

DISSERTATION

AN EFFICIENT AUTOMATED MACHINE LEARNING FRAMEWORK FOR GENOMICS AND PROTEOMICS SEQUENCE ANALYSIS

From DNA to RNA and Protein Sequence Analysis

Thesis approved by
Department of Computer Science
TECHNISCHE UNIVERSITÄT KAISERSLAUTERN
FOR THE AWARD OF THE DOCTORAL DEGREE
DOCTOR OF ENGINEERING (DR.-ING)

to

MUHAMMAD NABEEL ASIM

Date of Defense : December 19, 2022
Dean : Prof. Dr. Jens Schmitt
Reviewers : Prof. Dr. Prof. h.c. Andreas Dengel
: Prof. Dr. Sebastian Vollmer
: Prof. Dr. Sören Laue



DE-386

Executive Summary

Genomics and Proteomics sequence analyses are the scientific studies of understanding the language of Deoxyribonucleic Acid (DNA), Ribonucleic Acid (RNA) and protein biomolecules with an objective of controlling the production of proteins and understanding their core functionalities. It helps to detect chronic diseases in early stages, root causes of clinical changes, key genetic targets for pharmaceutical development and optimization of therapeutics for various age groups. Most Genomics and Proteomics sequence analysis work is performed using typical wet lab experimental approaches that make use of different genetic diagnostic technologies. However, these approaches are costly, time consuming, skill and labor intensive. Hence, these approaches slow down the process of developing an efficient and economical sequence analysis landscape essential to demystify a variety of cellular processes and functioning of biomolecules in living organisms.

To empower manual wet lab experiment driven research, many machine learning based approaches have been developed in recent years. However, these approaches cannot be used in practical environment due to their limited performance. Considering the sensitive and inherently demanding nature of Genomics and Proteomics sequence analysis which can have very far-reaching as well as serious repercussions on account of misdiagnosis, the main objective of this research is to develop an efficient automated computational framework for Genomics and Proteomics sequence analysis using the predictive and prescriptive analytical powers of Artificial Intelligence (AI) to significantly improve healthcare operations.

The proposed framework is comprised of 3 main components namely sequence encoding, feature engineering and discrete or continuous value predictor. The sequence encoding module is equipped with a variety of existing and newly developed sequence encoding algorithms that are capable of generating a rich statistical representation of DNA, RNA and protein raw sequences. The feature engineering module has diverse types of feature selection and dimensionality reduction approaches which can be used to generate the most effective feature space. Furthermore, the discrete and/or continuous value predictor module of the proposed framework contains a wide range of existing machine learning and newly developed deep learning regressors and classifiers. To evaluate the integrity and generalizability of the proposed framework, we have performed a large-scale experimentation over diverse types of Genomics and Proteomics sequence analysis tasks (i.e., DNA, RNA and proteins).

In Genomics analysis, Epigenetic modification detection is one of the key component. It helps clinical researchers and practitioners to distinguish normal cellular activities from malfunctioned ones, which can lead to diverse genetic disorders such as metabolic disorders, cancers, etc. To support this analysis, the proposed framework is used to solve the problem of DNA and Histone modification prediction where it has achieved state-of-the-art performance on 27 publicly available benchmark datasets of 17 different species with best accuracy of 97%. RNA sequence analysis is another vital component of Genomics sequence analysis where the identification of different coding and non-coding RNAs as well as their subcellular localization patterns help to demystify the functions of diverse RNAs, root causes of clinical changes, develop precision medicine and optimize therapeutics. To support this analysis, the proposed framework is utilized for non-coding RNA classification and multi-compartment RNA subcellular localization prediction. Where it achieved state-of-the-art performance on 10 publicly available benchmark datasets of Homo sapiens and Mus Musculus species with best accuracy of 98%.

Proteomics sequence analysis is essential to demystify the virus pathogenesis, host immunity responses, the way proteins affect or are affected by the cell processes, their structure and core functionalities. To support this analysis, the proposed framework is used for host protein-protein and virus-host protein-protein interaction prediction. It has achieved state-of-the-art performance on 2 publicly available protein protein interaction datasets of Homo Sapiens and Mus Musculus species with best accuracy of 96% and 7 viral host protein protein interaction datasets of multiple hosts and viruses with best accuracy of 94%. Considering the performance and practical significance of proposed framework, we believe proposed framework will help researchers in developing cutting-edge practical applications for diverse Genomic and Proteomic sequence analyses tasks (i.e., DNA, RNA and proteins).

DEDICATION AND ACKNOWLEDGEMENTS

First and foremost, I want to express my sincere gratitude to the Almighty Allah, for enabling me to complete the research at hand and making it possible for me to write this acknowledgment today. A special gratitude to Prof. Dr. Prof. h.c. Andreas Dengel for his tremendous guidance and support. You have consistently given me inspiration and motivation, I feel privileged to have the opportunity to collaborate with your research group. You are undoubtedly a remarkable inspiration for aspiring researchers, I could not have imagined having such a better mentor during my Ph.D. journey. I am greatly thankful to other members of my Ph.D. review committee, including Prof. Dr. Sebastian Vollmer, and Prof. Dr. Sören Laue, for reviewing and evaluating the quality of my work and providing insightful feedback for further improvement of my future research.

I would also acknowledge Dr. Sheraz Ahmed who has always been firm support for me throughout this entire study endeavor. I'm thankful that you supported my research and assisted me in growing as a research scientist. Your suggestions for my research study and career have been invaluable sources of inspiration.

A special thanks to all of my colleagues at the KM group, DFKI Kaiserslautern. I would like to thank Ali Ibrahim. His continuous motivation and support have always been my asset. I would also like to thank my friends who always gave me strength and motivation to work hard.

Last but not the least, I really admire the support of my family: my parents, my brother, my sisters, my nephews and my nieces. Words cannot express how grateful I am to all of you for all of the sacrifices that you have made on my behalf. Your prayers and support sustained me thus far. Without this support, it was never possible for me to think of this day. A huge bouquet of love, thanks and gratitude for my family.

PUBLICATIONS AS PART OF THIS THESIS

Parts of the research and material (including figures, tables and algorithms) in this thesis have already been published or submitted in:

Patent

1. Muhammad Nabeel Asim, Christoph Zehe, Olivier Cloarec, Johan Trygg, Sheraz Ahmed. "METHOD, COMPUTER PROGRAM PRODUCT AND SYSTEM FOR OPTIMIZING PROTEIN EXPRESSION", Application under review in European patent; MB&P Ref. S15547EU - hb / yma

Peer Reviewed Journal Articles

1. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Ahtisham Fazeel, Andreas Dengel, and Sheraz Ahmed. "DNA-MP: a generalized DNA modifications predictor for multiple species based on powerful sequence encoding method." *Briefings in Bioinformatics* (2022). DOI: [10.1093/bib/bbac546](https://doi.org/10.1093/bib/bbac546), Impact factor: 13.994
2. **Muhammad Nabeel Asim**, Ahtisham Fazeel, Muhammad Ali Ibrahim, Andreas Dengel, and Sheraz Ahmed. "MP-VHPPI: Meta predictor for viral host protein-protein interaction prediction in multiple hosts and viruses." *Frontiers in Medicine* 9 (2022) DOI: [10.3389/fmed.2022.1025887](https://doi.org/10.3389/fmed.2022.1025887), Impact factor: 5.058
3. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Christoph Zehe, Johan Trygg, Andreas Dengel and Sheraz Ahmed. "A Robust and Precise ConvNet for small non-coding RNA classification (RPC-snRC)" , *IEEE Access* 9 (2020): pp. 19379-19390 DOI: [10.1109/ACCESS.2020.3037642](https://doi.org/10.1109/ACCESS.2020.3037642), Impact Factor: 3.476
4. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Christoph Zehe, Johan Trygg, Andreas Dengel and Sheraz Ahmed "MirLocPredictor: A ConvNet-Based Multi-Label MicroRNA Subcellular Localization Predictor by Incorporating k-Mer Positional Information", *Genes* 11(12), (2020) p.147 DOI: [10.3390/genes11121475](https://doi.org/10.3390/genes11121475), Impact Factor: 4.141
5. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "Advances in Computational Methodologies for Classifica-

tion and Sub-Cellular Locality Prediction of Non-Coding RNAs", *International Journal of Molecular Sciences* 22(16), (2021) p.8719 DOI: [10.3390/ijms22168719](https://doi.org/10.3390/ijms22168719), Impact Factor: 6.208

6. Marco Stricker, **Muhammad Nabeel Asim**, Andreas Dengel and Sheraz Ahmed. "CircNet: an encoder–decoder-based convolution neural network (CNN) for circular RNA identification", *Neural Computing and Applications* pp.1-12 (2021) DOI: [10.1007/s00521-020-05673-1](https://doi.org/10.1007/s00521-020-05673-1), Impact Factor: 5.102
7. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Imran Razzak, Andreas Dengel and Sheraz Ahmed. "Histone-Net: A Multi-Paradigm Computational Framework for Histone Occupancy and Modification Prediction", *Complex & Intelligent Systems* pp.1-21 (2022) DOI: [10.1007/s40747-022-00802-w](https://doi.org/10.1007/s40747-022-00802-w), Impact Factor: 6.7
8. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Christoph Zehe, Johan Trygg, Andreas Dengel and Sheraz Ahmed. "BoT-Net: A Lightweight Bag of Tricks based Neural Network for Efficient lncRNA–miRNA Interaction Prediction", *Interdisciplinary Sciences: Computational Life Sciences* pp.1-22 (2022) DOI: [10.1007/s12539-022-00535-x](https://doi.org/10.1007/s12539-022-00535-x), Impact Factor: 3.492
9. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "LGCA-VHPPI: A Local-Global Residue Context Aware Viral-Host Protein-Protein Interaction Predictor", *PLOS ONE* 17(7), p.e0270275 (2022) DOI: [10.1371/journal.pone.0270275](https://doi.org/10.1371/journal.pone.0270275), Impact Factor: 3.75
10. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "Circ-LocNet: A Computational Framework for Circular RNA Sub-Cellular Localization Prediction", *International Journal of Molecular Sciences* 23(15), p.8221 (2022) DOI: [10.3390/ijms2315822](https://doi.org/10.3390/ijms2315822), Impact Factor: 6.208
11. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "EL-RMLocNet: An Explainable LSTM Network for RNA-Associated Multi-Compartment Localization Prediction", *Computational and Structural Biotechnology Journal* (2022) DOI: [10.1016/j.csbj.2022.07.031](https://doi.org/10.1016/j.csbj.2022.07.031), Impact Factor: 6.155
12. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "CONR-NET: A Collection of Neural Refinements for Protein Protein Interaction Prediction", *iScience* (2022) , Impact Factor: 6.107

