction

Cells are the basic biological blocks of all living organisms, forming the structural and functional unit. Based on the type of organisms and their biological roles, cells can be categorised into two major types. ***Prokaryotic cells***, that form simple single-celled organisms which do not contain a nucleus, or other membrane-bound organelles. ***Eukaryotic cells***, unlike the former, are compartmentalized and form complex multi-cellular organisms, containing membrane-bound organelles with specific functions, most importantly a nucleus.

Both types of cell contain genetic material that contain diverse information about the organisms themselves, and of different biological activities that takes place inside living organisms. Genetic information inside the cell is preserved in the nucleic acids, which are one of the macro-molecules essential for life. One such molecule is ***Deoxyribonucleic acid (DNA)***, which consists of two strands of polynucleotide chains, that are twisted around each other forming a double-helix structure [31]. For all known life forms, the genetic instructions are contained in the DNA, which define form, function, growth, and reproduction for the organism [31]. Each nucleotide in the chains is composed of one of the four nucleobases (adenine [A], cytosine [C], guanine [G], or thymine [T]), and a phosphate-deoxyribose backbone [31]. Another such molecule is ***Ribonucleic acid (RNA)***, a single stranded polynucleotide chain folded on itself. It is essential for various genetic activities, like coding, decoding, expression and regulation of genes. Each nucleotide in the chain is composed of one of the four nucleobases (adenine [A], cytosine [C], guanine [G], or uracil [U]), and a phosphate-ribose backbone [31].

The central dogma of molecular biology states that DNA is copied to create RNA (***DNA Transcription***), and then RNA is used to produce proteins (***Translation***). In *DNA Transcription*, segments of DNA molecules are replicated into RNA molecules by RNA polymerase, thereby constructing messenger RNA (mRNA) and different non-coding RNAs (ncRNA). During transcription, all occurrences of the the nucleobase thymine (T) in the DNA segment is replaced by uracil (U). After Transcription, *Translation* occurs in the cell's nucleus, where proteins are synthesized from the RNA molecules. This entire process is *Gene Expression*.

Within cells, the regulation of genetic expression occurs during the transcription process, called ***Transcription Regulation***. This regulation is vital to all organisms, since it provides the mechanism to be versatile and adapt through cellular differentiation and morphogenesis, resulting in the wide diversity of cells. Regulation occurs due to Transcription Factors (TFs) and Cis-Regulatory Elements (CREs) working together to adjust the RNA being produced. TFs are proteins which bind with certain sequences of DNA causing regulation of gene expression for the specific gene. Transcription factors may act as activators, repressors, or

both. CREs are segments of a nucleic acid molecule which can increase or decrease the expression of specific genes within an organism, and play essential roles in development [19]. One such CRE, **Promoters**, are sequences of DNA to which the TF proteins bind to initiate *Transcription*, thus defining the *Transcription Start Sites (TSS)*. Another CRE, **Enhancers**, are non-coding segments of DNA that comprises of various activator and repressor binding sites. Through DNA looping (favorable folding of the genome in the three-dimensional space), active enhancers interact with specific promoters, based on the specificity of the latter (***Enhancer Promoter Interaction - EPI***), as visually depicted in Figure 1. Enhancer–promoter interactions (EPIs) play key roles in transcriptional regulation, and has implications in disease progression.

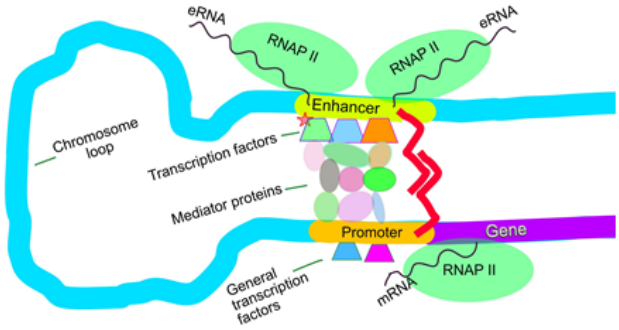

Figure 1: Illustration of Enhancer-Promoter Interaction [30]

RNA Modification (or *RNA editing*) is a change in the constituent nucleotides of the RNA molecule, that occurs after the creation of the RNA. Cells use this process to perform specific changes to certain sequences of nucleotides in the RNA molecule. This process affects different kinds of RNA, like Messenger RNA (mRNA), Transfer RNA (tRNA) and Ribosomal RNA (rRNA). During this process, changes like insertion, deletion, and substitution of the nucleobases occur within the RNA molecule. This process affects different aspects of RNAs, such as the activity of the RNA, their localization, and their structural stability, amongst others. It has also been linked to various diseases in human beings as well [27].

Pseudouridine (Ψ) Modification is the most abundantly occurring form of modification in RNA. It occurs at *Uridine* sites, where the fifth nucleotide base of RNA (5-ribosyluracil) is generated from the isomerization of a uridine base. This process of isomerization (*Pseudouridylation*) is a post-transcriptional mechanism, catalyzed by *Pseudouridine Synthases* (PUS). Another way of catalyzation depends on the H/ACA nucleic acid protein complex [17]. The binding process occurs in specific contexts and with the conformation of nucleotides [12]. During the reaction, Ψ synthase (PUS) enzyme cleaves a uridine residue from its original nucleoside (sugar) to add a Ψ residue, an isomer of uridine, by rotating a bonding angle along the N3–C6 axis at 180°. Finally, a new bond between the base's 5-carbon and the 1-carbon of the nucleoside is formed, as visually depicted in Figure 2.

Ψ-modification has significant functional and disease-related implications, as an example, for some types of cancers, it provides important biomarkers [12]. Changes of Ψ in rRNA can affect the susceptibility of bacteria to antibiotics. Furthermore, it plays a vital role in transcriptional activities due to its contribution in maintaining the functional structure of tRNA and the gene regulation machine (e.g., spliceosome) [11]. It can accelerate RNA-RNA or RNA-protein interaction and spliceosome assembling for producing functional mRNA [16]. However, the role of Ψ-modification in most RNA systems, its biological functions, and action mechanisms are yet to be identified [16]. Hence, gaining a holistic understanding of Ψ-modification can contribute to the development of drug discovery and gene therapies.

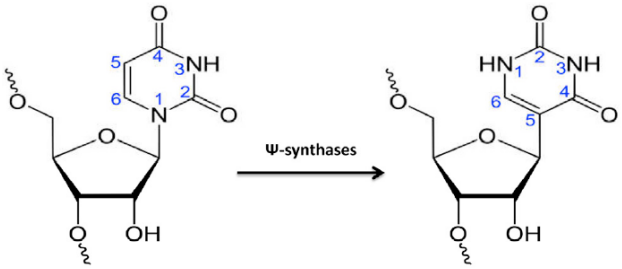

Figure 2: Illustration of Pseudouridine (Ψ) Modification [16]

2. Motivation for computational methods to predict Enhancer Promoter Interaction and Pseudouridine modification

For the two specific genetic mechanisms under study, Enhancer-Promoter Interaction (EPI) and Pseudouridine (Ψ) Modification, there exists numerous experimental approaches. As discussed earlier, the exploration of three-dimensional chromatin interaction provides important insight for gene regulation, cell differentiation, disease development, etc.

Due to the development of advanced high-throughput experimental approaches over decades, such as chromosome conformation capture-based (3C) [35], its variants of Hi-C (high-throughput chromosome conformation capture) [36] and ChIA-PET (chromatin interaction analysis by paired-end tag) [32], circularized chromosome conformation capture (4C) [33], carbon-copy chromosome conformation capture [34], the study of chromatin interactions has made significant progress. Hi-C and ChIA-PET can measure the entire genome DNA-DNA interactions, however, the genomic resolutions are often low, varying from few kilobases to tens of thousands bases. High-resolution Hi-C and ChIA-PET data are only available for a limited number of cell lines, and their acquisition is costly, time consuming, laborious and affected by theoretical limitations. But, the study of EPIs often require very high ($<$ 10 kilobases) resolution data, since resolution of data directly impacts the

effectiveness and accuracy of identification. The exponential data growth with increasing depth also brings along new analytical challenges. Due to the limitations of experimental approaches, the number of available experimental data of EPIs is still limited.

