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

Sourajyoti Datta
Supervisor: **Muhammad Nabeel Asim**
Seminar – Applied AI (DFKI)
Summer Semester 2021

German
Research Center
for Artificial
Intelligence

Concepts of biology

- **Cells**
  - Basic building block of all living organisms
  - Contains genetic material
- **Genetic information of an organism**
  - Defines the form and function of the organism
  - Defines all biological/biochemical activities
  - Preserved in nucleic acids
- **Nucleic acids**
  - One of the *macro-molecules essential for life*
  - Composed of Nucleotides
  - **Deoxyribonucleic acid (DNA)**
    - Two strands of polynucleotide chains in a *double helix structure*
  - **Ribonucleic acid (RNA)**
    - A single strand of polynucleotide chain *wrapped on itself*

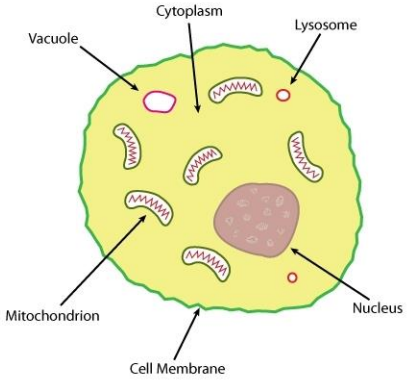


Fig: Outline of an animal cell showing the nucleus, cell membrane and some other organelles [1]

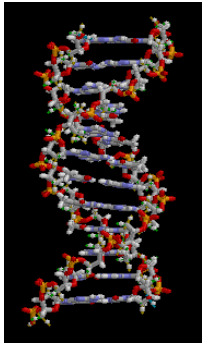


Fig: DNA

Fig: RNA

Central Dogma of Molecular Biology

- **Gene:**
  - Contained within chromosomes
  - *Segments of nucleotide sequence* in DNA or RNA
  - Encodes synthesis of a product

- **Gene expression:**
  - Produces functional gene products:
    - RNA
    - Proteins

- **Flow of genetic information:**
  1. DNA is copied to create RNA (**DNA Transcription**)
     - **Regulation of genetic expression** occurs
  2. RNA is used to produce proteins (**Translation**)

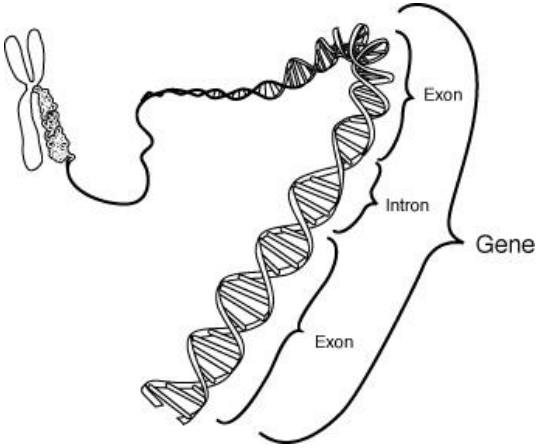


Fig: Gene, segment of DNA, part of chromosome [2]

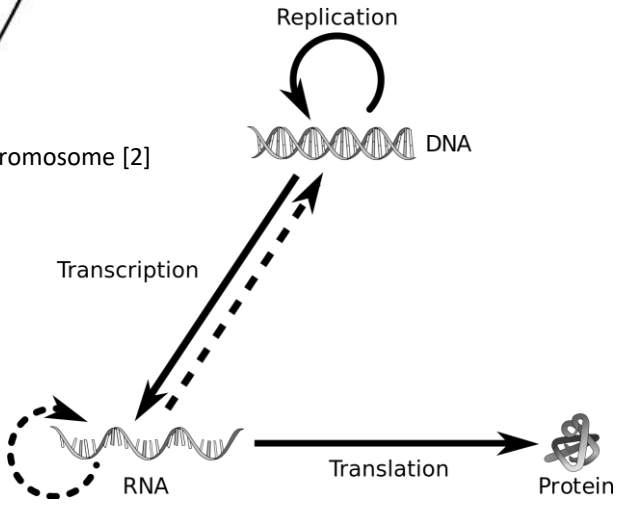


Fig: Flow of genetic information [3]

Regulation of DNA Transcription

- **Importance of Regulation for organisms:**
  - Provide attributes like **Versatility** and **Adaptability**
  - Through *Cellular differentiation* and *Morphogenesis*

- **Cis-regulatory Elements –** segments of DNA sequence**:**
  - **Enhancers**
  - **Promoters**
- Promoters define the *Transcription Start Sites (TSS)*

- **Enhancer-Promoter Interaction (EPI):**
  - Occurs through **DNA/Chromosome looping**
    - Favorable folding of the genome in the 3D space
  - *Active enhancers* interact with *specific promoters*
  - *Key role* in Transcription Regulation
  - Implications in *disease progression*