Conference Papers

13. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "K-mer Neural Embedding Performance Analysis Using Amino Acid Codons", *International Joint Conference on Neural Networks, (IJCNN)* (pp. 1-8). (2020) *IEEE*
DOI: [10.1109/IJCNN48605.2020.9206892](https://doi.org/10.1109/IJCNN48605.2020.9206892)
14. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "Enhancer-DSNet: A Supervisedly Prepared Enriched Sequence Representation for the Identification of Enhancers and Their Strength.", *27th International Conference on Neural Information Processing, (ICONIP-2020)* (pp. 38-48). Springer, Cham
DOI: [10.1007/978-3-030-63836-8_4](https://doi.org/10.1007/978-3-030-63836-8_4)
15. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Christoph Zehe., Cloarec, O., Sjogren, R., Johan Trygg, Andreas Dengel and Sheraz Ahmed. "L2S-MirLoc: A Lightweight Two Stage MiRNA Sub-Cellular Localization Prediction Framework", *International Joint Conference on Neural Networks, (IJCNN)* (pp. 1-8) (2021). *IEEE*
DOI: [10.1109/IJCNN52387.2021.9534015](https://doi.org/10.1109/IJCNN52387.2021.9534015)
16. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "ChrSLoc-Net: Machine Learning-Based Prediction of Channel-rhodopsins Proteins within Plasma Membrane", *IEEE EMBS International Conference on Biomedical and Health Informatics, (BHI)* (pp. 1-4) (2021). *IEEE*
DOI: [10.1109/BHI50953.2021.9508615](https://doi.org/10.1109/BHI50953.2021.9508615)

Other Publications

1. Faiza Mehmood, Muhammad Usman Ghani, Hina Ghafoor, Rehab Shahzadi, Muhammad Nabeel Asim, and Waqar Mahmood. "EGD-SNet: A computational search engine for predicting an end-to-end machine learning pipeline for Energy Generation Demand Forecasting." *Applied Energy* 324 (2022): 119754. [10.1016/j.apenergy.2022.119754](https://doi.org/10.1016/j.apenergy.2022.119754) Impact Factor: 11.446
2. Faiza Mehmood, Muhammad Usman Ghani, Muhammad Nabeel Asim, Rehab Shahzadi, Aamir Mehmood, and Waqar Mahmood. "MPF-Net: A computational multi-regional solar power forecasting framework." *Renewable and Sustainable Energy Reviews* 151 (2021): 111559. [10.1016/j.rser.2021.111559](https://doi.org/10.1016/j.rser.2021.111559) Impact Factor: 16.799
3. Muhammad Ali Ibrahim, Muhammad Usman Ghani Khan, Faiza Mehmood, Muhammad Nabeel Asim, and Waqar Mahmood. "GHS-NET a generic hybridized shallow neural net-

-
- work for multi-label biomedical text classification." *Journal of biomedical informatics* 116 (2021): 103699. DOI: [10.1016/j.jbi.2021.103699](https://doi.org/10.1016/j.jbi.2021.103699) Impact Factor: 8.00
4. Muhammad Nabeel Asim, Muhammad Ali Ibrahim, Muhammad Usman Ghani Khan, Waqar Mahmood, Andreas Dengel and Sheraz Ahmed. "Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification", *Neural Computing and Applications* 33(11), (2021) pp.5437-5469 DOI: [10.1007/s00521-020-05321-8](https://doi.org/10.1007/s00521-020-05321-8) Impact Factor: 5.102
 5. Faiza Mehmood, Muhammad Usman Ghani, Muhammad Ali Ibrahim, Rehab Shahzadi, Waqar Mahmood, and Muhammad Nabeel Asim. "A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis." *IEEE Access* 8 (2020): 192740-192759. [10.1109/ACCESS.2020.3030885](https://doi.org/10.1109/ACCESS.2020.3030885) Impact Factor: 3.476
 6. Muhammad Nabeel Asim, Muhammad Imran Malik, Muhammad Usman Ghani Khan, Andreas Dengel and Sheraz Ahmed. "A robust hybrid approach for textual document classification", *International conference on document analysis and recognition (ICDAR) (2019) (pp. 1390-1396). IEEE* DOI: [10.1109/ICDAR.2019.00224](https://doi.org/10.1109/ICDAR.2019.00224)
 7. Muhammad Wasim, Muhammad Nabeel Asim, Muhammad Usman Ghani Khan, and Waqar Mahmood. "Multi-label biomedical question classification for lexical answer type prediction." *Journal of biomedical informatics* 93 (2019): 103143. DOI: [10.1109/ICDAR.2019.00224](https://doi.org/10.1109/ICDAR.2019.00224) Impact Factor: 8.00
 8. Muhammad Nabeel Asim, Muhammad Imran Malik, Muhammad Usman Ghani Khan, Andreas Dengel and Sheraz Ahmed. "Two stream deep network for document image classification", *International conference on document analysis and recognition (ICDAR) (2019) (pp. 1410-1416). IEEE* DOI: [10.1109/ICDAR.2019.00227](https://doi.org/10.1109/ICDAR.2019.00227)

TABLE OF CONTENTS

Executive Summary	i
	Page
List of Tables	xv
List of Figures	xix
1 Introduction	1
1.1 Motivation	3
1.2 Problem Statement	5
1.3 Research Questions and Goals	6
1.4 Contributions	7
1.4.1 Genomics & Proteomics Sequence Analysis	8
1.4.2 Genomics Sequence Analysis	9
1.4.3 Proteomics Sequence Analysis	12
1.5 Dissertation Overview	14
2 A GENERIC FRAMEWORK For Genomics (DNA, RNA) and Proteomics (Protein) Sequence Analysis	17
2.1 Functional Scope and Description of Proposed Generic Framework	19
2.1.1 Feature Representation Module	21
2.1.2 Feature Engineering Module	22
2.1.3 Predictor Construction Module	29
2.1.4 Performance Evaluation Module	32
2.2 A Look Back & into Future: Functional Scope of the Existing and Proposed Generic Framework	35
I Genomics Sequence Analysis	37
3 DNA Modification Prediction	39
3.1 Related Work	40
3.2 Materials and Methods	42

TABLE OF CONTENTS

3.2.1 Proposed DNA Sequence Encoder	42
3.2.2 Benchmark Datasets	47
3.3 Evaluation Criteria	49
3.4 Results and Discussions	49
3.4.1 Performance Analysis of Proposed DNA Sequence Encoder at Different k-mers	49
3.4.2 Performance Impact of Proposed DNA Sequence Encoder on Different Classifiers	51
3.4.3 Intrinsic Performance Comparison of the Proposed DNA Sequence Encoder with Different Existing Encoders	56
3.4.4 Performance Comparison of Proposed DNA-MP Predictor with State-of-the-art Predictors	57
3.5 Conclusion	60
4 Histone Occupancy/Modifications Prediction and Enhancer Identification/Strength Prediction	61
4.1 Related work	63
4.1.1 Histone Occupancy and Modification Prediction	63
4.1.2 Enhancer Identification and Strength Prediction	64
4.2 Materials and Methods	66
4.2.1 Proposed Methodology	66
4.2.2 Benchmark Datasets	68
4.3 Evaluation Criteria	71
4.4 Results and Discussions	71
4.4.1 Evaluation of Histone-Net in Intra-Domain Setting using Binary Classification Paradigm	71
4.4.2 Performance Comparison of Histone-Net Predictor with Adapted and State-of-the-art Histone Occupancy and Modification Predictors	72
4.4.3 Evaluation of Histone-Net in Cross-Domain Setting	73
4.4.4 Intrinsic Evaluation of Histone-Net Predictor	74
4.4.5 Evaluation of Histone-Net in Multi-label Classification Paradigm	75
4.4.6 Performance Comparison of Proposed Enhancer-DSNet Approach with Existing Enhancer Identification and Strength Prediction Approaches	81
4.5 Conclusion	82
5 Small Non-Coding RNA Classification	85
5.1 Related Work	87
5.2 Materials and Methods	88
5.2.1 Proposed RPC-snRC Methodology	88
5.2.2 Benchmark Dataset	91

5.3 Evaluation Criteria	92
5.4 Results and Discussions	92
5.4.1 Class Level Performance Comparison of Proposed RPC-snRC and State-of-the-art nRC Methodologies	93
5.5 Conclusion	95
6 Circular RNA Identification	97
6.1 Related Work	98
6.2 Materials and Methods	99
6.2.1 Proposed Methodology	99
6.2.2 Benchmark Dataset	103
6.3 Evaluation Criteria	103
6.4 Results and Discussions	104
6.5 Conclusion	107
7 RNA Subcellular Location Prediction	109
7.1 Related Work	110
7.2 Materials and Methods	113
7.2.1 Proposed EL-RMLocNet Approach	113
7.2.2 A K-hop Neighbourhood Relation based Statistical Representation Scheme for RNA Sequences (GeneticSeq2Vec)	113
7.2.3 Explainable Deep Learning based RNA Associated Multi-Compartment Localization Predictor	117
7.2.4 Benchmark RNA-Associated Subcellular Localization Prediction Datasets	122
7.3 Evaluation Criteria	123
7.4 Results and Discussions	124
7.4.1 Performance Assessment of EL-RMLocNet for Multi-Compartment RNA Localization Prediction	125
7.4.2 Comparison of EL-RMLocNet with Existing Multi-Compartment RNA Localization Predictors	129
7.4.3 Visualization of Most Informative Nucleotide k-mers Patterns	131
7.5 Conclusion	132
II Proteomics Sequence Analysis	135
8 Protein-Protein Interaction Prediction	137
8.1 Related Work	137
8.2 Materials and Methods	141
8.2.1 Methodology of Proposed ADH-PPI Predictor	141

TABLE OF CONTENTS

8.2.2 Benchmark Datasets	149
8.3 Evaluation Criteria	150
8.4 Results and Discussions	152
8.4.1 A Comprehensive Performance Analysis of Traditional Sequence Preprocessing Strategies	152
8.4.2 Performance Analysis of Proposed Subsequence based Preprocessing Approaches	153
8.4.3 Performance Comparison of Proposed Subsequence Approaches with Traditional Sequence Fixed-Length Generation Approaches	155
8.4.4 Performance Impact of CNN Layer	158
8.4.5 Performance Assessment of ADH-PPI Robustness for Different Order Protein Sequence Pairs	159
8.4.6 Performance Comparison of Proposed ADH-PPI Predictor with Existing PPI Predictors using two Benchmark Core Datasets	161
8.4.7 Performance Comparison of Proposed ADH-PPI Predictor with Existing PPI Predictors using four Independent Test Sets	163
8.4.8 A Case Study: Objective Evaluation of Proposed Strategies for Fixed-Length Generation of Sequences	164
8.5 Explainability of Proposed ADH-PPI Predictor	164
8.6 Conclusion	166
9 Protein Virus Interaction Prediction	169
9.1 Related Work	170
9.2 Materials and Methods	173
9.2.1 Meta Predictor	173
9.2.2 Protein Sequence Encoding	174
9.2.3 Iterative Representation Learning	179
9.2.4 Benchmark Datasets	179
9.3 Evaluation Criteria	181
9.4 Results and Discussions	181
9.4.1 Performance Analyses of Proposed Meta Predictor using Different Representations at Property Level and Encoder Level	181
9.4.2 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on Barman’s Dataset	183
9.4.3 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on Denovo’s Dataset	184
9.4.4 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on SARS-CoV-2 Dataset	185