Similarly, experimental approaches for Ψ-Modification has been well established. High-throughput sequencing approaches developed for profiling transcriptome-wide distribution of Ψ, such as Pseudo-seq, can identify Ψ sites on a large scale at a single nucleotide resolution. However, it requires high sequencing depth and multiple biological replicates in order to do so accurately [12] [15]. Another method, namely Site-specific Cleavage And Radioactive labeling followed by Ligation-assisted Extraction and Thin-layer chromatography (SCARLET) identifies candidate (Ψ) sites, with the benefit of quantitatively detecting the extent to which a particular (Ψ) is modified. However, these site-specific mappings require prior knowledge of the Ψ-containing sequences, thus preventing an unbiased detection approach. Further more, transcriptome-wide Ψ-sequencing approaches using induced termination of reverse transcription, and quantitative mapping systems have also been developed.

Although successful, these experimental approaches in wet-labs can be technically challenging, time-consuming and have high false-negative rates. Especially, since detailed experimental exploration and acquisition of quality information is still a very expensive and difficult procedure. Moreover, these approaches are often labor-intensive, requiring skilled experience and maintenance. Furthermore, due to theoretical limitations, these methods provide only limited coverage. Due to the increasing availability of genomics and proteomics samples produced in the post-genomics era, accurate and robust computational approaches are desired, as an addition to experimental efforts, or even as an efficient and low-cost alternative [16]. As a result of huge amount of data generated by various high-throughput sequencing technologies, numerous computational methods have emerged for knowledge discovery targeting various biological issues, working with sequence and epigenomic data, implementing traditional machine-learning algorithms and deep-learning techniques.

3. An outlook towards advancements in Machine Learning and Deep Learning

Traditionally, *Machine Learning (ML)* and *Statistical* approaches implementing algorithms like Support Vector Machines (SVM), Extreme Gradient Boosting, Random Forest, Decision Trees, Logistic and Linear Regression, and other Statistical analyses techniques, has been developed for the processing of DNA/RNA sequences, and the genomic features extracted from these sequences or experiments. Although successful to an extent, it encouraged research and development in more advanced methods. Due to the suitability of *Natural Language Processing (NLP)* techniques, like *Word2vec* for sequence data, it has also sparked the development and use of NLP for feature extraction. However, all of these prior techniques require carefully hand-crafted, relevant and informative input features for the algorithms to work upon and generate good models. Building such features require extensive domain knowledge, thus requiring further labor and expertise. It can also become a cause of limitation for the performance of these models, since there is always a probability that knowledge of specific important features might not be available.

Deep Learning (DL) techniques provide several advantages in comparison to traditional ML approaches. They have been shown to outperform ML in numerous fields employing computational strategies. Furthermore, deep learning networks remove the need for hand-crafted feature engineering with their inherent ability to learn different kinds of important features, simple or complex, directly from the data itself. As an example, *Convolutional Neural Networks (CNN)* make use of kernels in convolution operations to learn important spatial features. Similarly, extraction of features from sequential or temporal data is performed using *Recurring Neural Networks (RNN)*. Prediction and Classification tasks are often performed using simpler, *Fully-Connected Neural Networks (FCNN)*. Today, most computational problems require application of more than one kind of Neural Network within the same architecture, giving rise to a host of *Hybrid Models*, where CNNs, RNNs and FCNNs are stacked to create a complex neural pipeline for the task at hand.

Hence, there is a boost in the use of DL techniques in genomic studies. However, the application of DL approaches in biological research has limitations due to the sheer complexity of the algorithms implemented (namely, neural networks). One drawback is that the explainability of the models are virtually impossible, and hence, such computational models might not be useful for any form of descriptive analysis. In addition, the appropriate method can be difficult to decide upon given the diversity of input data and a variety of software requirements. Furthermore, some computational approaches could be difficult and resource-intensive to reproduce and apply in extended research.

4. Existing computational methodologies for Enhancer Promoter Interaction and Pseudouridine modification prediction

Computational algorithms based on Machine Learning or Deep Learning techniques cannot process DNA or RNA sequence data directly, due to their inherent dependency on the use of numerical values. Hence, for data driven classification tasks, such as the ones discussed here, the algorithms need to be provided with numerical or statistical representations of the sequence data, for the algorithm to be able to learn discriminative patterns in the data. To that extent, both the classification approaches are segregated into two distinct stages, feature extraction and classification. The first stage is feature extraction where the aim is to generate an effective numerical or statistical representation that contains semantic data of the amino acid sequences of DNA or RNA, that

provides discriminative information for the classifier. In addition to determining important features from sequence data, other sources of data, such as genomic profiles, epigenomic profiles, 3D-chromatin interaction profiles amongst others are also employed to extract discriminative information for the sequences. In the second stage, algorithms are trained to identify and model the discriminative features based on their class labels, from the statistical vectors generated in the previous stage, to perform the final classification task. These classifiers are either developed from scratch, or employ transfer learning opportunities by optimising existing classifiers models. In this section, diverse feature encoding methods and DL architectures that have been used in existing research to generate suitable representations, along-with suitable ML and DL algorithms employed as classifiers have been presented.

4.1. Enhancer Promoter Interaction

Computational methods for EPI prediction focus extensively on relationships between regulatory elements. One key component of detecting functional EPI is to investigate whether the related CREs are activated under certain conditions, given genomic, epigenomic, and transcriptomic profiling data which provides abundant resources for the identification of active CREs and their interactions using supervised learning methods [20]. This provides the sequence profiling features required for EPI identification. Another key component is to identify reduced spatial distance due to the formation of the chromatin loop, which is a critical property of functional EPIs [20]. This provides the 3d chromatin structure features utilized for EPI prediction. With 3C-based techniques and its high throughput variants in numerous cells, it is possible to create models using supervised learning methods to describe EPIs across the entire genome and identify potential EPIs. With the evolution of 3D genomic techniques, genome wide detection of EPIs have gained further traction, by integrating specific epigenomic profiles and 3D genomic data into the computational models.

One of the ways EPI prediction is approached is by utilising DNA sequences, represented in the form of One-Hot Encoded (OHE) vectors, fed as input features directly to the models. In a second approach, the DNA sequences are utilized to extract sequence profiling information and other genomic/epigenomic features which are then fed as features to the model. As a third approach, structural context of the DNA sequences, such as 3D chromatin interaction caused by the 3D chromatin-loop formation of the DNA strand, are generated and fed as input feature to the models. However, some recent approaches also combine two or more of these approaches to build the feature vectors, so as to ensure availability of more information for a better discriminative model to be learnt. The drawback of some of the previous approaches, is the lack of any kind of semantic information that can be generated from the sequences themselves, which could play a role in EPI detection. With the advent of NLP, embedding schemes, such as word2vec and its derivative dna2vec, have been incorporated into the computational models to achieve better classification performance.

From an architectural perspective, in most recent researches, enhancer and promoter sequence fragments are treated separately, and two numerical vectors are generated for each pair. Then, these vectors are passed through CNN based spatial feature extraction step separately, to learn information about both the enhancer and promoter separately from each other as described in Figure 3. Often, RNN layers are also implemented to learn temporal features from the sequence derived features like in Figure 3. Finally the learned features are concatenated and passed on to the classifier stage. In contrast to that, few approaches instead accumulate all features from the enhancer and promoter into one vector which is used for feature extraction, and subsequently classification, especially when non-sequence features are being employed.