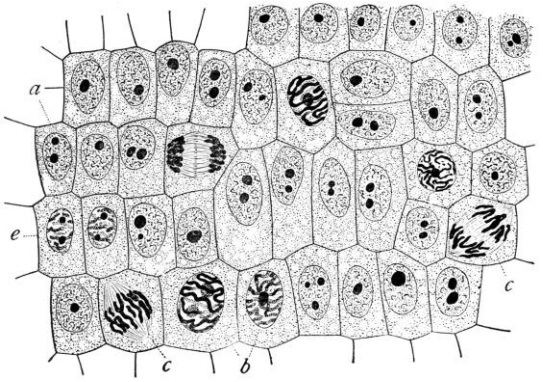


Fig: Variety of cells in root tissue of onion [4]

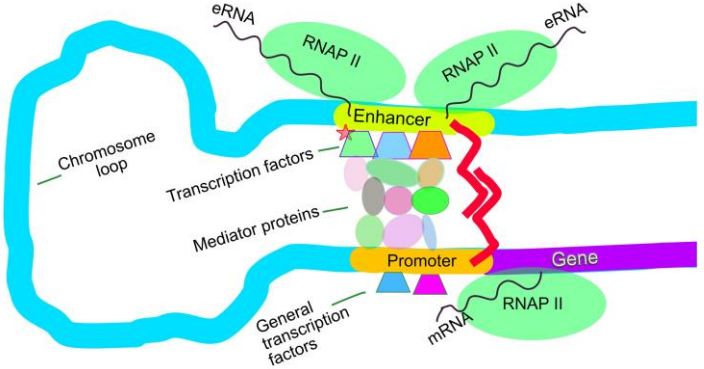


Fig: Regulation of transcription in mammals (EPI) [5]

RNA Modification/Editing

- *Post-transcriptional mechanism,* occurs in the RNA molecule
- Used by cells to perform specific changes in RNA sequences:
  - Nucleobase **Insertion, Deletion, Substitution**

- Importance of RNA Editing:
  - Affects different aspects of RNA
    - Activity, Localization, Structure, Stability
  - Role in **Regulation of Genetic Expression**
    - Supports genetic code rewiring
  - Linked to human diseases

- **Pseudouridine Modification (Ψ):**
  - *Most abundant* RNA modification in cellular RNA
  - Occurs at **Uridine sites**
  - **5-ribosyluracil** generated from **isomerization** of Uridine base
    - Catalyzed by *Pseudouridine Synthase* (PUS)

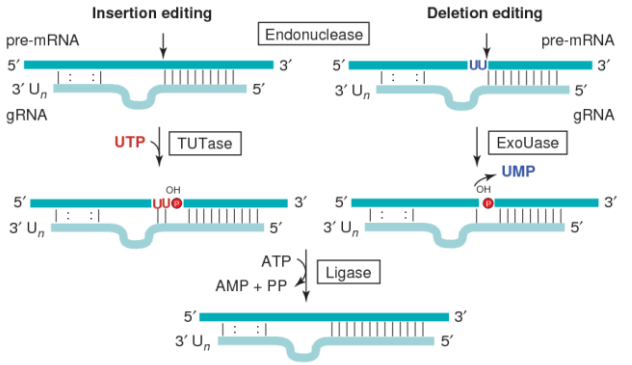


Fig: Sample RNA Editing: mRNA edited with gRNA [6]

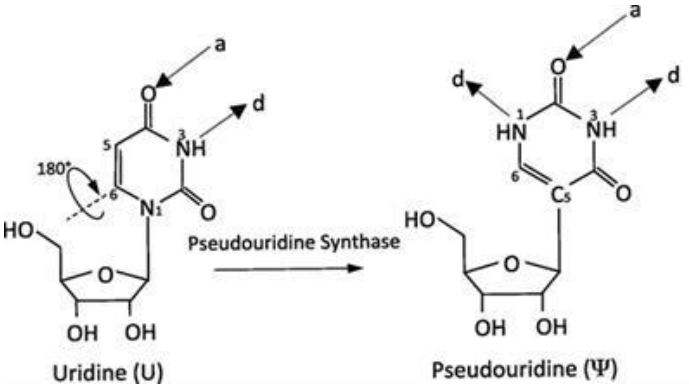


Fig: Pseudouridine Modification [7]

Motivation for computational methods

- **Experimental approaches** have been **successful** in EPI and Ψ-site identification
  - **Huge amounts of data generated** for knowledge discovery**:**
    - By high-throughput sequencing technologies
    - Genomic and Epigenomic data

- **Resulted in increasing use of Machine Learning and Deep Learning approaches:**
  - Due to overwhelming advantages of computational approaches
  - By high-throughput sequencing technologies

- **Wet lab experimental approaches:**
  - Technically challenging with Theoretical limitations
  - Labor intensive
  - Requires skilled laboratory experience
  - Time consuming
  - Expensive
  - High false-negative rates
  - ***Generates ground truth data***

- **Computational approaches:**
  - Easier development
  - Comparatively, much less labor intensive
  - Requires coding skills
  - Ease of reuse, hence, saves time
  - Economical
  - Accurate and Robust

EPI – Features

- **DNA Sequences as feature**
  - Long sequences, consisting of 4 nucleobases
    - Adenine [**A**] , Cytosine [**C**]
      Guanine [**G**] and Thymine [**T**]
  - Fixed length segments used as input