9.4.5 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on Unseen Viruses Test Sets	186
9.4.6 Discussion	188
9.5 Conclusion	192
10 Conclusion and Future Work	195
10.1 Conclusions	195
10.2 Limitations	197
10.3 Future Work	198
Bibliography	1
Appendix	1
Curriculum Vitae	1

LIST OF TABLES

TABLE	Page
2.1 Generic encoding methods for DNA, RNA and protein sequences	23
2.2 Encoding methods for DNA and RNA sequences	24
2.3 Encoding methods for protein sequences	25
2.4 Feature selection and dimensionality reduction methods in the proposed GFGPA framework	26
2.5 Classifiers and regressors available in proposed GFGPA framework	31
2.6 Confusion matrix for binary classification	33
2.7 A comprehensive functional scope analysis of proposed GFGPA and existing frameworks	36
3.1 A hypothetical dataset containing 15 sequences related to modification (c_1) and non-modification (c_2) classes. In the sequence samples, occurrence frequencies of a particular k-mer at 10 different positions	46
3.2 K-mers densities in, modification ($kmer_{ij}^{posden}$) and non-modification ($kmer_{ij}^{negden}$) classes, based on the k-mer densities, scores and ranks assigned by the existing PSTNPss [85] and proposed POCD-ND encoders to a particular k-mer present at 10 different positions	46
3.3 Extrinsic performance analysis of proposed and 32 existing encoders using 17 modification datasets. For each encoder accuracy values of 2 top performing classifier are given.	54
3.4 5-fold cross-validation based performance comparison of the proposed DNA modification predictor with existing generic and type specific predictors i.e., iDNA-MS [282], DCNN-4mc [342] and Bert-6ma [392], across 17 different benchmark datasets in terms of 5 evaluation measures.	57
3.5 Performance comparison of the proposed DNA modification predictor with existing generic and type specific predictors i.e., iDNA-MS [282], DCNN-4mc [342], iDNA-MT [435] and Bert-6ma [392], based on independent test sets in terms of 5 evaluation measures for 17 different DNA modification datasets.	59

4.1 Statistical summary of 10 benchmark datasets including 2 datasets for histone occupancy detection, 3 datasets for acetylation and 5 datasets for methylation level prediction.	69
4.2 Statistics of enhancer identification and strength prediction datasets	70
4.3 Performance statistics of proposed Histone-Net [22] predictor over 10 benchmark datasets using 5 different k-mers	72
4.4 Accuracy comparison of proposed Histone-Net [22] approach with state-of-the-art HCNN [443] and adapted DeepHistone [444] approach. Accuracy values of DeepHistone are obtained by processing raw histone sequences of various histone markers using convolutional neural network model presented by the authors [444] and accuracy values of HCNN are taken from Table 4.3 of Yin et al. [443] study.	72
4.5 Performance of proposed Histone-Net [22] predictor in cross-domain setting using different degree higher order residue based sequence representation.	73
4.6 Performance statistics of Histone-Net predictor [22] using different size k-mers over imbalanced and balanced version of multilabel dataset	76
4.7 Performance statistics of proposed Histone-Net [22] and adapted DeepHistone predictors using optimal size k-mer, over imbalanced and balanced versions of multilabel dataset in terms of 11 distinct evaluation metrics	76
4.8 5-Fold cross-validation based performance comparison of proposed Enhancer-DSNet [20] and latest existing predictor [379] for enhancer/non-enhancer and strong/weak enhancer prediction.	82
4.9 Performance comparison of proposed Enhancer-DSNet [20] with existing Enhancer/Non-Enhancer and Strong/Weak Enhancer predictors over independent test sets	82
5.1 Architecture summary of Res18-nRC and Res50-nRC	90
5.2 Characteristics of non-coding RNA classification dataset, where Max-seq length and Min-seq length illustrate maximum and minimum length of nucleotides in each class.	91
5.3 Performance statistics of the proposed RPC-snRC, adapted (Res18-nRC, Res50-nRC) and state-of-the-art (nRC [296] and RNAGCN [345]) methodologies on the benchmark small non-coding RNA dataset.	92
6.1 Scaling of two different sequences ATAG and ATATGUAT to length of 6 by either addition or removal with three different methods namely Pre, Middle and Post.	100
6.2 Statistics of benchmark dataset, where minimal and maximal sequence length represents the length of shortest and longest sequences, respectively. On the other hand average and standard deviation of sequence length illustrate the mean and standard deviation of sequences in the corresponding classes.	103
6.3 Performance statistics of CircNet [372] based on different adjacent nucleotides, scaling methods and sequence lengths.	105

7.1 A summary of existing computational subcellular localization predictors for miRNA, lncRNA, mRNA and circular RNA molecules	111
7.2 Optimal parameter values of proposed EL-RMLocNet approach for 8 benchmark datasets belonging to 4 different RNA classes and 2 species	123
7.3 Comparative analysis of 6 different fixed-length sequence generation approaches based on proposed EL-RMLocNet [23] approach over 8 benchmark datasets of 2 different species in terms of average precision	126
7.4 Performance comparison of proposed EL-RMLocNet approach with state-of-the-art approach for multi-compartment localization prediction of miRNA, mRNA, snoRNA and lncRNA using 8 benchmark datasets of Homo sapiens (Human) and Mus Musculus (Mouse) species	130
8.1 Optimal values of different hyperparameters of proposed ADH-PPI methodology for 2 core datasets and 4 independent test sets for the task of PPI prediction.	151
8.2 Performance analysis of proposed model using pipeline of LSTM, CNN and Attention layers and only LSTM and Attention layers over H.pylori species dataset to quantify the impact of CNN layer	159
8.3 Performance comparison of proposed ADH-PPI predictor with 12 existing PPI predictors on benchmark S. cerevisiae dataset, where results of existing PPI predictors are taken from Yu et al. [450] paper.	161
8.4 Performance comparison of proposed ADH-PPI predictor with 10 existing predictors on benchmark H. pylori dataset, where results of existing PPI predictors are taken from Yu et al. [450] paper.	162
9.1 Performance comparison of different statistical representations across 1 st stage RF classifier and with iterative feature learning based 2 nd stage SVM classifier.	182
9.2 Performance comparison of proposed MP-VHPPI with existing viral-host PPI predictors over a benchmark Barman dataset in terms of 7 different evaluation measures. Performance figures of Barman et al. SVM [35], Barman et al. RF [35], Alguwzizani et al. SVM [9] and Yang et al. RF [433] are taken from Yang et al. [433] work.	184
9.3 Performance comparison of proposed MP-VHPPI with existing viral-host PPI predictors over benchmark DeNovo dataset [121] in terms of 7 different evaluation measures. Performance figures of DeNovo SVM [121], Alguwzizani et al. SVM [9] and Yang et al. RF on DeNovo dataset [121] are taken from Yang et al. work [433].	185
9.4 Performance comparison of the proposed predictor with existing Yang et al. predictor [434] over the SARS-CoV-2 dataset.	186

9.5 Performance comparison of the proposed MP-VHPPI with existing virus-Host PPI predictors over 4 datasets developed by Zhou et al., [473], to assess the applicability on the unseen viruses. The performance values of the existing approaches i.e., Zhou et al., [473] and Tsukiyama et al. (LSTM-PHV) [393] are taken from their corresponding studies [393, 473]

187

LIST OF FIGURES

FIGURE	Page
2.1 The complete workflow of generic GFGPA framework proposed for efficient Genomics and Proteomics sequence analysis.	20
3.1 Working paradigm of proposed POCD-ND encoding method	42
3.2 Contour plots for PSTNPss encoder where contour lines parallel to diagonal reveals PSTNPss encoder assigns same scores to k-mers that have same positive to negative class density differences	45
3.3 Graphical representation of a particular k-mer at 10 different positions	46
3.4 Distribution of sequences in 17 benchmark datasets related to 4mc, 5hmc and 6ma modifications	48
3.5 Random Forest classifier based extrinsic performance analysis of statistical representations generated at different k-mers using proposed encoder for different modification datasets.	50
3.6 5-fold cross-validation based performance comparison of 10 classifiers using 17 modification datasets related to (a) 4mc modification datasets, (b) 5hmc modification datasets and (c) 6ma modification datasets .	52
3.7 Intrinsic performance analysis of proposed POCD-ND and 32 existing encoders using 4mc modification prediction dataset <i>C.equisetifolia</i> .	55
4.1 Histone octamer and nucleosome formation	62
4.2 Discriminative and overlapping k-mers in positive and negative classes	66
4.3 Workflow of Histone-Net [22] approach	67
4.4 Development workflow of imbalanced and balanced multi-label classification datasets for histone occupancy and modification prediction	70
4.5 Internal representation of Histone-Net predictor [22] for three distinct datasets at 11-mers	75
4.6 Confusion metrics of Histone-Net [22] predictor for unbalanced version of multi-label classification dataset, where each confusion matrix illustrates correct and wrong predictions of a particular class.	77

4.7 Confusion metrics of Histone-Net predictor [22] for balanced version of multi-label classification dataset, where each confusion matrix illustrates correct and wrong predictions of a particular class.	78
4.8 Performance analysis of proposed Histone-Net predictor [22] in terms of sequence label distributions using imbalance and balanced versions of multilabel classification datasets	80
5.1 A comprehensive taxonomy of RNA families	86
5.2 Proposed RPC-snRC methodology for small non-coding RNA classification. In figure, (128,16,18) indicates there are 128 kernels, each of width 16 and length 18 in a convolutional layer and (1,4) indicates kernel width and length are set to 1 and 4, respectively in a pooling layer. Remaining layers of network also follow same dimensionality pattern.	89
5.3 Accuracy confusion matrix of the proposed RPC-snRC [28], adapted Res18-nRC and state-of-the-art nRC [296] classification methodologies.	94
5.4 Class level performance comparison of proposed RPC-snRC [28] approach and state-of-the-art nRC [296] approach using small ncRNA classification dataset.	95
6.1 Circular RNA sequence extension by adding adjacent nucleotides	100
6.2 Graphical representation of the employed autoencoder.	101
6.3 Graphical representation of the employed classifier.	102
6.4 CircNet [372] performance in terms of AUROC for different experimental settings by taking subsequences from different positions along with fusion of genome adjacent nucleotide information	106
6.5 Performance comparison of proposed CircNet predictor with existing circular RNA classification approaches.	106
7.1 Results	115
7.2 A k-hop neighbourhood relation based statistical representation scheme for RNA sequences	116
7.3 Workflow of an explainable deep learning model for RNA associated multi-compartment subcellular localization prediction	118
7.4 Information flow in standard LSTM cell	119
7.5 Architecture of the Attention model	120
7.6 A comparison of variations in sequence length across 8 benchmark RNA associated multi-compartment subcellular localization datasets	123
7.7 Statistical distribution of benchmark RNA associated multi-compartment localization prediction datasets for Homo sapiens (A-D) and Mus Musculus species (E-H)	124
7.8 Multi-compartment localization prediction performance produced by EL-RMLocNet on 4 benchmark Homo spaien datasets of mRNA, miRNA, snoRNA and lncRNA corresponding to unique sequence-compartment distribution	127