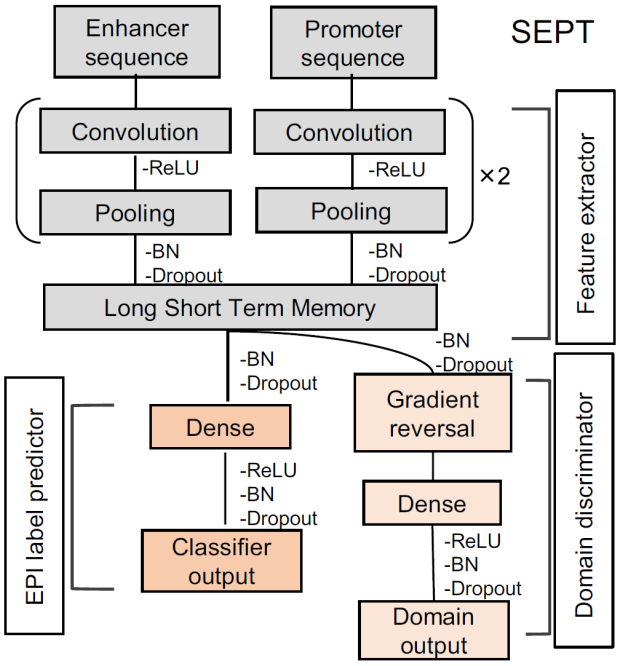

Figure 3: Architecture of SEPT [2], proposed by Jing et al.

The methodology proposed by Zhuang et al. [6] uses OHE vectors, and builds a hybrid multi-path CNN to learn spatial features from the OHE vectors of promoters and enhancers separately. The learned features are then passed through a single layered FCNN for classifying the EP pair as interacting or not. Singh et al.'s proposed methodology, SPEID [10], builds on the previous architecture, by extracting features for the enhancer and promoter separately using CNNs. Then, the extracted features are passed through an RNN layer to capture temporal information. Finally, the temporal features are passed through a single layered FCNN for classification. Jing et al.'s proposed methodology, SEPT [2], builds on this architecture furthermore by adding a Gradient Reversal Layer (GRL) with transfer learning mechanism, called the domain discriminator, to adapt the

entire feature extraction network (CNN layer followed by RNN layer) to remove cell line specific features and retain features independent of cell line for the final classification layer as described in 3.

In contrast, Zeng et al.'s technique, EP2vec [9], is an unsupervised NLP based word embedding technique with attention mechanism, to extract fixed-length feature representation of the DNA sequences from variable-length sequences. It is applied on enhancer and promoter sequences separately to generate two feature vectors. Finally, these two feature vectors are combined and an ensemble ML algorithm, Gradient Boosted Regression Trees (GBRT), is used to train the classifier model. EPnet, proposed by Wang et al. [1], promotes the use of NLP further by incorporating *DNA2vec*, which is based on the *word2vec* architecture, to embed fragments of DNA sequences into a 100-dimensional continuous vector space using the entire human genome sequence as the learning corpus. Each of the two 100-d vectors for an EP pair are passed separately through a CNN layer with a multi-path pooling layer to learn spatial features, which are finally concatenated and passed through a bi-directional RNN layer to complete the feature extraction step. Finally, the extracted temporal features are fed into a FCNN layer for classifying the EP pair.

Apart from sequence based techniques, various methods have also been developed which are based on genomic and epigenomic features for modelling and predicting EPIs. Gao et al.'s proposed method, EAGLE [8], defines six genomic features, Enhancer activity and gene expression profile correlation (EGC), Gene score (GS), Distance (DIS), Enhancer window signal (EWS), Gene window signal (GWS) and Weight of enhancer-enhancer correlations (WEEC) derived for each EP pair. The features are then fed to an ensemble learning algorithm, Adaboost [], for classifying the EP pair as interacting or not. Another technique, EPIP, proposed by Talukder et al. [5], uses distance between an EP pair computed using normalized Hi-C contact matrices, a conserved Synteny score, the correlation of epigenomic signals of an EP pair, and 14 additional types of data that contains DNase-seq and ChIP-seq data for histone modifications. These features are fed to an ensemble ML algorithm implementing Decision Trees (DTs) for the modelling the classification of EP pairs. Belokopytova et al. proposes another technique, 3DPredictor [4], which utilises ChIP-seq profiles that describe chromatin binding of architectural proteins or histone modifications, RNA-seq profiles that describe gene expression levels, and genomic distance of three dimensional contacts as epigenetic features. The features are then modelled using an ensemble Gradient Boosting Regression algorithm to predict EPIs.

Building on the success of using sequence based features, and genomic & epigenomic features, further development sparked by incorporating both the aforementioned types of features to predict EPIs. DeepTACT, proposed by Li et al. [7], is a bootstrapped deep learning approach utilizing OHE DNA sequences and chromatin accessibility scores derived from DNase-seq. Two CNNs are used separately to extract spatial information from the two OHE vectors of enhancer and promoter of an EP pair. Similarly, two more CNNs are used to extract epigenomic information separately from the chromatin accessibility scores of enhancer and promoter of an EP pair. Then, the output of the four CNNs

TABLE 1: **Overview of recent EPI prediction techniques**

Year	Author and Method	Features	Classifiers
2021	Wang et al. - EPnet	DNA sequence feature: - DNA2vec	Hybrid model: Multi-path multi-stage CNN with RNN, and multi-layer FC
2020	Jing et al. - SEPT	DNA sequences - One-hot-encoded (OHE)	Hybrid model: Multi-path CNN with RNN, and FC layer
2020	Liu et al. - EPIHC	DNA sequence features: - DNA2vec - Communicative learning Genomic features: - Open chromatin, - DNA methylation, - Gene expression - ChIP-seq peaks	Multi-path Architecture, CNN with MLP
2019	Belokopytova et al. - 3DPredictor	Epigenetic features: - RNA-seq profiles - ChIP-seq profiles - Distance measures	Ensemble ML algorithm: - Gradient Boosting Regression
2019	Talukder et al. - EPIP	- Normalized Hi-C contact matrices - Distance measures - DNase-seq and ChIP-seq for histone modification	Ensemble ML algorithm implementing Decision Trees
2019	Zhuang et al.	DNA sequences - OHE	Transfer Learning approach using Multi-path CNN with FC layer
2019	Li et al. - DeepTACT	DNA sequence features: - OHE Epigenomic information: - DNase-seq Chromatin accessibility	Bootstrapping ensemble of Hybrid models: Multi-path CNN with attentive RNN, and FC layer
2019	Gao et al. - EAGLE	Genomic features: - EGC, GS, DIS, EWS, GWS, WEEC	Ensemble boosting algorithm: - AdaBoost
2018	Zeng et al. - EP2vec	DNA sequences - EP2vec embedding	Ensemble ML algorithm: - Gradient Boosted Regression Trees classifier
2018	Singh et al. - SPEID	DNA sequences - OHE	Hybrid model: Multi-path CNN with RNN, and FC layer

are merged and passed to a bi-directional RNN layer with attention mechanism to capture temporal information, which are finally fed to a FCNN layer for classifying as interacting or not. In addition to predicting EPIs, the method is also able to identify promoter-promoter interactions as well. Another method, EPIHC, proposed by Liu et al. [3], uses sequence derived feature vectors with DNA2vec embedding, passed through a CNN layer and a custom defined *Communicative learning* module to learn the communicative features of EP pairs. In parallel, measures of open chromatin, DNA methylation, gene expression, ChIP-seq peaks for transcription factors, architectural proteins, and modified histones are fused for EP pairs to learn the genomic features. Both type of features are merged, and finally processed through a Multi-Layer Perceptron (MLP) to model EPI classification.