- **Natural Language Processed embeddings**
  - DNA2vec
  - EP2vec

- **3D Chromatin Interaction features**
  - ChIA-PET and Hi-C Contact Matrices
  - Distance Measures

- **Genomic features**
  - Gene Expression profiles
    - Enhancer Gene Correlation
    - Gene Score
  - DNA Methylation
  - ChIP-seq peaks
  - Window Signal (Enhancer and Gene)

- **Epigenomic features**
  - ChIP-seq profiles
  - RNA-seq profiles
  - Chromatin accessibility
    - DNase-seq
    - ChIP-seq

EPI – Classifiers

- **Machine Learning approaches**
  - Uses feature sets directly
    - Requires crafted features
      - Requires domain expertise
  - **Feature Selection** performed
    - To filter relevant and important features only
    - **Forward feature selection** widely used
  - Prevalence of **Ensemble** algorithms for classification:
    - Gradient Boosting Regression
    - Decision Trees
    - AdaBoost

- **Deep Learning approaches**
  - Uses feature sets directly
    - Do not require hand crafted features
  - **Feature Extraction** layer
    - Learns and extracts features directly from data
    - **Convolutional Neural Network** layer
      - Extracts *spatial features*
    - **Recurrent Neural Network** layer
      - Extracts *temporal dependencies*
  - **Fully Connected Neural Network** layer used for final classification

EPI – State-of-the-Art ML approach

- **Input for model**
  - **Enhancer-Promoter** pair features for 7 cell lines
    - 3D Chromatin Interaction features
    - Epigenomic features

- **Partitions**
  - Partitions with **overlapping feature set** are used
  - 11 in total

- **Incremental Learner (IL)**
  - Each IL trained on each partition
  - Contains 200 Weak leaners ($\mathbf{W_i}$), Decision Trees
  - **Result** is the **voting** of all the $\mathbf{W_i}$(s)

- **Model Prediction**
  - **Result** is the **voting** of all **11 ILs**

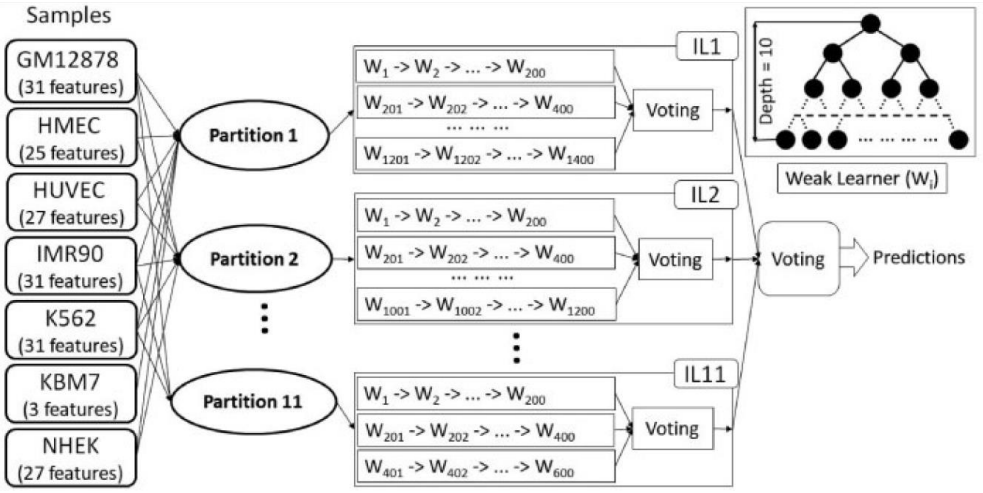


Fig: Architecture of **EPIP**, proposed by Talukder et. al. [8]

EPI – State-of-the-Art DL approach

- **Input to network**
  - **Enhancer** and **Promoter** sequence segment from DNA
  - **One-Hot Encoding** of segment
- **Feature Extractor**
  - Multi-path, multi-layer architecture
  - **CNN** processes Enhancer and Promoter sequences separately, parallelly
  - **RNN** processes combined features
- **Domain Discriminator**
  - Removes cell-line specific features
  - Retains global features for EPI
  - **Gradient Reversal Layer**
- **EPI Label Predictor**
  - **FCNN** layer