7.9 Multi-compartment localization prediction performance produced by EL-RMLocNet [23] on 4 benchmark Mus Musculus datasets of mRNA, miRNA, snoRNA and lncRNA corresponding to unique sequence-compartment distribution	128
7.10 Most and least informative nucleotide k-mers patterns for 4 different RNAs belonging to Homo sapiens and Musculus species identified by attention layer of proposed EL-RMLocNet [23] approach	132
8.1 Workflow of unsupervised transfer learning applied using 6 datasets of distinct species to learn distributed representation of higher order sequence amino acids	142
8.2 A variety of experimental settings to generate fixed-length sequences based on traditional copy padding or sequence truncation and proposed bag of most informative amino acids distribution tricks.	144
8.3 Workflow of proposed attention based deep hybrid methodology ADH-PPI for protein-protein interaction prediction	146
8.4 Statistics of 2 core protein-protein interaction prediction datasets	150
8.5 Performance comparison of proposed ADH-PPI approach across 2 different datasets including S.Cerevisiae and H.Pyloir using 6 traditional sequence fixed-length generation approaches	152
8.6 Proposed classifier performance analysis under the hood of 4 different subsequences based strategies used to generate fixed-length sequences. Here 'PA' represents protein A and 'PB' refers to protein B, whereas S indicates the starting amino acids of protein sequence and E represents the ending amino acids of protein sequence.	154
8.7 Impact of 5 different settings on the performance of proposed ADH-PPI approach across 2 different datasets including S.Cerevisiae and H.Pyloir for the task of PPI prediction in terms of area under receiver operating characteristics. Setting 1 is based on traditional copy padding, sequence truncation and hybrid approaches. Settings 2, 3, 4 and 5 are based on subsequences criteria where X number of amino acids from starting and ending regions of protein A and protein B are taken. The value of X varies from 10 to 70 amino acids with the difference of 10 amino acids. Setting 2 takes X number of amino acids from starting region of protein A and ending region of protein B. Setting-3 takes X number of amino acids from ending region of protein A and protein B. Setting 4 takes X number of amino acids from starting region of protein A and protein B. Setting 5 takes X number of amino acids from starting and ending region of protein A and starting and ending region of protein B.	157
8.8 Performance assessment of most optimal informative subsequence generation setting using 2 differently ordered protein sequence pairs over S.Cerevisiae and H.Pyloir core datasets. Here P_A represents the protein A and P_B refers to protein B, whereas start and end represent the starting and ending region of respective protein	160

8.9 Performance assessment of most optimal informative subsequence generation setting using 2 differently ordered protein sequence pairs over C.Elegans, H.sapiens, M.musculus and E.coli independent test sets after training the model on core S.cerevisiae dataset.	160
8.10 Accuracy comparison of ADH-PPI and recent PPI predictors on 4 independent test sets	163
8.11 Most and least informative amino acid and k-mers patterns identified by Attention layer of proposed ADH-PPI predictor in two test protein sequences belonging to benchmark S.cerevisiae and H.pylori datasets	165
9.1 The overall working paradigm of the proposed VH-PPIs predictor. Dataset Construction To begin with, different datasets are collected from existing studies based on VH-PPIs from several databases such as, HPID, intact and VirusMentha. Feature Representation Obtained protein sequences are encoded on the basis of two physicochemical properties based protein sequence encoders i.e., QS order and APAAC. Feature Analyses Appropriate physicochemical properties are selected for the APAAC and QS order on the basis of feature analyses. Model Construction The VH-PPIs predictor is a SVM model formed on the basis of probabilistic vectors obtained from the RF and ET classifiers. Finally, a web server is established for fast and easy on-go analyses of VH-PPIs.	175
9.2 The process of computing amino acid combinations based on the lag values.	176
9.3 Distribution of sequences in interactive and non-interactive classes.	179
9.4 Distribution of amino acids in 7 different datasets. For each dataset the distribution of amino acids is shown across interactive and non-interactive protein samples	189
9.5 Clusters formation with representations of protein sequences based on APAAC and Qsorder without dimensionality reduction (a), with dimensionality reduction (b) and 6D representations from ET and RF classifier (c).	191

ACRONYMS

DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
ncRNA	Non Coding Ribonucleic Acid
lncRNA	long Non Coding Ribonucleic Acid
SNcRNA	Small Non Coding Ribonucleic Acid
mRNA	messenger Ribonucleic Acid
miRNA	micro Ribonucleic Acid
PPI	Protein Protein Interaction
VHPPI	viral Host Protein Protein Interaction
NLP	Natural Language Processing
DNN	Deep Neural Network
CNN	Convolution Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-term Memory
SVM	Support Vector Machine
RF	Random Forest
ET	Extra Tree
SGD	Stochastic Gradient Descent
bP	base Pair
AE	Autoencoder
ACC	Accuracy
PR	Precision
SP	Specificity
SN	Sensitivity
MCC	Mathew's Correlation Coefficient
FPR	False Positive Rate
TPR	True Positive Rate
AUC	Area Under The Curve
RMSE	Root Mean Square Error

INTRODUCTION

Humans (and other living organisms) usually rely on different organs (brain, heart, lungs, etc.) and biological systems (nervous system, immune system, reproductive system, etc.) in which organs communicate with each other to carry out vital functions such as pumping blood and keeping blood oxygenized. Organs comprise different tissues (e.g., skin, bones and nerve tissues) and tissues are made up by the grouping of similar cells, hence cell is the basic structural and functional unit of living organisms. Cells can grow, move, adapt to different environmental changes and can replicate themselves. Despite these similarities, cells differ in size, shape and capabilities to perform the functions of life. Cells contain Deoxyribonucleic Acid (DNA) that carries instructions about various activities such as synthesis of proteins which are essential to maintain the health and working of organs. From birth to death, DNA contains all the information organisms needed to grow, survive, reproduce and perform diverse types of functions. A 100 trillion meters long DNA is mainly comprised of repetitions of 4 different nitrogenous bases called Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). In order to compress long DNA within a tiny 100 micrometer (μm) cell, negatively charged DNA is wrapped around histone proteins octamer (2 copies of H2A, H2B, H3 and H4) known as Histone proteins octamer. The resulting DNA-protein complex is known as chromatin, fundamental structural and functional unit of which is known as nucleosome. In chromatin, thousands of nucleosomes organise DNA in 23 pairs of thread-like structures called chromosomes. On average, every chromosome comprises more than 100 million nucleotides and each nucleosome contains 146 nucleotides sequence wrapped around histone octamers.

Based on cell's elementary structures, living organisms are broadly classified into 2 categories namely eukaryotes and prokaryotes. Eukaryotes are multi-cellular organisms in which linearly arranged DNA lies inside the nucleus of cell. A few examples of such organisms are animals, plants, protists, fungi and humans. Prokaryotes are unicellular organisms in which circularly

arranged DNA lies in cytoplasm compartment of cell. A few examples of prokaryotes are bacteria and archaea. In eukaryotes such as humans, 50 trillion cells are broadly classified into 200 classes (e.g., nerve cells, muscle cells, skin cells, etc.) where cells of each class are responsible to perform distinct biological functions such as nerve cells transfer messages from brain to muscles for muscular activity like speaking. All 50 trillion cells contain the same DNA which is mainly comprised of 3 distinct regions: genes/coding DNA, regulatory elements and non-coding DNA. Within DNA, different combinations of 4 nucleotides ranging from a few hundred to more than 2 billion represent 1,40,000 functional units called genes which produce proteins essential for organism's growth, survival and health. Genes are a core part of genetic material which control functional modules of organ systems. Regulatory and non-coding DNA supports the process of gene expression regulation that controls unique types and numbers of proteins across different cells.

According to principal dogma of molecular biology, DNA replicates to provide genetic code for newborn cells and it is transcribed to RNA that further translates to proteins. In the process of transcription, coding DNA/genes are transformed into messenger RNA (mRNA) and non-coding DNA generates non-coding RNA. During the process of translation, mRNA and some particular types of non-coding RNAs make proteins, which are essential parts of life. Without the consumption of proteins, we cannot perform any activity including speaking, hearing, walking and even in the process of breathing, we are constantly burning a lot of proteins. Furthermore, even regulatory network that follows the instructions of genetic code to control the production of proteins cannot work without the support of proteins. For example, enzymes that metabolize nutrients to generate energy that is necessary for organs to perform different functions and also required for DNA polymerases which create the copies of DNA during the cell division, are all proteins. Regulatory network controls the expression and repression of genes to produce right type and amount of proteins within different cells. Within the regulatory network, irregularities at the transcription and translation levels lead to malfunctioning of molecular processes that eventually cause chronic diseases. To better understand and control the production of mRNAs in different cells, researchers are trying to find different factors such as diet, lifestyle, drug routine etc., non-coding DNA segments and regulatory elements roles that largely influence the process of transcription. Similarly, they are investigating non-coding RNA roles in translation to decode and control the synthesis of proteins across unique cell types. Genomics and Proteomics sequence analysis is the scientific study for understanding the language of DNA, RNA and protein biomolecules with an objective to control the expression and repression of genes at transcriptional as well as post-transcription levels. Through this analysis, biomedical researchers and practitioners are not only successfully forecasting the likelihood of different diseases in patients, but they have also developed genetic therapies to cure genetic diseases by suppressing or replacing the diseases specific genes. Furthermore, this analysis also paves way to develop novel drugs, optimize existing therapeutics and understand their impact on living organisms. More accurate and comprehensive

Genomics and Proteomics sequence analysis can potentially open new horizons for personalized healthcare, facilitating optimal treatment options to ensure healthier and longer lifespans.