4.2. Pseudouridine modification

Computational methods for Ψ-modification prediction focus on RNA sequences and their Ψ-sites. Numerous methods employ the use of these sequences directly, whereas others employ the use of the nucleotide composition surrounding the Ψ-sites, and their physical and chemical properties. Even structural context of RNA fragments are employed. Along with these, several specialized feature extraction techniques are employed, such as pseudo-k-tuple nucleotide composition (PseKNC) type I and II, n-gram and multivariate mutual information (MMI) amongst others. Some also focus on the use of genomic features, such as conservation, gene annotation, and mRNA binding, etc. Different kind of feature selection methods based on importance are implemented as well, such as forward feature selection. However, since these features require specialized skills and expertise to build, more advanced methods implementing CNN based feature extraction techniques developed. Furthermore, structural information of RNAs were also incorporated.

One of the earlier methods proposed by Nguyen et al., iPseU-NCP [18], employs the use of Nucleotide Chemical Properties (NCP) encoding scheme, a vector of 63 dimensions derived from the RNA sequence. The feature set was used to train an ensemble classification algorithm, Random Forest, as the classifier. Another methodology, XG-PseU, proposed by Liu et al. [17], builds further by collating together various kinds of RNA sequence derived features along with the OHE vectors of the RNA sequences. The derived feature set incorporates Nucleotide Composition (NC), Dinucleotide Composition (DNC), Trinucleotide Composition (TNC) and Nucleotide Density (ND) encoding schemes. Through two feature selection techniques, forward and incremental, the optimal feature set is determined, which is then passed to an eXtreme Gradient Boosting (XGboost) algorithm for the final classification task. iPseU-Layer, proposed by Mu et al. [14], merges together five kinds of feature extraction methods which are DNC, TNC, NCP, Nucleic Acid Composition (NAC) and Position-specific Trinucleotide Propensity based on Single-Strand (PSTNPss), derived from RNA sequences. Important features are determined using the Correlation based Feature Selection (CFS)

TABLE 2: **Overview of recent Ψ-site prediction techniques**

Year	Author and Method	Features	Classifiers
2021	Aziz et al. - Deep learning	RNA sequence data - One-hot encoded - Merged-seq	Multi-stage CNN with Multi-layer FC
2020	Khan et al. - MU-PseUDeep	RNA sequence Structure context of mRNA	Multi-path multi-stage CNN with FC
2020	Lv et al. - RF-PseU	RNA sequence data, and derived features: - Nucleotide Binary Profiles - Accumulated Nucleotide Frequency - Electron-Ion Interaction Pseudopotentials - Enhanced Nucleic Acid Composition - Xmer k-Spaced Ymer Composition Frequency	- Feature selection using Light Gradient Boosting Machine (LGBM) - Random Forest
2020	Mu et al. - iPseU-Layer	RNA sequence data (OHE), and derived features: - Nucleic Acid Composition (NAC) - Di-Nucleotide Composition (DNC) - Tri-Nucleotide Composition (TNC) - Position-specific trinucleotide propensity based on single-strand (PSTNPss) - Nucleotide Chemical Property (NCP)	Multi-layer ensemble of Machine Learning algorithms: - Naive bayes - Bayes network - Linear logistic regression - Sequential minimal optimization - C4.5 decision tree - Random forest
2020	Song et al. - PIANO	RNA Sequence derived features: - NCP - Position-specific nucleotide propensity (PSNP) - Clustering information Genome derived features: - WHISTE - PIANO	Support Vector Machine (SVM) with Radial Bias kernel
2019	Tahir et al. - iPseU-CNN	RNA sequence data - OHE	Multi-stage CNN with Multi-layer FC
2019	Liu et al. - XG-PseU	RNA sequence data (OHE), and derived features: - NAC - DNC - TNC - Nucleotide Density - NCP	eXtreme Gradient Boosting (XGboost)
2019	Nguyen-Vo et al. - iPseU-NCP	RNA sequence derived feature: - NCP	Random Forest

algorithm. Then, the subset of selected features are passed to an ensemble model based on six ML algorithms, namely, Native Bayes (NB), Bayes Network (BN), linear Logistic Regression (LR), Sequential Minimal Optimization (SMO), C4.5 Decision Tree (C4.5), and random forest (RF), for the extraction and fusion of features. Finally, the generated features are passed to an ensemble model based on Random Forest to perform the classification task. Another method proposed by Lv et al., RF-PseU [13], widened the use of various features by introducing Nucleotide Binary Profiles, Accumulated Nucleotide Frequency (ANF), Electron-Ion Interaction, Pseudopotentials (EIIP), enhanced Nucleic Acid Composition (eNAC), and Xmer k-Spaced Ymer Composition Frequency derived from the RNA sequences. Then, features were ranked and thresholded by their importance values calculated using the Light Gradient Boosting Machine (LGBM) algorithm. Finally the features are fed to a Random Forest based ensemble model to perform the classification task.

Building on the success of using RNA sequence based features, a hybrid feature set, bringing genomic features into the mix was employed by PIANO, proposed by Song et. al [15]. It used three different RNA sequence encodings, Position-specific nucleotide propensity (PSNP), NCP and Clustering information. Furthermore, WHISTLE, a genomic encoding scheme with 35 features, along with 7 more genomic features defined by PIANO were also used. The entire feature set was used to train a classification algorithm using Support Vector Machine (SVM) with Radial Bias kernel.

With the advent of DL techniques, the possibility of learning features directly from data, without having to handcraft them, was utilized by iPseU-CNN, proposed by Tahir et al. [16]. It represented the RNA sequences only as OHE vectors, which were then passed through a two stage CNN to learn spatial features from the input feature. Then, a two stage FCNN is used to classify the site.

Apart from purely RNA sequence based approaches, secondary structural context of RNAs were also incorporated. MU-PseUDeep [12], proposed by Aziz et al., introduces the Merged-Seq representation, incorporating RNA sequences merged with secondary structures of the RNA fragment predicted using *RNAfold* into a 12 dimensional OHE vector for each sequence. Then, a multi-layered CNN is used to extract spatial features from the Merged-seq representation, which is then passed onto a two-stage FCNN to perform the final classification of sites. Another method, proposed by Aziz et al. [11], uses two CNNs to extract spatial features from the OHE vectors of the RNA fragments. Similarly, it also uses two more CNNs to extract features from the OHE vectors of the secondary structures of the RNA fragment predicted using *RNAshapes*. All the features generated from the four CNNs are concatenated and passed to a single FCNN layer to perform the classification task.

5. Performance analysis of Enhancer Promoter Interaction and Pseudouridine modification prediction

The evaluation of performance of the multitude of methods discussed in Section 4, for both EPI and Ψ-site prediction, employ numerous statistical measures. These measures are often evaluated on multiple datasets derived from DNA and RNA sequences of various cell lines of same or different species. To validate the performance of these models across iterations, strategies such as k-fold cross validation are implemented, and often the average across the iterations is reported for each metric. Furthermore, certain methods for EPI prediction distinguish the performance of the models between cell-line specific prediction models and global prediction models. However, since the classification tasks surveyed in this study often encompass imbalanced data, the choice of appropriate measures is of utmost importance to make correct analysis of these models.

For classification tasks, there exists a variety of performance metrics which can be used to evaluate the models. Some of these metrics are derived using a *confusion matrix* of the prediction generated by a classifier model, containing two success terms defined by True Positives (TP) and True Negatives (TN), and two error terms defined by False Positives (FP) and False Negatives (FN) as described in Table 3. *Accuracy* is a measure of the proportion of successful predictions amongst all the instances examined. *Sensitivity*, also known as *Recall* or *True Positive Rate* (*TPR*), is the proportion of successful positive predictions amongst all the instances that are actually positive. Similarly, *Specificity*, also known as *True Negative Rate* (*TNR*), is the proportion of successful negative predictions amongst all the instances that are actually negative. *Precision*, also known as *Positive Predictive Value* (*PPV*), is the proportion of successful positive predictions amongst all the instances that are predicted as positive by the classifier. *F1 score* is the harmonic mean of Precision and Recall.