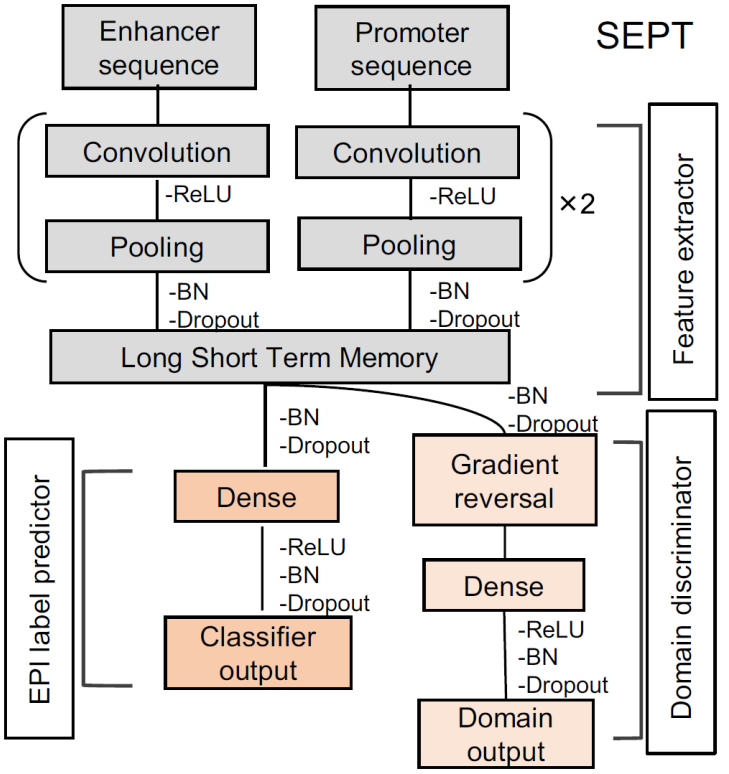
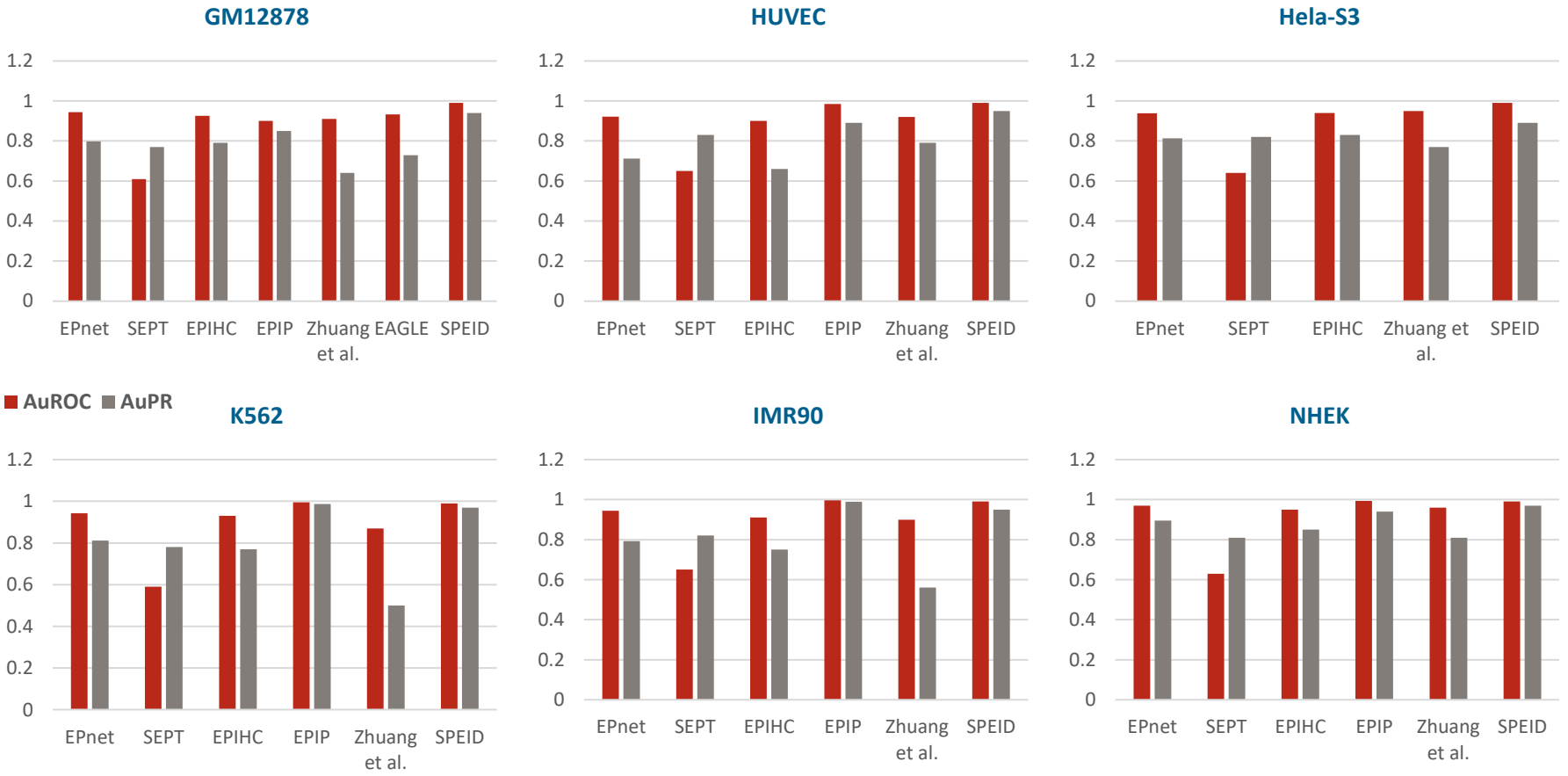