Pre-dominantly, Genomics and Proteomics sequence analysis is being performed using typical wet lab experiments. However, these approaches require controlled environments based on genetic diagnostic technologies in order to process the Genomes of different living organisms which is why these approaches are not easily adaptable for a very large scientific community. Furthermore, these approaches are skill and labor intensive, hence these approaches are not appropriate for efficient genetic sequence analysis. This dissertation utilizes the powers of Artificial Intelligence methods to develop a generic framework competent in performing efficient Genomics and Proteomics sequence analysis at large-scale. The generic nature of the presented framework makes it applicable to a variety of Genomics (Histone occupancy prediction, Histone and DNA modification prediction, RNA classification and subcellular location prediction) and Proteomics (host protein-protein and viral host protein-protein interaction prediction) sequence analysis tasks across multiple species.

1.1 Motivation

The main motivation behind this study is author's observation that there is a very limited availability of robust computational frameworks for Genomics and Proteomics sequence analysis despite their immense need in research and industry. Aim of this work is to develop a generic computational framework, capable of performing different types of Genomics and Proteomics sequence analysis tasks. By unraveling the nature of diverse sequence analysis tasks, it is observed that majority of the sequence analysis tasks are related to classification problems such as enhancer identification and their strength prediction, histone occupancy detection, acetylation and methylation level prediction in histone and DNA bio-molecules, RNA classification and their subcellular location prediction, host protein-protein and viral host protein-protein interaction prediction. Building on this, author decides to develop automatic methods in generic framework which can identify and learn discriminative information of nucleic and amino acids from raw sequences and use such information to categorize sequences into relevant classes.

Furthermore, author observes that few sequence analysis tasks such as antibody sequence performance prediction and prediction of channelrhodopsins proteins within plasma membrane fall under the hood of regression problems. In this regard, author aims to develop efficient methods which can learn and extract distribution of nucleic and amino acids in the raw sequences and use this information to predict float values associated with the sequences.

Following the working paradigm of machine and deep learning approaches that do not operate on raw DNA, RNA and protein sequences, author observes that a variety of sequence encoders have been proposed for different biomolecules that convert raw biomedical sequences into statistical vectors by extracting comprehensive information of nucleic and amino acids. In

addition, author observes the most dominantly used machine learning classifiers and regressors across different sequence analysis tasks. In order to truly reveal the effectiveness of new predictive methods, researchers need to perform a comprehensive benchmarking of new methods using multiple approaches as baseline as well as an unbiased performance comparison of new predictive methods with existing predictive methods for various sequence analysis tasks. In such a comprehensive analysis, researchers face multiple challenges such as time constraint to implement existing approaches from scratch due to unavailability of source codes, missing or hard to find hyperparameters and model training information, etc. To facilitate a more generic framework which can potentially make the lives of biomedical researchers and practitioners easier, author aims to implement all sequence encoders published for three different biomolecules including DNA, RNA and protein, the most widely used machine learning classifiers, regressors and evaluation metrics, all at one platform. Furthermore, author's aim is to organize the pool of existing sequence encoders and predictors in an automated setting which exhaustively evaluates all possible combinations of sequence encoders and predictors to facilitate comprehensive benchmarking performance. This framework facilitates development of practical end-to-end pipelines for certain biomedical applications without considering any human effort at any level.

Moreover, author finds that existing sequence encoders lack in extracting comprehensive discriminative and positional information of nucleic or amino acids while transforming raw sequences into statistical vectors. Considering, Genomics and Proteomics sequences have large number of constant regions where the distribution of nucleic and amino acids remains almost the same and very few regions have distinguishing distribution of nucleic and amino acids which is important for accurate sequence analysis. Author aims to develop and incorporate novel sequence encoding methods in generic framework, capable of finding position specific discriminative distribution of nucleic and amino acids which largely assist the predictor in achieving optimal performance for diverse DNA, RNA and protein sequence analysis tasks.

Furthermore, author notes that few biomedical researchers treat the processing of DNA, RNA and protein sequences similar to Natural Language Processing (NLP) and adopt different neural embedding methods from NLP domain. Author aims to address this research direction by facilitating multiple neural embedding generation methods in generic framework which have shown great performance in different NLP and Bioinformatics tasks, in order to quantify their performance potential for Genomics and Proteomics sequence analysis tasks.

With the availability of huge number of methods in published literature, researchers and practitioners preference for quick customization of predictive pipeline such as combining the benefits of multiple sequence encoders to enhance predictive performance, generalizability, etc., is inevitable. Author aspires to provide automated settings in generic framework which are capable of combining homogeneous as well as heterogeneous sequence encoders using different strategies to generate more appropriate statistical vectors. Furthermore, apart from traditional deep learning predictors, author aims to develop and incorporate a variety of novel deep learning

based classifiers that can achieve state-of-the-art performance in different types of Genomics and Proteomics sequence analysis tasks related to binary/multi-class or multi-label classification.

1.2 Problem Statement

Unlike the tasks related to different domains (e.g., NLP and energy forecasting), Genomics and Proteomics sequence analysis tasks are highly sensitive by nature where false analysis can have very far-reaching as well as serious repercussions on human lives. To perform an accurate and large scale analysis of different biomolecules sequences (DNA, RNA, protein) related to multiple species, development of robust Artificial Intelligence (AI) based approaches for different biomedical classification and regression tasks face multiple challenges.

Raw sequences of DNA/RNA consist of repetitive patterns of 4 unique nucleic acids while protein raw sequences are made up of repetitive patterns of 20 unique amino acids. For any particular sequence analysis task, distribution of unique nucleic or amino acids remain same in sequence samples that belong to same class, whereas this distribution slightly varies among sequence samples of different classes. Therefore, one of the challenges which AI based approaches need to address is to develop an effective sequence encoder which can generate statistical vectors of raw sequences by capturing comprehensive local and global relations of nucleic or amino acids. The performance of AI predictors mainly relies on the quality of generated statistical vectors. A simple classifier can precisely discriminate sequence samples into different classes when it is fed with statistical vectors which contain comprehensive discriminative patterns. Contrarily, a sophisticated classifier may fail to discriminate sequence samples into multiple classes when it is fed with in-effective statistical representations.

Another challenge that impacts the performance of AI predictors is how they deal with variable length sequence samples. For some sequence analysis tasks, the length of sequences largely varies from hundreds to thousands of nucleic or amino acids. Most of the sequence encoding methods either require fixed-length sequences as input, or they generate variable length statistical vectors for input sequences. Few sequence encoders take variable length sequences and generate fixed-length statistical vectors of sequences. Particularly, diverse type of textual classification approaches in the domain of NLP and most of the existing DNA/RNA or protein sequence classification approaches generate fixed-length sequence samples using three different tricks. Copy padding trick maps shorter length sequence samples to maximum possible sequence length by adding a constant. Sequence truncation trick discards words or nucleic/amino acids in sequence samples whose lengths are larger than minimum possible sequence length. The average length based copy padding/truncation method computes average length of sequence samples and according to that length, it discards words or nucleic/amino acids from the sequence samples that are larger than defined length and adds constant to the sequence samples that have length less than defined length. However, pre-processing of biomedical sequences is different from the

preprocessing of textual sequences, as unlike textual sequences, lengths of biomedical sequences samples vary by thousands of nucleic or amino acids. Here, addition of a significant number of constant to shorter length sequences makes them almost similar whereas truncation may discard important regions where distribution of nucleic/amino acids is highly variable and useful. Building on this, another open challenge for AI based approaches is to generate fixed-length biomedical sequences without losing important distribution of nucleic/amino acids.

In addition, most genetic sequence analysis datasets are highly imbalanced in nature where few target classes have a large number of sequence samples while other classes have few numbers of sequence samples. Furthermore, the magnitude and diversity of features importance for accurately predicting target classes also vary across multiple species, indicating one set of features may be good enough to determine target classes for humans but not sufficiently good to predict same target classes for other species like mouse. Hence, third biggest challenge for AI based Genomics and Proteomics sequence analysis approaches is to develop a more robust and generalized machine or deep learning model which can precisely and most accurately predict target classes across multiple species.

In order to effectively resolve the aforementioned challenges, the prime focus of this dissertation is to develop a robust and adaptable machine or deep learning predictors for various Genomics and Proteomics sequence analysis tasks which make best use of different strategies to precisely predict target classes of sequence samples across a variety of species. Furthermore, another focus of this dissertation is to facilitate comprehensive existing approaches and newly developed approaches at a single platform with customization capabilities to empower wide scientific community.

1.3 Research Questions and Goals

Following are the research questions that are investigated to make the proposed framework generic and capable of developing diverse types of practical applications for Genomics and Proteomics sequence analysis tasks.

- 1 What is the most effective sequence preprocessing strategy to fix the length of fluctuating biomedical sequences?

Goal: To fix the length of highly variable DNA/RNA and protein sequences without losing the important information about distribution of nucleic or amino acids, this dissertation investigates the performance impact of traditional fixed-length sequence generation approaches. Through this analysis, it develops a novel mechanism to generate fixed-length sequences by keeping only the most informative distribution of nucleic/amino acids.

- 2 How to generate comprehensive statistical representation of DNA/RNA and protein sequences by precisely capturing the distribution and heterogeneous relations of nucleic/amino acids

in the raw sequences?

Goal: This research explores the potential of existing encoding methods that made use of physicochemical properties, occurrence frequency, local and global context information of nucleic/amino acids to generate a rich statistical representation of DNA/RNA and protein sequences. Through this exploration and by performing a deep performance comparison of existing encoders, this research provides two different paradigms to generate comprehensive statistical representation of raw sequences. First paradigm goal is to reap the benefits of different top performing encoding methods. Second paradigm focuses on the development of novel standalone encoders capable of generating statistical representation by capturing occurrence and positional information of nucleic or amino acids.

3 Is it possible to develop a unified deep learning based predictor that can be used for multiple species?

Goal: Here the goal is to deeply evaluate the functional paradigms and performances of existing computational approaches over multiple datasets belonging to distinct species and identify the factors which contributed to decline the generalizability of existing approaches. This analysis is performed with an aim to develop an improved approach by extracting task and specie specific features and perform a rich performance comparison with other multi-species genetic sequence analysis approaches.

4 Is it possible to highlight important patterns of nucleic/amino acids in the raw sequences to interpret classifier decisions?

Goal: One of the many important goals of this dissertation is to develop interpretable deep learning based predictors that can explain which nucleic/amino acid distribution patterns are most informative in the sequences and what exactly they represent, which is a crucial information to make predictor decisions more transparent and understandable for wide scientific community.

5 Do the raw DNA/RNA and protein sequences contain sufficient information to perform computational Genomics and Proteomics sequence analysis?

Goal: The aim of this research is to compare the efficacy and potential of raw sequences based computational approaches with other computational approaches which rely on cellular, molecular, or structural profiles of biomolecules to identify them, their families, their localization inside the cell and the probability to interact with other biomolecules.