TABLE 3: **Confusion Matrix**

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Furthermore, metrics predominantly used in classification tasks also include AuROC and AuPR. A Receiver Operating Curve (ROC) is a plot of TPR vs. FPR calculated at various probability threshold settings for a classifier model, depicting the trade-off between the TPR and FPR. The *Area under the Receiver Operating Curve* (AuROC) is equivalent to the probability that the classifier will rank a random positive instance higher than a random negative instance, and hence is a metric that can be used to compare classifier models. Similarly, the Precision-Recall Curve plots the trade-off between the Precision and Recall at various probability threshold setting for a classifier model, and the *Area under the Precision-Recall Curve* (AuPR) is used to

TABLE 4: **Performance comparison of recent cell-line specific EPI prediction techniques**

Year	Author and Method	Datasets	Evaluation Metrics				
			Sn. (%)	PPV	F1	AuROC	AuPR
2019	Talukder et al. - EPIP	GM12878	0.8885	0.9946	0.9385	0.9873	0.9835
		HUVEC	0.9810	0.5387	0.6955	0.9920	0.9816
		KBM7	0.9836	0.7377	0.8431	0.9892	0.9826
		IMR90	0.9938	0.8714	0.9286	0.9979	0.996
		K562	0.9881	0.8421	0.9093	0.9955	0.9921
		NHEK	0.9879	0.6275	0.7675	0.9942	0.9791
		HMEC	0.9889	0.6256	0.7664	0.9951	0.9905
2019	Zhuang et al.	GM12878	-	-	-	0.98	0.92
		HeLa-S3	-	-	-	0.98	0.95
		HUVEC	-	-	-	0.99	0.95
		IMR90	-	-	-	0.98	0.92
		K562	-	-	-	0.99	0.95
		NHEK	-	-	-	0.99	0.96
2019	Gao et al. - EAGLE	GM12878	-	-	-	0.9403	
		K562	-	-	-	0.9565	0.9528
		MCF-7	-	-	-	0.9325	0.9198

determine the quality of the classifier model. Although both these measures are considered to be better for classification tasks compared to the previously defined metrics, AuPR is considered to be even better at determining classifier performance on very imbalanced datasets since it can be focused on the minority class.

Apart from these confusion matrix based metrics, several other metrics have also been employed. *Pearson Correlation Coefficient* (PCC) is a measure of linear correlation between two sets of features or data, as the ratio between the covariance of two variables and the product of their standard deviations. *Matthews Correlation Coefficient* (MCC) is the correlation coefficient between the observed and predicted binary classifications represented as vectors. It can also be calculated directly from the confusion matrix. *Stratum-adjusted Correlation Coefficient* (SCC) is another statistic that assess the reproducibility of Hi-C matrices. Furthermore, various standard statistical error metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Relative Error (MRE) are also analyzed.

5.1. Performance comparison of Enhancer Promoter Interaction predictors

For EPI prediction techniques, the performance of the classifier models are mostly evaluated on the original dataset that was processed by *TargetFinder* [22], consisting of data from six human cell lines: GM12878, HUVEC, HeLa-S3, IMR90, K562 and NHEK illustrated in Tables 5 and 4. Additionally, some studies also take the KBM7 cell line into consideration as well. The genomic locations in these sequences are annotated using the human genome reference *hg19* [2] [1]. Furthermore, the enhancer and promoters are identified using various segmentation-based annotations from ENCODE Segway [23] and ChromHMM of RoadMap Epigenomics [24]. For each of the cell lines, the data comprises of enhancer-promoter pairs that are annotated using high-resolution genome-wide measurements of chromatin contacts in the respective cell line based on Hi-C matrices [10]. The annotations are labelled such that the interacting pairs are denoted as positive, and the non interacting pairs as negative, thus defining the positive class as the minority, which helps generate better evaluation metrics (for example, AuPR) by focusing on the minority class. Certain methods also address the imbalance by applying data augmentation methods on the positive class instances. In contrast, the method DeepTACT reports performance on Promoter capture Hi-C (PCHi-C) data in total B cells (tB), monocytes (Mon), fetal thymus (FoeT), total CD4+ T cells (tCD4), naive CD4+ T cells (nCD4), total CD8+ T cells (tCD8) from the study conducted by Javierre et al. [25], since DNase-seq data are needed in the modeling process and available only for these 6 cell lines [7]. Hence, DeepTACT cannot be qualitatively compared with the other methods under study here.

The most frequently reported metrics for EPI prediction using global models are AuROC and AuPR. For the GM12878 cell line, SPEID reports the best AuROC value ($\approx$0.99) and the lowest is reported by SEPT (0.61) as described in Table 5. However, the point of interest here is that SEPT trains the classifier model on a dataset comprising all the cell-lines but leaving out the cell-line for which the performance values are predicted and calculated upon, whereas the other methods generally combine data from all cell-lines to build the global models. Similarly, for all the other cell lines, SPEID reports the best AuROC values of $\approx$0.99 across the board. However, SEPT did showcase that using the leave out cell-line for train-test split, the performance of SPEID falls significantly (from $\approx$0.99 to $\approx$0.65) compared to the original performance numbers reported by Singh et al. Considering reported AuPR values, SPEID and EPIP seems to have considerably better performance (in the range of 0.98-0.99) than the other methods, as observed in Table 5. However, two methods, EP2vec and 3DPredictor do not report these numbers and hence cannot be included in the comparative analysis. Additionally, Li et al. compared the performance of their method DeepTACT with SPEID using the 6 different cell-lines (tB, Mon, FoeT, tCD4, nCD4,

TABLE 5: **Performance comparison of recent global EPI prediction techniques.** The "+" means that the measurement value has been extrapolated from charts provided by the authors, in the absence of absolute values.

Year	Author and Method	Datasets	Evaluation Metrics					
			Acc.(%)	Sn.(%)	PPV	F1	AuROC	AuPR
2021	Wang et al. - EPnet	GM12878	-	-	-	-	0.944	0.797
		HUVEC	-	-	-	-	0.922	0.712
		HeLa-S3	-	-	-	-	0.938	0.813
		IMR90	-	-	-	-	0.945	0.792
		K562	-	-	-	-	0.942	0.811
		NHEK	-	-	-	-	0.969	0.895
		HMEC	-	-	-	-	-	-
2020	Jing et al. - SEPT	GM12878	55.00	-	-	0.62	0.61	0.77
		HUVEC	55.83	-	-	0.69	0.65	0.83
		HeLa-S3	55.67	-	-	0.65	0.64	0.82
		IMR90	57.50	-	-	0.71	0.65	0.82
		K562	53.83	-	-	0.66	0.59	0.78
		NHEK	55.67	-	-	0.70	0.63	0.81
		HMEC	56.17	-	-	0.67	0.63	0.82
2020	Liu et al. - EPIHC	GM12878	-	-	-	-	0.925+	0.79+
		HUVEC	-	-	-	-	0.900+	0.66+
		HeLa-S3	-	-	-	-	0.940+	0.83
		IMR90	-	-	-	-	0.910+	0.75+
		K562	-	-	-	-	0.930+	0.77+
		NHEK	-	-	-	-	0.950+	0.85+
2019	Talukder et al. - EPIP	GM12878	-	0.9979	0.8772	0.9337	0.9007	0.8502
		HUVEC	-	0.9915	0.1824	0.3081	0.9842	0.8904
		KBM7	-	0.9940	0.3629	0.5317	0.9856	0.9285
		IMR90	-	0.9985	0.4672	0.6365	0.9963	0.9894
		K562	-	0.9938	0.3887	0.5589	0.9947	0.9874
		NHEK	-	0.9926	0.2897	0.4485	0.9942	0.9394
		HMEC	-	0.9938	0.2733	0.4287	0.9872	0.9038
2019	Zhuang et al.	GM12878	-	-	-	-	0.91	0.64
		HeLa-S3	-	-	-	-	0.95	0.77
		HUVEC	-	-	-	-	0.92	0.79
		IMR90	-	-	-	-	0.90	0.56
		K562	-	-	-	-	0.87	0.50
		NHEK	-	-	-	-	0.96	0.81
2019	Li et al. - DeepTACT	tB	-	-	-	-	0.946	0.815
		Mon	-	-	-	-	0.931	0.801
		tCD4	-	-	-	-	0.941	0.815
		tCD8	-	-	-	-	0.937	0.800
		nCD4	-	-	-	-	0.942	0.823
		FoeT	-	-	-	-	0.939	0.837
2019	Gao et al. - EAGLE	GM12878 balanced	-	-	-	-	0.9338	0.9236
		GM12878 imbalanced	-	-	-	-	0.9327	0.7289
2018	Zeng et al. - EP2vec	GM12878	-	-	-	0.867	-	-
		HeLa-S3	-	-	-	0.920	-	-
		HUVEC	-	-	-	0.875	-	-
		IMR90	-	-	-	0.872	-	-
		K562	-	-	-	0.882	-	-
		NHEK	-	-	-	0.933	-	-
		FANTOM	-	-	-	0.841	-	-
2018	Singh et al. - SPEID	GM12878	-	-	-	0.85	0.99+	0.94+
		HeLa-S3	-	-	-	0.81	0.99+	0.89+
		HUVEC	-	-	-	0.75	0.99+	0.95+
		IMR90	-	-	-	0.78	0.99+	0.95+
		K562	-	-	-	0.85	0.99+	0.97+
		NHEK	-	-	-	0.94	0.99+	0.97+