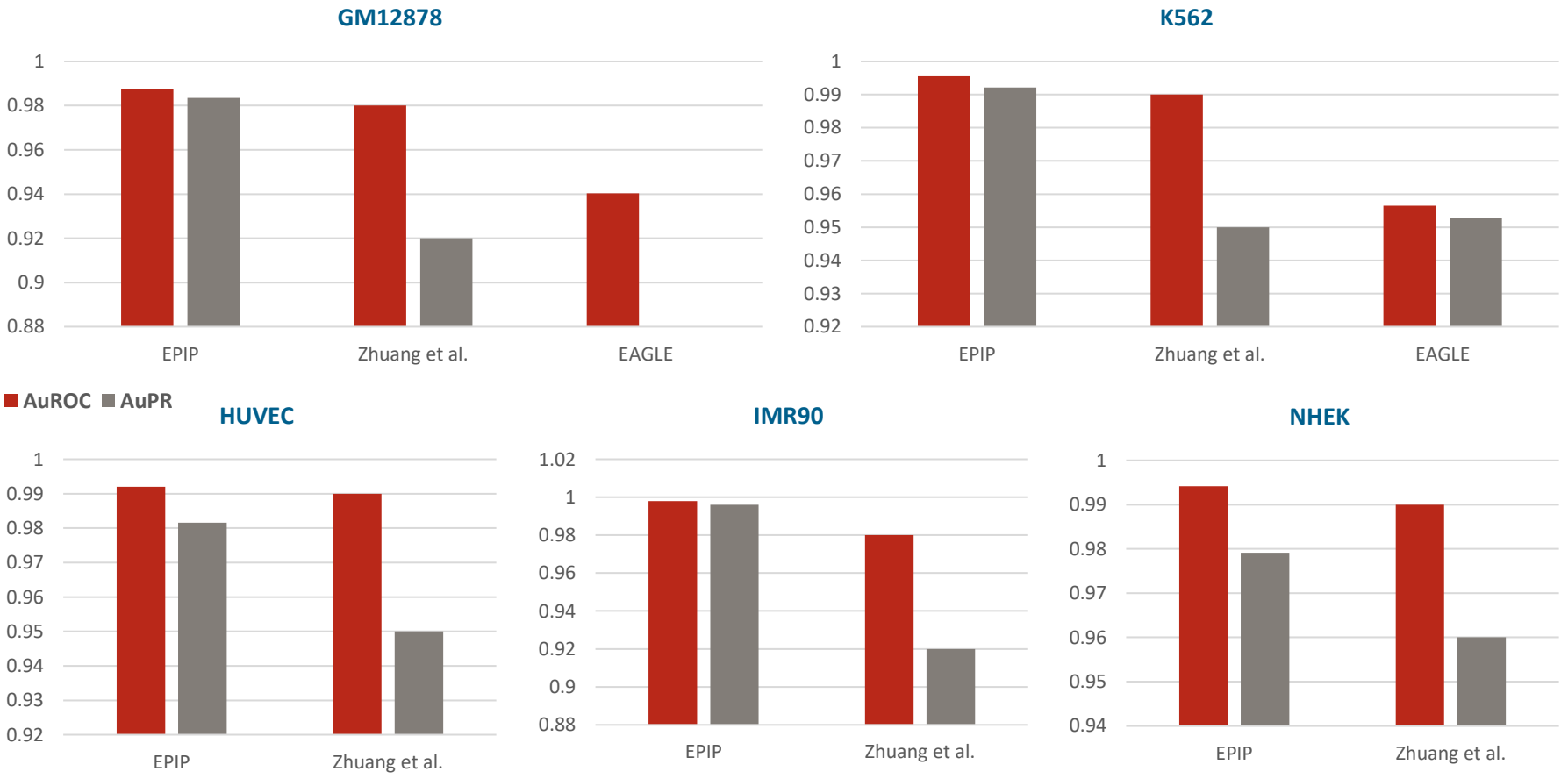