1.4 Contributions

The main contribution of this dissertation is the conceptualization and implementation of a computational framework using the power of Artificial intelligence approaches. The generic

nature of the presented framework makes it applicable to a variety of Genomics (DNA, RNA) and Proteomics (protein) sequence analysis tasks. A premier on presented framework is that, it only requires raw DNA/RNA and protein sequences to perform a more accurate biomedical sequence analysis. A high-level overview of dissertation contributions is provided in the following subsections.

1.4.1 Genomics & Proteomics Sequence Analysis

A high level overview in terms of Genomics & Proteomics Sequence analysis framework needs and manifold contributions of this dissertation for the development of computational framework is summarized below.

1.4.1.1 Contribution 1: An Efficient Automated Machine Learning Framework for Genomics and Proteomics Sequences Analysis

Background:

An efficient comprehensive Genomics and Proteomics sequence analysis is important to understand the functional dynamics of different biomolecules, execution of various biological processes and irregularities in core cellular behaviors. To provide automated methods for biomedical sequence analysis, several computational frameworks have been developed. However, these frameworks are not generic in nature as they only facilitate a handful of existing sequence encoders and predictors for two types; of tasks binary or multi-class classification and clustering. They neither provide regressors nor multi-label classifiers. Furthermore, under the hood of binary classification, they also lack to facilitate robust pipelines for biomolecule interaction prediction tasks. The influx of biological sequences demands a generic computational framework which can accurately perform multi-dimensional analysis of Genomics and Proteomics sequences such as predicting multiple subcellular localization compartments of biomolecules, biomolecules interaction possibility, antibody sequence performance values, etc.

Contribution:

This dissertation provides a computational framework capable of automatically developing pipelines for accurate biomedical sequence analysis tasks related to 4 different major categories: regression, clustering, binary/multi-class classification and multi-label classification. The proposed computational framework facilitates 221 existing and novel DNA, RNA and protein sequence encoders, 25 dimensionality reduction and feature selection methods, 25 widely used machine learning classifiers/regressors and a variety of deep learning based architectures. To ensure the generalization and applicability of the proposed framework, it is evaluated on different tasks related to multiple species. Furthermore, it facilitates 5 web applications¹ capable of performing efficient histone sequence analysis, RNA subcellular localization prediction, host

¹https://sds_genetic_analysis.opendfki.de/

protein-protein interaction prediction and viral-host protein-protein interaction prediction on new sequence data related to multiple species.

1.4.2 Genomics Sequence Analysis

In Genomics sequence analysis, proposed framework is evaluated on 5 different classification tasks and produces state-of-the-art performances.

1.4.2.1 Contribution 2: DNA Modification Prediction

Background:

Accurate prediction of DNA modifications is essential to explore and discern the process of cell differentiation, gene expression and epigenetic regulation. Pre-dominant DNA modification predictors are suitable to predict only one particular type of modification. Only two generic predictors support the prediction of multiple DNA modifications. Both type specific and generic DNA modification predictors produce suboptimal performance across multiple species. This is mainly due to the use of ineffective sequence encoding methods based on nucleotide frequency and physicochemical properties that lack to capture comprehensive discriminative nucleotide distributions.

Contribution:

This dissertation develops a generic approach “DNA-MP” to most accurately predict 4-Methylcytosine (4mc), 5-Hydroxymethylcytosine (5hmc) and N6-methyladenine (6mA) modifications across multiple species only using raw DNA sequences. DNA-MP makes use of a novel encoding method “POCD-ND” to generate a comprehensive statistical representation of DNA sequences by capturing position specific discriminative distribution of nucleotides and a deep forest classifier for modification prediction.

To perform a large scale performance comparison of proposed sequence encoder with existing 31 most widely used sequence encoders, an intrinsic performance comparison is performed by visualizing the feature space of proposed sequence encoder and existing 31 sequence encoders. Extrinsic performance comparison of proposed sequence encoder and existing 31 sequence encoders is performed using 10 different machine learning classifiers on 17 benchmark DNA modification prediction datasets of 12 different species. The proposed DNA-MP predictor outperforms state-of-the-art type-specific and generic modification predictors by an average accuracy of 7% across 4mc datasets, 1.35% across 5hmc datasets and 10% for 6ma datasets. DNA-MP is deployed as an interactive web application at ² which can be used to predict DNA modifications on the go and can be adapted for other epigenetic modification prediction tasks.

²https://sds_genetic_analysis.opendfki.de/DNA_Modifications/

1.4.2.2 Contribution 3: Prediction of Histone Occupancy and Modifications along with Enhancer Identification and their Strength Detection

Background:

In order to deeply understand and control the production of proteins and for the development of novel genetic therapies to treat complicated diseases like Cancer by addressing the issues of under or over expression of genes, determining histone occupancy, modifications, enhancers and their strengths are extremely important sequence analysis tasks. To perform this analysis, existing computational approaches produce suboptimal predictive performance due to their inability to extract discriminative features from DNA sequences. Furthermore, existing histone sequence analysis approaches face the huge overhead of training separate predictors for different histone markers.

Contribution:

Following the success of the supervised FastText classifier in the domain of Natural Language Processing, this dissertation develops a novel deep learning predictor Histone-Net [22] which can more accurately predict histone occupancy, methylation and acetylation levels across 10 different datasets in intra-domain and cross-domain binary classification paradigms. Histone-Net outperforms state-of-the-art histone occupancy, acetylation and methylation prediction approach by an average accuracy of 7.5% on 10 different datasets. Another promising contribution is the development of a multi-label classification dataset which will accelerate the research direction of simultaneously predicting histone occupancy, acetylation and methylation levels across 10 histone markers using a single predictor. Apart from traditional binary classification paradigm, proposed Histone-Net approach also produces promising performance in multi-label classification paradigm. To analyze generalizability and applicability of the proposed Histone-Net approach in other Genomics sequence analysis tasks, it is evaluated for enhancers identification and their strength prediction tasks [20] where it outperforms state-of-the-art enhancer sequence analysis predictor on 2 benchmark datasets by an average accuracy of 9%. To enable the wide scientific community to perform efficient histone sequence analysis under the hood of different paradigms, Histone-Net is deployed as a web application at ³.

1.4.2.3 Contribution 4: Small Non-Coding RNA Classification

Background:

The development of biomolecular devices which interface with the biological systems to unveil novel insights and produce new functions is one of the most promising goals of synthetic biology. Accurate classification of small non-coding RNAs (sncRNAs) is a pre-requisite for the development of powerful riboregulators competent in providing modular, tunable, as well as precise control of the gene expression. Existing sncRNA classification approaches mark suboptimal predictive

³<https://histone.opendfki.de/>

performance due to their dependency on secondary structural features which lack to capture comprehensive local and global relations of nucleotides.

Contribution:

This dissertation develops a robust and precise deep learning model namely RPC-snRC [28] for accurate classification of small ncRNAs into relevant families by utilizing their primary sequences. Contrary to existing predictors, RPC-snRC makes use of working paradigm of DenseNet architecture which is competent to learn hierarchical representation of features by providing comprehensive paths for the flow of gradients to all the previous layers. To prove the effectiveness of proper gradient flow while learning better representations, similar to DenseNet architecture, two ResNet architectures are adopted to perform small non-coding RNA classification. RPC-snRC outperforms adapted approaches by an accuracy margin of 3% and state-of-the-art approaches with an accuracy margin of 10% on a public benchmark dataset.

1.4.2.4 Contribution 5: Circular RNA Identification

Background:

Circular RNAs (circRNAs) have emerged as useful regulators for physiological development and disease parthenogenesis due to their cell specific expression associations with a plethora of biological functions. Accurate detection of circRNAs is essential to decode transcription and splicing regulation which provide new opportunities for clinical applications development. State-of-the-art circular RNA identification approach relies on handcrafted features which do not only increase the feature space but also extract irrelevant and redundant features.

Contribution:

This dissertation develops an end-to-end deep learning framework namely CircNet [372] which utilizes encoder-decoder strategy to learn lower-dimensional latent representations and convolutional operations to automatically extract the most discriminative features from latent feature space. Another distinguishing contribution is that, unlike existing circRNA identification approaches, CircNet finds and highlights different regions of genome which contain the most useful information for accurate detection of circRNAs. CircNet significantly outperforms state-of-the-art approaches with a considerable margin of 10.29% in terms of F1 measure on a public benchmark dataset.

1.4.2.5 Contribution 6: RNA Subcellular Localization Prediction

Background:

Subcellular localization of Ribonucleic Acid (RNA) molecules provides significant insights into the functionality of RNAs and helps to explore their associations with various diseases. Pre-dominantly developed single-compartment localization predictors (SCLPs) lack to demystify RNA associations with diverse biochemical and pathological processes mainly happen through RNA co-localization in multiple compartments. Few multi-compartment localization predictors

(MCLPs) mark suboptimal generalizability. Furthermore, both types of predictors have below par practical significance due to a low degree of model explainability.

Contribution:

To empower the process of RNA subcellular localization prediction, contributions of this dissertation are manifold. 1) It briefly, illustrates the working paradigms of existing RNA subcellular localization predictors in terms of their pros and cons [19]. 2) It develops two computational predictors for miRNA subcellular localization prediction [24] [27]. 3) It develops first computational predictor for circular RNA subcellular localization prediction [306]. Considering the need of a robust and generalized predictor, it develops a robust deep learning predictor namely EL-RMLocNet [23] which can more accurately predict multi-compartment localizations of 4 different RNA classes across two different species. EL-RMLocNet makes use of a unique graph based encoding method to capture comprehensive local and global interaction patterns and translational in-variances of nucleotides. Furthermore, it reaps the benefits of Long Short-Term Memory (LSTM) and attention layers to capture the most informative features and their heterogeneous relations which are important for accurate multi-compartment localization prediction for target species. EL-RMLocNet outperforms state-of-the-art predictor by an average accuracy of 8% for Homo sapiens species and 6% for Mus Musculus species. To the best of author knowledge, this is the very first explainable predictor which is competent in handling multiple RNA types across different species in a more complicated multi-compartment subcellular localization prediction setting where it highlights which nucleic acids patterns are responsible for certain model decisions. Furthermore, it facilitates a web application ⁴ to enable the researchers and practitioners to predict multi-compartment localizations of new RNA sequences.

1.4.3 Proteomics Sequence Analysis

In Proteomics sequence analysis, proposed framework is evaluated on 2 different applications areas where it produces state-of-the-art performances.

1.4.3.1 Contribution 7: Protein-Protein Interaction Prediction

Background:

Protein-protein interaction (PPI) prediction is essential to understand the functions of proteins in various biological processes and their roles in the development, progression and treatment of different diseases. Existing PPI prediction approaches have suboptimal predictive performance and practical significance as they lack the ability to extract comprehensive discriminative features and aptitude to explain the decision making of predictor.