tCD8), and concluded that DeepTACT (AuROC: 0.93 to 0.94, AuPR: 0.80 to 0.83) significantly outperforms SPEID (AuROC: 0.66 to 0.75, AuPR: 0.59 to 0.72).

Furthermore, some methods also report F1 score, Accuracy, Sensitivity, and PPV. Considering the reported F1 scores, EP2vec and SPEID shows better performance over SEPT and EPIP across all cell-lines using their global model. However, the difference in train and test dataset creation with SEPT makes it difficult to complete the analysis. The other metrics are only reported by single methods, thus preventing their use in performance comparison of the models. Furthermore, 3DPredictor only reports PCC,

TABLE 6: **Performance evaluation of Belokopytova et al. for EPI prediction**

Year, Author and Method	Datasets	Evaluation Metrics				
		PCC	SCC	MSE	MAE	MRE
2019	GM12878	0.971	0.640	0.001	0.001	0.329
- 3DPredictor	GSE95116	0.935	0.625	0.005	0.001	1.13

SCC, MSE, MAE and MRE which no other method reports as shown in Table 6, due to which the comparison of it's performance with other methods cannot be undertaken unless the experiment is repeated and the required metrics are generated.

For cell-line specific EPI classification models, 3 methods, namely EPIP, EAGLE and the method by Zhuang et al., implement and report the model's performance as illustrated in Table 4. Both EPIP and Zhuang et al.'s method report very high AuROC values (ranging from 0.98 to 0.99) across all cell lines. However, EPIP outperforms all the methods in terms of AuPR (from 0.97 to 0.99) across all cell lines. Even though comparatively lower than the other models under study here, EAGLE still reports high AuROC (from 0.93 to 0.95) and AuPR (from 0.91 to 0.95) for all cell lines predicted upon.

5.2. Performance comparison of Pseudouridine modification predictors

For Ψ-site identification techniques, the performance of the classifier models are predominantly evaluated on the dataset that was built by Chen et al. in 2016, for their work with Ψ-site identification called *iRNA-PseU* [26]. The data consists of RNA sequences from RNA Modification Base (RMBase) database for three different species, Homo sapiens (H_990), Saccharomyces cerevisiae (S_628), and Mus musculus (M_944) [27]. Additionally, Chen et al. also provided two independently generated datasets on which the performances were reported as well, Homo sapiens (H_200) and Saccharomyces cerevisiae (S_200) [26]. However, two methods report performance for datasets different from the ones mentioned above. Khan et al.'s MU-PseUDeep [12] employs data for hg19 (human), mm10 (mouse), and sacCer3 (yeast) from RMBase v2.0 [28]. PIANO, developed by Song et al. [15], uses human Ψ-sites detected from four different base-resolution Ψ profiling techniques, including Ψ-Seq, RBS-Seq, CeU-Seq, and Pseudo-Seq from Gene Expression Omnibus (GEO) [29], for cell lines HEK293, HEK293T and Hela. Due to this difference of datasets, the performance of MU-PseUDeep and PIANO cannot be directly compared with the other methods under study here.

The most frequently reported metrics for Ψ-site prediction are Accuracy, Sensitivity, Specificity and MCC. For the five datasets built by Chen et al., Merged-seq, RF-PseU, iPseU-Layer, iPseU-CNN, XG-PseU and iPseU-NCP report all the aforementioned 4 performance metrics as illustrated in Table 7. For H_990, iPseU-Layer reports the highest accuracy (79.70%), specificity (88.22%) and MCC (0.60), and Merged-seq reports highest sensitivity (85.61%). For S_628, iPseU-Layer reports highest accuracy (89.34%), specificity (93.76%) and MCC (0.79), and Merged-seq reports highest sensitivity (88.29%). For M_944, iPseU-Layer reports highest accuracy (80.08%), sensitivity (77.92%), specificity (81.82%) and MCC (0.60). For the independent dataset, S_200, Merged-seq reports highest accuracy (76.50%), sensitivity (80.00%) and MCC (0.53), and iPseU-Layer and PseU-NCP reports highest specificity (78.00%). For H_200, iPseU-NCP reports highest accuracy (75.00%), sensitivity (73.00%) and MCC (0.50), iPseU-Layer reports highest specificity (79.00%).

A distinguishable trend that can be deduced from the performance comparison is that the two methods, Merged-seq and iPseU-Layer, consistently outperform the other methods for the 5 datasets built by Chen et al. Furthermore, the performance reported by Khan et al. for their method, MU-PseUDeep, although not comparable with the other methods due to difference in dataset, still have qualitatively high performance. Similarly, PIANO, proposed by Song et al., also reports significant performance in terms of AuROC.

6. Conclusion

The advent of in-silico computational methods has provided the ability to precisely predict 3D chromatin interaction in DNA and RNA modification in the respective sequences, thereby replacing expensive, skill and labor intensive wet lab approaches. These computational methods based on DNA/RNA sequences, genomic and epigenomic data, 3D structural contexts and distance measures, and physico-chemical properties have advanced the understanding of chromatin interaction, it's architecture and role in transcriptional regulation, and of pseudouridine modification and it's functional implications. In this study, recent methods for the two classification tasks have been collated, and the comparison of their performance evaluation has been undertaken.

As reviewed, the computational methods employ various Machine Learning algorithms. The trend seems to focus more on the implementation of ensemble techniques like Random Forest, AdaBoost, amongst others. This could be attributed to the fact that Ensemble models often outperform singular constituent algorithms used in ML. However, these methods primarily depend on informative features which often require expertise and a prior feature engineering step. With the advent of Deep Learning, the feature engineering step can be automatised by the use of hybrid architectures containing CNNs and RNNs which can extract spatial features and temporal dependencies respectively from the DNA/RNA sequences. The analysis of the performance shows that these DL techniques can match up to, or even outperform the ensemble ML techniques, based on the metrics in focus.