Fig: Architecture of **SEPT**, proposed by Jing et. al. [9]

EPI – Performance of Global models



EPI – Performance of Cell-line-specific models



Ψ-site – Features

- **RNA Sequences as feature**
  - Long sequences, consisting of 4 nucleobases
    - Adenine [**A**] , Cytosine [**C**], Guanine [**G**] and Uracil [**U**]
  - Fixed length segments used as input
- **Genomic features**
  - WHISTLE
  - PIANO
- **Structural context of mRNA**
- **Clustering information of RNA segments**
- **RNA Sequence encoding schemes as features:**
  - Nucleotide Binary Profiles
  - Nucleotide Density
  - Accumulated Nucleotide Frequency
  - Electron-Ion Interaction Pseudopotentials
  - Nucleic Acid Composition
  - Enhanced Nucleic Acid Composition
  - Di-Nucleotide Composition
  - Tri-Nucleotide Composition
  - Nucleotide Chemical Property
  - Position Specific Nucleotide Propensity
  - Xmer k-spaced Ymer Composition Frequency

Ψ-site – Classifiers

- **Machine Learning approaches**
  - Uses feature sets directly
    - Requires crafted features
      - Requires domain expertise
  - **Feature Selection** performed
    - To filter relevant and important features only
    - **Forward feature selection**
    - **Light Gradient Boosting Machine**
  - Prevalence of **Ensemble** algorithms for classification:
    - Random Forest
    - Extreme Gradient Boosting (XGboost)
    - Ensemble of different algorithms

- **Deep Learning approaches**
  - Uses feature sets directly
    - Do not require hand crafted features
  - **Feature Extraction** layer
    - Learns and extracts features directly from data
    - **Convolutional Neural Network** layer
      - Extracts spatial features from sequences
  - **Fully Connected Neural Network** layer used for final classification

Ψ-site – State-of-the-Art ML approach

- **Input for model**
  - **RNA Sequence encoding schemes** on segments
- **Feature Selection layer**
  - **Correlation based Feature Selection** (CFS) algorithm
  - Greedy hill-climbing augmented with backtracking
- **3 Feature Extraction and Fusion layers**
  - 6 ensemble ML algorithms
  - Voting based feature generation
  - Contains Skip connections
- **Prediction layer**
  - **Random Forest** based final classifier

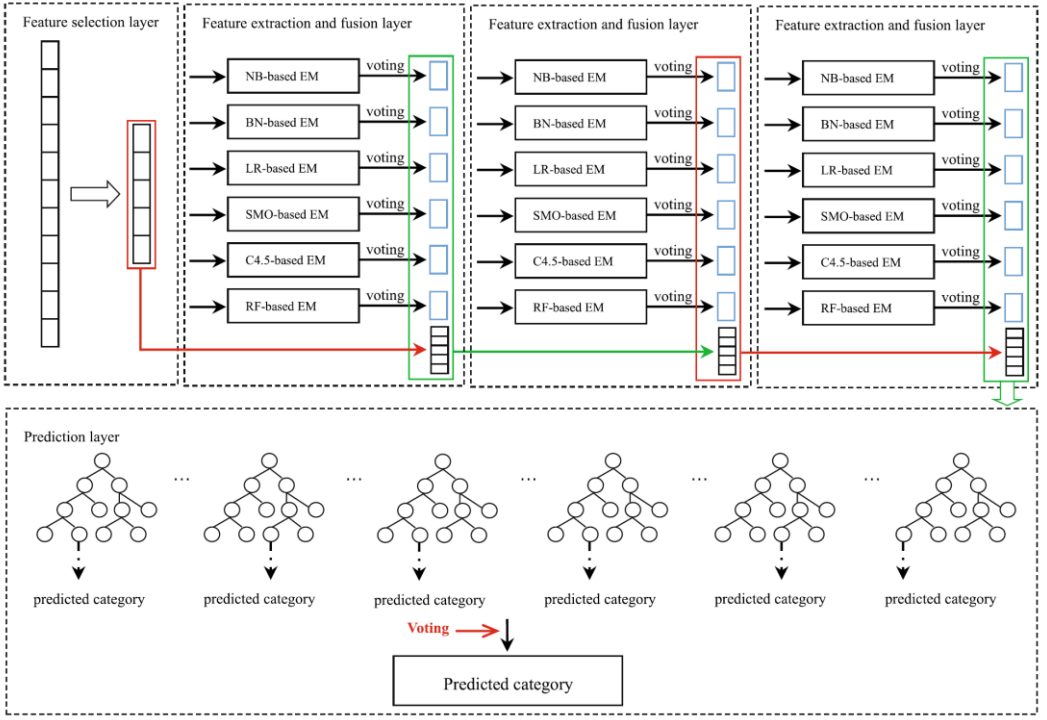


Fig: Architecture of **iPseU-Layer**, proposed by Aziz et. al. [10]

Ψ-site – State-of-the-Art DL approach

- **Input to network**
  - **RNA Sequence** encodings, concatenated:
    - **One-hot Encoding**
    - **Merged-seq Encoding**
- **Feature Extractor**
  - Multi-layer architecture
  - **CNN** processes the sequence vectors
  - Different sized kernels capture spatial information at different resolutions
- **Ψ-site Label Predictor**
  - 2-layer architecture
  - **FCNN** layer