Contribution:

This dissertation develops a novel PPI predictor namely ADH-PPI that uses Long Short-Term Memory layer to extract short and long range dependencies of features, convolutional layer to

⁴https://rna_subcellular_predictor.opendfki.de/

extract comprehensive hidden informative features and Self-Attention layer to focus on most valuable features which contribute the most for accurate PPI prediction. Another, promising contribution is that it generates the k-mer amino acid representations using FastText in an unsupervised manner. In addition, it performs a comprehensive performance comparison using 2 benchmark core datasets and 4 independent test sets related to different species where ADH-PPI outperforms existing PPI predictors by an overall accuracy of 4% and 7%, respectively. One more contribution is that it finds and highlights which amino acid distributions are more important for accurate PPI prediction across multiple species which helps to decode the decision making of predictor. It facilitates the PPI prediction web application⁵ which can be used to make accurate and explainable predictions in multiple species.

1.4.3.2 Contribution 8: Virus-Host Protein-Protein Interaction Prediction

Background:

Viral-host protein-protein interaction (VHPPI) analysis is essential to decode molecular mechanism of viral pathogens and host immunity processes which eventually helps to control the viral diseases and optimize therapeutics. Multiple AI-based approaches have been developed to predict VH-PPIs interactions across a wide range of viruses and hosts, however, these approaches produced better performance only for specific types of hosts and viruses. The influx of viruses from heterogeneous sources including farm or wild animals, arthropods, etc., leading to plethora of deadly infectious diseases implies the desperate need of a generic predictor that can efficiently determine the viral-host PPIs across diverse hosts and viral species.

Contribution:

To supplement the process of Viral-host protein-protein interaction prediction, this dissertation develops LCGA-VHPPI predictor [21] that makes use of a deep forest classifier and novel sequence encoding method capable of capturing local and global context of amino acids. Furthermore, it develops a robust meta predictor capable of more accurately predicting VHPPI across multiple hosts and viruses. Proposed meta predictor makes use of two well-known encoding methods APAAC and QS order that captures and encodes sequence order and distributional information of amino acids to statistical vectors. Feature agglomeration method is utilized to transform original feature space to more comprehensive feature space. Random forest and Extra tree classifiers are trained on optimized feature space by combining encodings generated by APAAC and QS order encoders. Furthermore, predictions of both classifiers are utilized to train SVM classifier that makes final predictions. Proposed meta predictor is evaluated, over 7 different benchmark datasets, where it outperforms existing VH-PPI predictors with average performance figures of 3.864%, 8.434%, 8.857% and 7.365% in terms of accuracy, MCC, precision and sensitivity, respectively. Furthermore, it develops an interactive web server⁶, which enables

⁵https://sds_genetic_analysis.opendfki.de/PPI/

⁶https://sds_genetic_analysis.opendfki.de/MP-VHPPI/

the scientific community to predict viral-host PPIs across multiple viral and host species in no time.

1.5 Dissertation Overview

This dissertation makes great efforts to develop cutting-edge Artificial Intelligence based automated Genomics and Proteomics sequence analysis approaches and centralise them in a single platform to ease the lives of different end users. It also makes notable efforts to make sequence analysis approaches applicable to different biomolecules and species using only raw genetic of sequences. This dissertation is organized into 10 different chapters. After providing a bird's eye view the core motivations behind biological sequence analysis and author major contributions to address different problems in Chapter 1. Chapter 2 provides the complete workflow of proposed framework and performs a detailed comparison of proposed framework with existing frameworks on the basis of different core qualities and acceptance criteria. Chapter 3 discusses the importance of determining DNA modifications to better understand the process of cell differentiation and gene expression regulation. It also sheds light on a novel predictor developed to infer three different modifications: 4-Methylcytosine (4mc), 5-Hydroxymethylcytosine (5hmc) and N6-methyladenine (6mA) across multiple species and its overall effectiveness as compared to existing single type and multi-type DNA modifications predictors. Chapter 4 describes the significance of histone and enhancer sequence analysis for controlling the production of proteins and development of naive genetic therapies for complex diseases. It discusses a novel predictor developed for histone occupancy, histone modifications, enhancer identification and their strength prediction tasks and its efficacy as compared to existing predictors. Chapter 5 discusses the worth of small non-coding RNAs classification for the development of powerful riboregulators. It describes the working paradigm of novel predictor developed for classifying small non-coding RNAs into their respective families and their practical significance as compared to existing predictors. Chapter 6 describes the emerging regulatory roles of circular RNA in physiological development and disease parthenogenesis. It discusses the novel predictor developed for circular RNA identification and its performance as compared to state-of-the-art predictors. Chapter 7 describes the importance of identifying multi-compartment subcellular localization of different RNAs to better understand their functions and associations with diverse diseases. It provides details about a novel predictor developed to infer multi-compartment subcellular localization of different RNAs, its generalization across multiple species and overall efficacy as compared to existing predictors. Chapters 8 and 9 discuss the significance of determining host protein-protein interactions and viral-host protein-protein interactions to understand the functions of proteins, molecular mechanisms of viral pathogens and host immunity processes. These chapters shed light on the working paradigms of developed novel predictors, their performances on different datasets and species and overall effectiveness as compared to existing predictors. Final chapter

10 provides conclusive remarks, major limitations of developed novel predictors and compelling future directions of current work.

A GENERIC FRAMEWORK FOR GENOMICS (DNA, RNA) AND PROTEOMICS (PROTEIN) SEQUENCE ANALYSIS

Genetic sequence analysis technologies that are competent in analyzing a large number of DNA molecules in a massively parallel manner are labeled as next-generation sequencing (NGS) technologies [17, 152]. NGS technologies enable the exploration of hundreds and thousands of genes to demystify the associations of genetic variations with different diseases and biological phenomenon [17, 152]. NGS technologies provide the basis to control the regulation of gene expression and have revolutionized the development of new applications in clinical and genomic research, reproductive health, environmental, agricultural and forensic science [64, 104, 145, 175, 234, 364, 465].

The high throughput paradigm of NGS technologies has given birth to exponentially increasing Genomics and Proteomics data which is of great significance [120]. The humongous Genomics and Proteomics data analysis is useful to comprehensively understand diverse biological processes, gene expression patterns and their associations with the initiation, progression and treatment of different diseases through optimizing therapeutics [318]. An accurate exploration of genetic data will not only take our understanding regarding life science to an advanced level that we cannot imagine yet, but it will also make personalized healthcare a reality [318].

Considering the aptitude of Artificial Intelligence (AI) approaches to automatically extract important hidden patterns, these approaches have been extensively utilized to explore the hidden potential of biological sequences for the establishment of economical large-scale Genomics and Proteomics sequence analysis landscape [106, 289]. Within this landscape, a closer look at the problem nature of various genetic sequence analysis tasks reveals that most of the tasks fall under three different paradigms: 1) Classification, 2) Clustering and 3) Regression. In classification, the primary goal of AI approaches is to forecast discrete values such as families of biomolecules,

possibility of biomolecules interaction, localization of biomolecules within the cell, distinguishing normal cells from disease cells [154, 430], etc. In clustering, the main focus of AI approaches is to group different sequences that have similar characteristics [154, 430]. Whereas, in regression, the focus of AI based genetic sequence analysis approaches is to estimate continuous values such as genomic prediction of disease risk, quantifying the drug response and estimating DNA copy number variations [154, 430].

With an aim to perform accurate genetic sequence analysis, diverse AI frameworks have been developed [29, 45, 48, 74, 85–87, 224, 267]. Despite the diversity, genetic sequence analysis pipelines of all the frameworks are based on five different modules 1) Sequence data collection and preprocessing, 2) Feature Representation, 3) Feature Engineering, 4) Predictor construction, 5) Predictor Evaluation and intrinsic or extrinsic performance visualization. To facilitate researchers by providing diverse types of algorithms related to all 5 modules at a single platform, according to our best knowledge, 9 different generic frameworks have been developed.

For instance, Selene [74] is a command-line Pytorch based deep learning framework which provides sequence sampling module, existing model training and improvement modules for only multi-class sequence classification tasks. Another framework Janggu [224] provides few existing encoding schemes and deep learning models for predicting transcription factors, chromatin effects and promoter usage, most of which fall under the hood of multi-class classification tasks. Kipoi [29] provides 2,194 ready-to-use trained deep learning based predictive models for transcriptional and post-transcriptional gene regulation. BioSeq-Analysis [45] is the first machine learning-based genetic sequence analysis platform that supports the development of end-to-end pipelines for classification tasks. Its more recent version called BioSeq-Analysis-2.0 [267], contains 35 sequence encoding algorithms to further improve the process of analyzing genetic sequences. In 2018, researchers from different institutes collaborated to release the first comprehensive framework named iFeature [86]. iFeature provides the implementation of 53 feature representation methods for only protein and peptide sequences. Later in 2020, an extended version of iFeature named iLearn [87] has been published. iLearn is designed to generate statistical representations of all three DNA, RNA and protein sequences. It also contains some feature engineering approaches along with traditional machine learning classifiers. Recently, in 2021, the advanced version of iLearn named iLearn plus [85] is released that provides more encoding methods to transform DNA, RNA and protein sequences into statistical vectors. Recently, another framework named math feature [48] has been published, which contains a variety of feature encoding methods and a pool of machine learning classifiers.

Prime focus of existing computational frameworks is to provide an effective platform which researchers and practitioners can use to perform diverse types of Genomics and Proteomics sequence analysis tasks. To accomplish this goal, existing computational frameworks facilitate different sequence encoding methods, however, there is not even a single computational framework that provides word embedding methods to generate statistical representations of Genomics

and Proteomics sequences despite considering their success in Natural Language Processing and Bioinformatics. Furthermore, these frameworks only support the development of pipelines for two major tasks namely clustering and binary/multi-class classification. Also, these frameworks do not facilitate pipelines for biomolecule interaction prediction tasks despite considering the fact that pre-dominant biomolecule interaction prediction tasks such as host protein-protein interaction prediction, virus-host protein-protein interaction prediction, RNA-protein interaction prediction, lncRNA-miRNA interaction prediction, etc., are binary in nature. Another major pitfall is existing frameworks neither support regression nor multi-label classification, neglecting that a significant number of tasks fall under the hood of multi-label classification such as RNA multi-compartment subcellular localization prediction and protein multi-compartment subcellular localization prediction. It is widely accepted that feature engineering is important to achieve good predictive performance. However, existing frameworks only support a few generic feature engineering methods. One more downfall is existing frameworks have suboptimal practical significance because they are evaluated on limited case studies and few species.