For EPI prediction methods, although there are various types of features used, like the sequences themselves, genomic and epigenomic features, features extracted using Natural Language Processing, and 3D interactions, it is not

TABLE 7: **Performance comparison of recent Ψ-Modification prediction techniques**

Year	Author and Method	Datasets	Evaluation Metrics					
			ACC (%)	Sn. (%)	Sp. (%)	MCC	F1	AuROC
2021	Aziz et al. - Merged-seq	H_990	78.83	85.61	72.07	0.59	-	-
		S_628	85.85	88.29	83.37	0.72	-	-
		M_944	77.23	76.62	77.60	0.54	-	-
		S_200	76.50	80.00	73.00	0.53	-	-
		H_200	74.00	73.00	75.00	0.48	-	-
2020	Khan et al. - MU-PseUDeep	hg19	89.40	70.90	81.50	0.37	0.42	-
		hg19 - balanced	72.60	70.90	81.00	0.52	0.75	-
		mm10	85.40	80.00	73.00	0.38	0.44	-
		mm10 - balanced	76.00	80.00	73.00	0.54	0.77	-
		saCcer3	86.90	74.20	78.80	0.36	0.40	-
		saCcer3 - balanced	76.80	74.20	79.80	0.55	0.76	-
2020	Lv et al. - RF-PseU	H_990	64.30	66.10	62.60	0.29	-	0.700
		S_628	74.80	77.20	72.40	0.50	-	0.810
		M_944	74.80	73.10	76.50	0.50	-	0.796
		S_200	77.00	75.00	79.00	0.540	-	0.838
		H_200	75.00	78.00	72.00	0.501	-	0.800
2020	Mu et al. - iPseU-Layer	H_990	79.70	71.18	88.22	0.60	-	-
		S_628	89.34	84.68	93.76	0.79	-	-
		M_944	80.08	77.92	81.82	0.60	-	-
		S_200	72.50	68.00	78.00	0.45	-	-
		H_200	71.00	63.00	79.00	0.43	-	-
2020	Song et al. - PIANO	Ψ-Seq - full transcript	-	-	-	-	-	0.957
		Ψ-Seq - mRNA	-	-	-	-	-	0.859
		RBS-Seq - full transcript	-	-	-	-	-	0.978
		RBS-Seq - mRNA	-	-	-	-	-	0.770
		CeU-Seq- full transcript	-	-	-	-	-	0.914
		CeU-Seq - mRNA	-	-	-	-	-	0.864
		Pseudo-Seq- full transcript	-	-	-	-	-	0.972
		Pseudo-Seq - mRNA	-	-	-	-	-	0.857
2019	Tahir et al. - iPseU-CNN	H_990	66.68	65.00	68.78	0.34	-	-
		S_628	68.15	66.36	70.45	0.37	-	-
		M_944	71.81	74.29	69.11	0.44	-	-
		S_200	69.00	77.72	60.81	0.40	-	-
		H_200	73.50	68.76	77.82	0.47	-	-
2019	Liu et al. - XG-PseU	H_990	66.05	63.45	68.65	0.32	-	0.70
		S_627	71.10	65.92	76.30	0.43	-	0.77
		M_944	73.42	77.35	69.48	0.47	-	0.74
2019	Nguyen et al. - iPseU-NCP	H_990	62.92	58.79	65.05	0.24	-	-
		S_628	69.59	77.07	62.10	0.40	-	-
		M_944	71.82	67.37	76.27	0.44	-	-
		S_200	74.00	70.00	78.00	0.48	-	-
		H_200	75.00	73.00	77.00	0.50	-	-

yet clear if a certain type of feature, or a hybrid set of features outperforms others. As an example, it can be seen that EPIP employing 3D interaction and histone modification features with an ensemble ML model performs at par, in in certain metrics better than most other methods. Similarly, for Ψ-site prediction, iPseU-Layer, implementing a multi-layer ensemble model using sequence and derived chemical composition and properties, outperforms all other models in most metrics, closely matched by Merged-seq encoding with CNN based network built by Aziz et al. Additionally, for both the classification tasks, it is also observed that the use of Natural Language Processing techniques, like DNA2vec, EP2vec, etc. have also showcased performance boosts over previous approaches. Hence, further experiments need to be designed with comparable classifier architecture to identify the predictive power that each of the different types of features contain, which in turn would enable parallel paths of research based on optimum combination of these features.

During analysis of the performance metrics, it was realized that AuROC and AuPR are the most widely reported metric for EPI prediction. However, certain methods, like EP2vec, 3DPredictor report other measures such as F1 score, different kinds of correlation metrics and statistical error metrics. Similarly for Ψ-site prediction, Accuracy, Sensitivity, Specificity and MCC are widely reported, however, some methods like PIANO only report AuROC. Due to this, the methods cannot be directly compared with the other methods without recreating the experiment to generate the desired metrics. Furthermore, certain EPI prediction methods, like EAGLE, DeepTACT and 3DPredictor, do not employ the widely used TargetFinder dataset for evaluation of their method. Similarly, for Ψ-site prediction as well, PIANO and MU-PseUDeep deviates from using the widely popular datasets provided by Chen et al. This also makes it impossible to directly compare their performance with the others methods in the study. Hence, there needs to be a

consensus in the community to structure these components of research in such a way that identifying the progress becomes clear and concise.

Acknowledgments

The author would like to thank his mentor, Mr. Muhammad Nabeel Asim for his guidance throughout the literature survey, and DFKI Kaiserslautern for providing the opportunity to expand his knowledge in the domain by taking part in the Applied AI Seminar, SS2021.

References

[1] Z. Wang, L. Zhou, S. Jiang & W. Huang. (2021). *EPnet: A general network to predict enhancer-promoter interactions*. 11th International Conference on Information Science and Technology (ICIST), 2021, pp. 119-124, doi: 10.1109/ICIST52614.2021.9440647.

[2] Jing, Fang & Zhang, Shihua. (2020). *Prediction of enhancer–promoter interactions using the cross-cell type information and domain adversarial neural network*. BMC Bioinformatics. 21. 10.1186/s12859-020-03844-4.

[3] Liu, Shuai & Xu, Xinran & Yang, Zhihao & Zhao, Xiaohan & Zhang, Wen. (2020). *EPIHC: Improving Enhancer-Promoter Interaction Prediction by using Hybrid features and Communicative learning*.

[4] Belokopytova, Polina & Nuriddinov, Miroslav & Mozheiko, Evgeniy & Fishman, Daniil & Fishman, Veniamin. (2019). *Quantitative prediction of enhancer-promoter interactions. Genome Research*. 30. gr.249367.119. 10.1101/gr.249367.119.

[5] Talukder, Amlan & Saadat, Samaneh & Li, Xiaoman & Hu, Haiyan. (2019). *EPIP: A novel approach for condition-specific enhancer-promoter interaction prediction*. Bioinformatics (Oxford, England). 35. 10.1093/bioinformatics/btz641.

[6] Zhuang, Zhong & Shen, Xiaotong & Pan, Wei. (2019). *A simple convolutional neural network for prediction of enhancer-promoter interactions with DNA sequence data*. Bioinformatics (Oxford, England). 35. 2899-2906. 10.1093/bioinformatics/bty1050.

[7] Li, Wenran & Wong, Wing & Jiang, Rui. (2019). *DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning*. Nucleic acids research. 47. 10.1093/nar/gkz167.

[8] Gao, Tianshun & Qian, Jiang. (2019). *EAGLE: an algorithm that utilizes a small number of genomic features to predict tissue/cell type-specific enhancer-gene interaction*. 10.1101/781427.

[9] Zeng, Wanwen & Wu, Mengmeng & Jiang, Rui. (2018). *Prediction of enhancer-promoter interactions via natural language processing*. BMC Genomics. 19. 10.1186/s12864-018-4459-6.

[10] Singh, Shashank & Yang, Yang & Poczos, Barnabas & Ma, Jian. (2019). *Predicting enhancer-promoter interaction from genomic sequence with deep neural networks*. Quantitative Biology. 7. 10.1007/s40484-019-0154-0.

[11] Bin Aziz, Abu Zahid & Hasan, Md. Al & Shin, Jungpil. (2021). *Identification of RNA pseudouridine sites using deep learning approaches*. PLOS ONE. 16. e0247511. 10.1371/journal.pone.0247511.