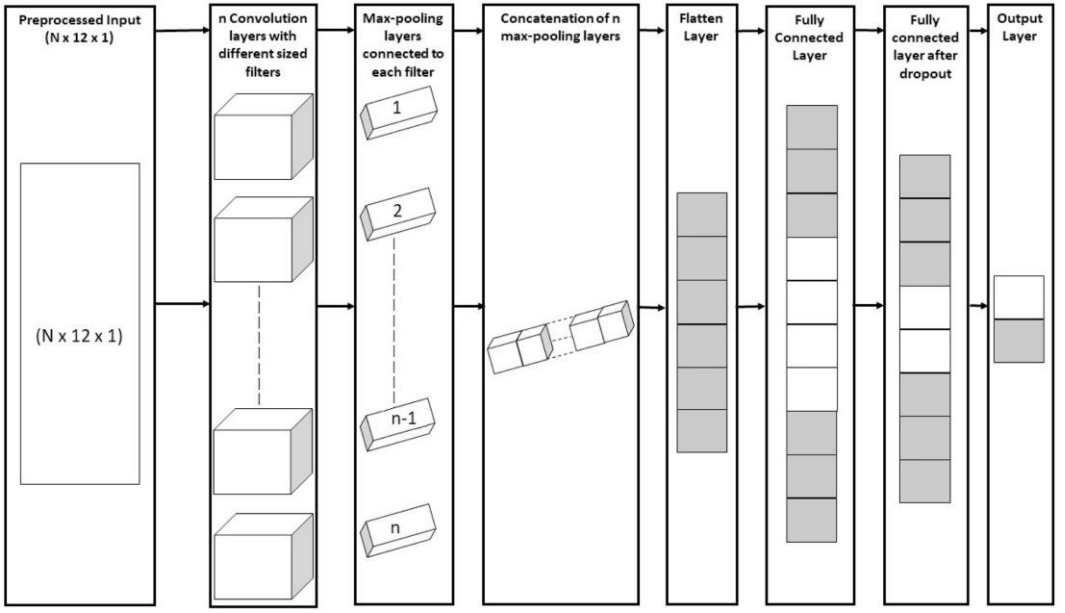
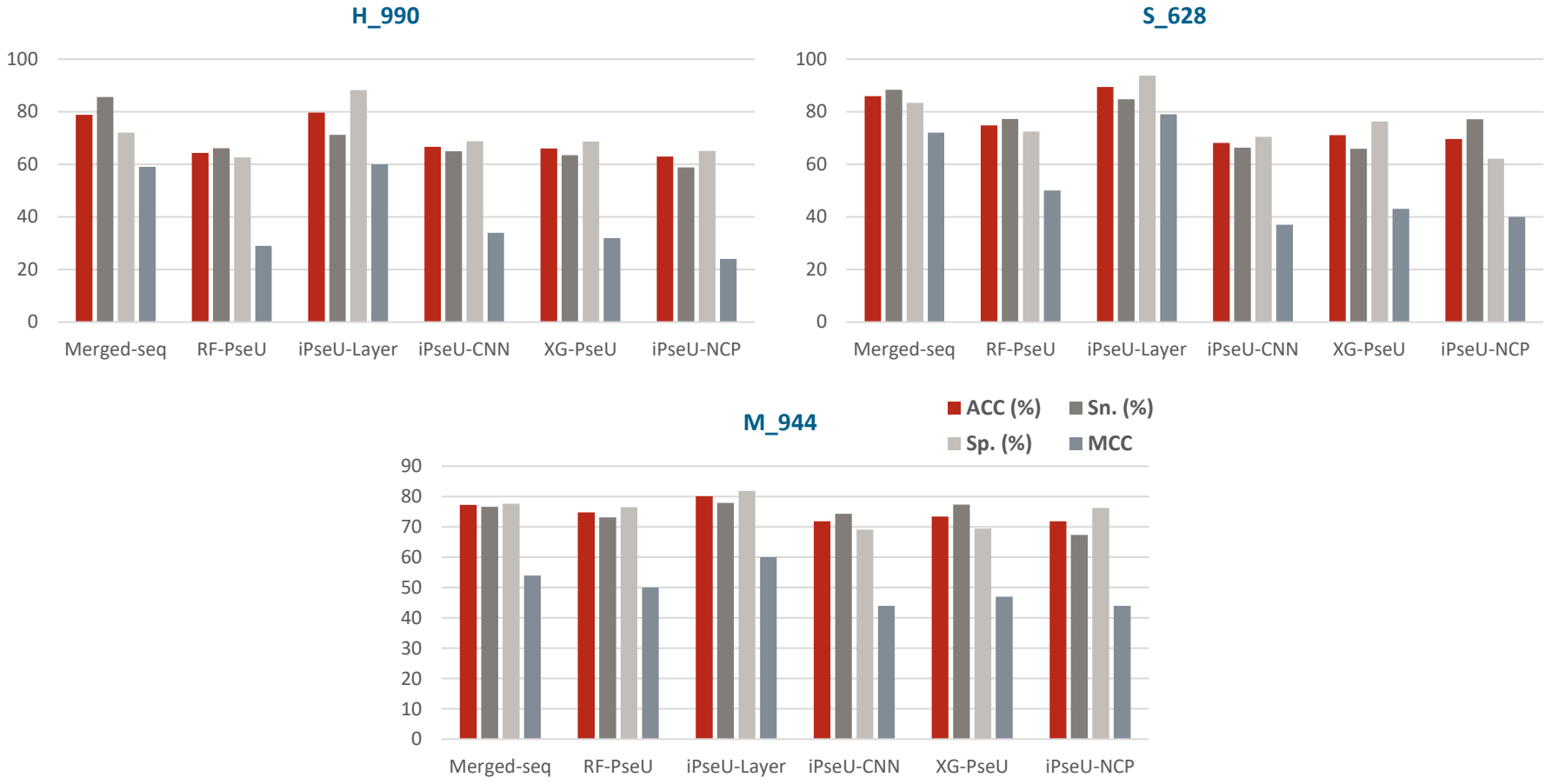
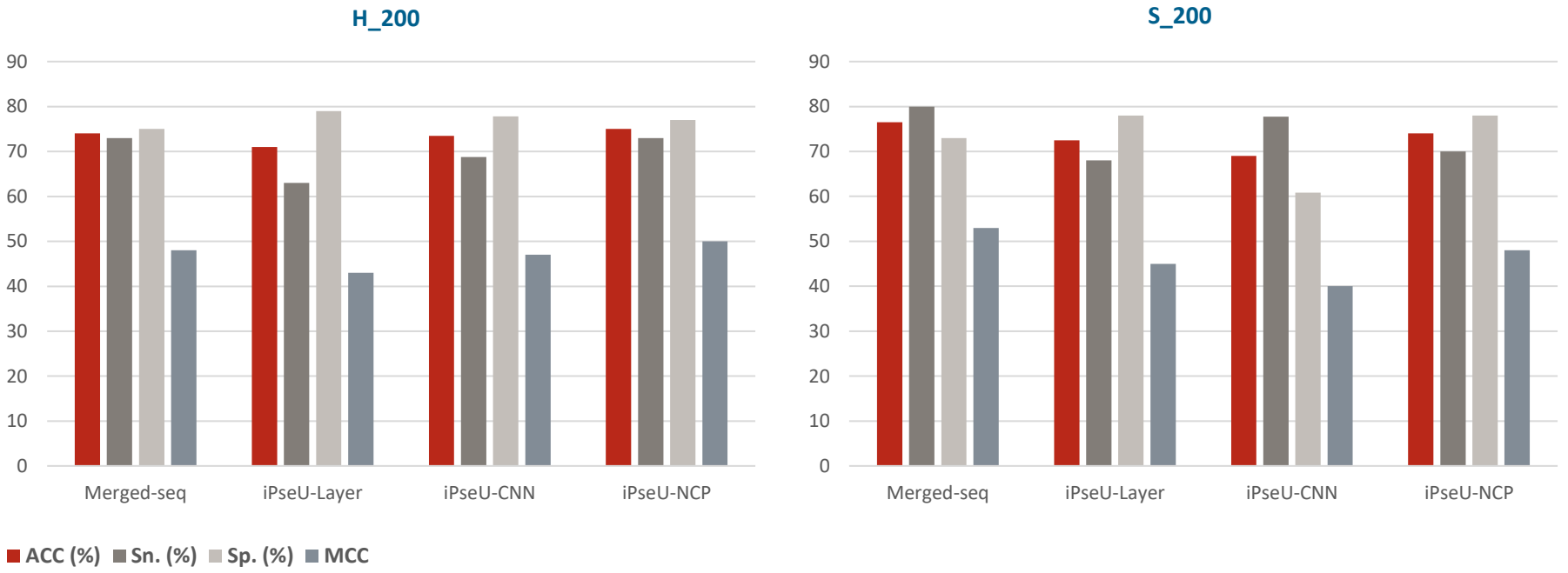