[12] Khan, Saad & He, Fei & Wang, Duolin & Chen, Yongbing & Xu, Dong. (2020). *MU-PseUDeep: A deep learning method for prediction of pseudouridine sites*. Computational and Structural Biotechnology Journal. 18. 10.1016/j.csbj.2020.07.010.

[13] Lyu, Zhibin & Zhang, Jun & Ding, Hui & Zou, Quan. (2020). *RF-PseU: A Random Forest Predictor for RNA Pseudouridine Sites*. Frontiers in Bioengineering and Biotechnology. 8. 10.3389/fbioe.2020.00134.

[14] Mu, Yashuang & Zhang, Ruijun & Wang, Lidong & Liu, Xiaodong. (2020). *iPseU-Layer: Identifying RNA Pseudouridine Sites Using Layered Ensemble Model*. Interdisciplinary Sciences: Computational Life Sciences. 12. 10.1007/s12539-020-00362-y.

[15] Song, Bowen & Tang, Yujiao & Zhen, Wei & Liu, Gang & Su, Jionglong & Meng, Jia & Chen, Kunqi. (2020). *PIANO: A Web Server for Pseudouridine-Site (Ψ) Identification and Functional Annotation*. Frontiers in Genetics. 11. 10.3389/fgene.2020.00088.

[16] Tahir, Muhammad & Tayara, Hilal & Chong, Kil. (2019). *iPseU-CNN: Identifying RNA Pseudouridine sites Using Convolutional Neural Networks*. Molecular Therapy - Nucleic Acids. 16. 10.1016/j.omtn.2019.03.010.

[17] Liu, Kewei & Chen, Wg & Lin, Hao. (2020). *XG-PseU: an eXtreme Gradient Boosting based method for identifying pseudouridine sites*. Molecular Genetics and Genomics. 295. 10.1007/s00438-019-01600-9.

[18] Nguyen Vo, Thanh Hoang & Quang, Nguyen & Do, T. & Rahardja, Susanto & Nguyen, Binh. (2019). *iPseU-NCP: Identifying RNA pseudouridine sites using random forest and NCP-encoded features*. BMC Genomics. 20. 971. 10.1186/s12864-019-6357-y.

[19] Wittkopp, P. & Kalay, G. (2012). *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nature Reviews Genetics 13, 59–69 https://doi.org/10.1038/nrg3095

[20] Xu, Hang & Zhang, Shijie & Xianfu, Yi & Plewczynski, Dariusz & Mulin Jun, Li. (2020). *Exploring 3D chromatin contacts in gene regulation: The evolution of approaches for the identification of functional enhancer-promoter interaction.* Computational and Structural Biotechnology Journal. 18. 10.1016/j.csbj.2020.02.013.

[21] Davis, Jesse & Goadrich, Mark. (2006). *The Relationship Between Precision-Recall and ROC Curves*. Proceedings of the 23rd International Conference on Machine Learning, ACM. 06. 10.1145/1143844.1143874.

[22] Kiełbasa, Szymon & Blüthgen, Nils & Fähling, Michael & Mrowka, Ralf. (2010). *Targetfinder.org: A resource for systematic discovery of transcription factor target genes.* Nucleic acids research. 38. W233-8. 10.1093/nar/gkq374.

[23] Harrow, Jennifer & Frankish, Adam & González, José & Tapanari, Electra & Diekhans, Mark & Kokocinski, Felix & Aken, Bronwen & Barrell, Daniel & Zadissa, Amonida & Searle, Stephen & Barnes, If & Bignell, Alexandra & Boychenko, Veronika & Hunt, Toby & Kay, Mike & Mukherjee, Gaurab & Rajan, Jeena & Despacio-Reyes, Gloria & Saunders, Gary & Hubbard, Tim. (2012). *GENCODE: the reference human genome annotation for the ENCODE project.* Genome Res. 22. 1760-74. 10.1101/gr.135350.111.

[24] Kundaje, Anshul & Meuleman, Wouter & Ernst, Jason & Bilenky, Mikhail & Yen, Angela & Heravi-Moussavi, Alireza & Kheradpour, Pouya & Zhang, Zhizhuo & Wang, Jianrong & Ziller, Michael & Amin, Viren & Whitaker, John & Schultz, Matthew & Ward, Lucas & Sarkar, Abhishek & Quon, Gerald & Sandstrom, Richard & Eaton, Matthew & Wu, Yi-Chieh & Lin, Yiing. (2015). *Integrative analysis of 111 reference human epigenomes.* Nature. 518. 317-30. 10.1038/nature14248.

[25] Javierre, Biola & Burren, Oliver & Wilder, Steven P & Kreuzhuber, Roman & Hill, Steven & Sewitz, Sven & Cairns, Jonathan & Wingett, Steven & Várnai, Csilla & Thiecke, Michiel & Burden, Frances & Farrow, Samantha & Cutler, Antony J & Rehnström, Karola & Downes, Kate & Grassi, Luigi & Kostadima, Myrto & Freire-Pritchett, Paula & Wang, Fan & Fraser, Peter. (2016). *Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters.* Cell. 167. 1369-1384.e19. 10.1016/j.cell.2016.09.037.

[26] Chen, Wg & Tang, Hua & Ye, Jing & Lin, Hao & Chou, Kuo-Chen. (2016). *iRNA-PseU: identifying RNA pseudouridine sites.* Molecular Therapy. Nucleic Acids. 5. 10.1038/mtna.2016.37.

[27] Wenju, Sun & Li, Jun-Hao & Liu, Shun & Wu, Jie & Zhou, Hui & Qu, Liang-Hu & Yang, Jianhua. (2015). *RMBase: A resource for decoding the landscape of RNA modifications from high-throughput sequencing data.* Nucleic acids research. 44. 10.1093/nar/gkv1036.

[28] Xuan, Jia-Jia & Wenju, Sun & Lin, Peng-Hui & Zhou, Keren & Liu, Shun & Zheng, Ling-Ling & Qu, Lianghu & Yang, Jianhua. (2017). *RMBase v2.0: Deciphering the map of RNA modifications from epitranscriptome sequencing data.* Nucleic acids research. 46. 10.1093/nar/gkx934.

[29] Barrett, Tanya & Troup, Dennis & Wilhite, Stephen & Ledoux, Pierre & Evangelista, Carlos & Kim, Irene & Tomashevsky, Maxim & Marshall, Kimberly & Phillippy, Katherine & Sherman, Patti & Muertter, Rolf & Holko, Michelle & Ayanbule, Oluwabukunmi & Yefanov, Andrey & Soboleva, Alexandra. (2011). *NCBI GEO: Archive for functional genomics data sets - Update.* Nucleic acids research. 39. D1005-10. 10.1093/nar/gkq1184.

[30] Bernstein. (2021). *Regulation of transcription in mammal.* CC-BY-SA-4.0 Wikimedia Commons, Creative Commons Attribution-Share Alike 4.0 International license. https://bit.ly/2WyfJ0k

[31] Fowler, Samantha & Roush, Rebecca & Wise, James, & OpenStax College. (2017). Concepts of biology. ISBN-13 978-1-947172-03-6. CB-2013-005(03/16)-RS. https://openstax.org/details/books/concepts-biology

[32] Sati S, Cavalli G. Chromosome conformation capture technologies and their impact in understanding genome function. Chromosoma 2017;126:33–44

[33] Simonis M, Klous P, Splinter E, et al. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 2006;38:1348–54.

[34] Dostie J, Richmond TA, Arnaout RA, et al. Chromosome conformation capture carbon copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 2006;16:1299–309.

[35] Dekker J, Rippe K, Dekker M, et al. Capturing chromosome conformation. Science 2002;295:1306–11.

[36] Lieberman-Aiden E, van Berkum N, Williams L, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 2009;326:289–93.