Fig: Architecture of **Merged-seq**, proposed by Aziz et. al. [11]

Ψ-site – Performance of approaches



Ψ-site – Performance of approaches



Conclusion

- **State of the art**
  - Huge repository of **in-silico methods** for prediction of EPI and Ψ-sites in DNA and RNA sequences respectively
  - **Trend: Machine Learning** approaches focus on **Ensemble** methods
  - Advent of **Deep Learning**, to learn important features from data
  - DL approaches out-perform ML approaches

- **Issues with surveyed researches**
  - Performance of approaches **not comparable** directly in some instances
    - **Non-uniform set of metrics** reported by researches
    - **Variety of datasets** (cell-lines) used by researches
  - Need consensus in community to structure these aspects to **identify progress** clearly

- **Further research opportunities**
  - Comparison between **predictive power** of different types of features
    - Analyze if certain type of feature, or a hybrid set of features outperforms others
  - Application of **RNN** and hybrid layers for Ψ-site prediction, like in EPI
  - Application of **Natural Language Processing** embeddings for Ψ-site prediction, like in EPI

Thank you for your time!

Questions?

References

1. **Science Learning Hub** – Pokapū Akoranga Pūtaiao, University of Waikato, www.sciencelearn.org.nz
2. **Cytogenetics**, Università degli Studi di Napoli Federico II
3. Emmanuel Barillot, Laurence Calzone, Philippe Hupé, Jean-Philippe Vert, Andrei Zinovyev, **Computational Systems Biology of Cancer** Chapman & Hall/CRC Mathematical & Computational Biology , 2012
4. Wilson, Edmund B. (1900) **The cell in Development and Inheritance (2nd ed.)**, Category: New York: The Macmillan Company
5. 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
6. Stuart, K. et al. **Complex management: RNA editing in trypanosomes.** Trends in biochemical sciences 30 2 (2005): 97-105 .
7. Zhao, Yang & Dunker, William & Yu, Yi-Tao & Karijolich, John. (2018). **The Role of Noncoding RNA Pseudouridylation in Nuclear Gene Expression Events.** Frontiers in Bioengineering and Biotechnology. 6. 10.3389/fbioe.2018.00008.
8. 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.
9. 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.
10. 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.
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.