2.1 Functional Scope and Description of Proposed Generic Framework

Considering the diversity of Genomics and Proteomics sequence analysis tasks and the sensitive nature of sequence analysis which can have serious repercussions on human health in the case of incorrect findings and suboptimal accuracy, this dissertation develops a robust generic framework for Genomics and Proteomics Sequence Analysis (GFGPA). GFGPA framework supports the automated development of predictive sequence analysis pipelines and meta predictors to handle tasks of 4 different major categories including regression, clustering, binary/multi-class classification and multi-label classification. Besides facilitating all existing DNA, RNA and protein sequence encoders, widely used dimensionality reduction algorithms, feature selection algorithms, machine learning classifiers, different neural architecture based deep learning predictors, it also facilitates novel sequence encoders and deep learning predictors to perform more accurate sequence analysis across multiple species.

Figure 2.1 illustrates complete workflow of the proposed generic GFGPA framework. To perform any genetic sequence analysis task after data collection in preprocessing stage, an important task is to fix the length of sequence samples. Genetic sequence analysis tasks can be considered similar to Natural language processing tasks. Just like text classification where the words present within sentences determine the context and make sense of the sentences, in genetic sequence analysis, we have a string of letters that is segregated into small subsequences where each subsequence acts as a feature to make biological sense of sequences. However, in text classification, length of documents does not vary as much as sequence samples length varies in genetic analysis tasks. Traditional machine and deep learning algorithms require fixed-length

sequence samples as an input. Hence, in the proposed GFGPA framework, we incorporate 3 traditional fixed-length generation approaches copy padding at maximum length, truncation at minimum length and copy padding or truncation sequences at average length. Furthermore, we notice that most informative distributional information of nucleic and amino acids lies at the starting region and at the ending region of sequences. To more effectively handle high length variability of genomic and proteomic sequences, in the GFGPA framework, we integrate a unique way of generating fixed-length sequences based on the most informative bins of sequences.

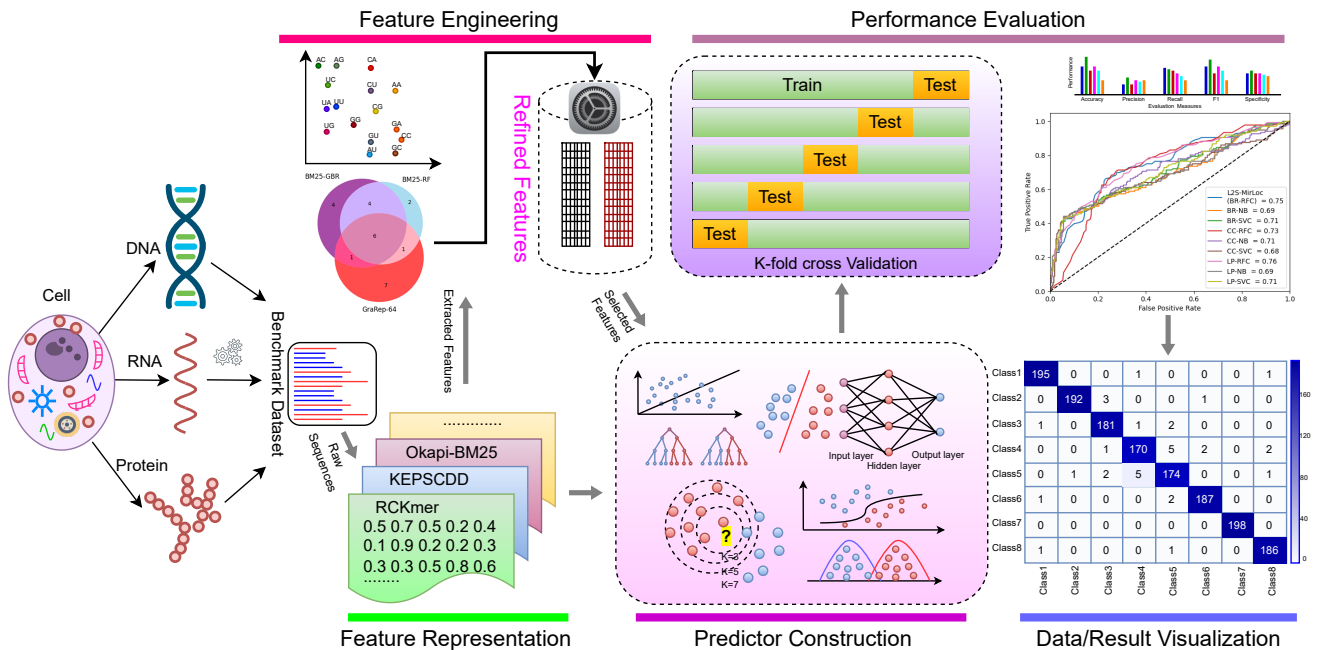


Figure 2.1: The complete workflow of generic GFGPA framework proposed for efficient Genomics and Proteomics sequence analysis.

After generating fixed-length sequences, the next step is to generate k-mers of the sequences. To generate k-mers, we slide a fixed-size window with particular stride size over sequences. If the window size is equal to the stride size, non-overlapping k-mers are generated and by taking different window and stride sizes, overlapping k-mers are generated. Proposed framework supports the generation of overlapping as well as non-overlapping k-mers of sequences.

After k-mer generation, next phase is to transform k-mer sequences into statistical vectors. The proposed framework supports several published heterogeneous statistical representation generation approaches and also provides novel statistical representation generation methods, details of which are given in section 2.1.1. The statistical vectors of k-mer sequences are either directly passed to deep learning based predictors which automate the process of feature engineering or they are passed to feature engineering module before feeding to machine learning based predictors. Feature engineering module analyzes whether the statistical vectors contain redundant or

noisy features. It is widely accepted that not all input features significantly contribute in the accurate estimation of target class labels for hand on task [163]. Feeding predictors with highly informative features significantly assist them to find useful correlations for accurate predictive modeling. It is commonly acknowledged that even a simple predictor performs better when it is fed with a relevant and informative subset of input features. Whereas, even a sophisticated predictor underperforms on account of noisy features.

Generally, there are two prominent ways to select the most informative subset of features from noisy data. One way is to apply feature selection that removes irrelevant and redundant features and retain the most informative features from original subset of features. While other way is to apply dimensionality reduction approaches that transform original feature space into reduced feature space by eliminating redundant correlations of features. Proposed GFGPA framework facilitates a variety of feature selection and dimensionality reduction algorithms that are summarized in section 2.1.2.

Using the optimized statistical vectors, the GFGPA framework supports the training and evaluation of a variety of existing and novel predictors for accurate genetic sequence analysis, details of which are summarized in section 2.1.3. With an aim to most effectively quantify the performance of predictors, the comprehensive performance evaluation metrics used by proposed GFGPA framework for different genetic sequence analysis tasks are briefly described in section 2.1.4.

2.1.1 Feature Representation Module

In comparison to existing frameworks, a high-level contribution at this particular stage is that it contains novel sequence encoder that is briefly described in chapter 3. Furthermore, following the success of word embedding methods in NLP, the author incorporated 11 diverse types of word embedding generation methods in the developed framework, these methods are not available in any of the existing framework. A comprehensive working paradigm of word embedding methods is described in chapters 7 and 8. Similarly, in NLP domain from information retrieval task, author adopted a well-known method okapi-BM25 for k-mers representation. This method is used for in house DFKI industrial project in the development of an application for antibody sequence performance prediction. Overall, sequence representation module contains 35 generic encoders that can be utilized to generate statistical representation of all three types of sequences DNA, RNA and protein. Furthermore, it facilitates 29, 23 and 64 encoding methods for specifically DNA, RNA and protein sequences, respectively. A comprehensive details about categories and names of generic encoders are illustrated in Table 2.1, DNA and RNA based encoding methods are summarized in Table 2.2 and encoding methods for protein sequences are summarized in Table 2.3.

Most of the DNA/RNA and protein sequence encoding methods make use of pre-computed physicochemical properties. In these encoding methods, it is important to find which particular

set of properties should be used to generate statistical representations of raw sequences. For instance, AAINDEX encoder has 512 physicochemical properties, existing frameworks require manual input of names for a particular set of properties out of 512 which should be used to generate statistical representation of protein sequences. Manual selection of different subsets of properties is a tedious and time consuming task.

To fully utilize the potential of physicochemical properties based encoders, GFGPA framework facilitates a strategy similar to forward feature selection method, with an aim to find out the most appropriate physicochemical properties inside the encoders. For instance, from 3 properties of APAAC encoder, first it generates statistical vectors by using one property and compute performance of classifier. Similarly, it repeats the same process for the second and third properties in order to record the performance of classifier. On the basis of higher performance, it takes the property specific statistical vectors and combine them with the second best performing property vectors. This is followed by the evaluation on the basis of combined features, if this does not yield any performance gains then the iterative process stops and individual property based statistical vectors with the highest performance are selected. In contrast, if there are any performance gains with such combinations then the combined encodings are retained and utilized further. A more comprehensive detail about property selection method is provided in chapter 9.

Apart from property selection, a similar working paradigm is also developed to reap the benefits of diverse types of sequence encoding methods. Proposed GFGPA framework provides option to run different encoders in a batch manner, where from the batch of encoders it first generates and evaluates the encodings of each encoder one by one. Further, it combines the statistical vectors of different encoders similar to property selection criterion.

2.1.2 Feature Engineering Module

In the domain of feature engineering, researchers have always been striving to develop innovative techniques for selecting relevant and more appropriate features. In this marathon, a variety of feature selection and dimensionality reduction approaches have been proposed. This section briefly describes different feature selection and dimensionality reduction approaches that GFGPA framework facilitates to scientific community.

2.1.2.1 Feature Selection

The working paradigm of feature selection approaches can be categorized into three main classes: filter [163], wrapper [221] and embedded [236]. To precisely analyze which feature selection approach is more appropriate to find optimal subset of features for a particular Genomics and Proteomics sequence analysis task, the proposed framework facilitates the top performing approaches from each of the three categories. The filter based feature selection techniques filter the corpus features based on their general properties, such as correlation with the dependent variable. It is considered the fastest and the best approach when the corpus has a large number

2.1. FUNCTIONAL SCOPE AND DESCRIPTION OF PROPOSED GENERIC FRAMEWORK

Table 2.1: Generic encoding methods for DNA, RNA and protein sequences

Descriptor Category	DNA/RNA/protein Encoding Method	Reference
Nucleic/Amino Acid composition	Basic Kmer (Kmer)	[42, 245]
	Tri-Peptide Composition (TPC)	[42, 350]
	Enhanced Amino Acid Composition (EAAC)	[87, 472]
	Accumulated Nucleotide Frequency (ANF)	[79]
Pseudo nucleic/amino acid composition	Pseudo KNC (PseudoKNC)	[269, 270]
Residue composition	Binary (binary)	[84, 88] [84, 88]
Nucleic/Amino acid Mapping	Electron-ion interaction pseudopotentials value(MappingClass_eiip_fourier)	[237, 307]
	MappingClass_integer_fourier (MappingClass_integer_fourier)	[237, 307]
Nucleic/Amino acid Distribution	Term Frequency Inverse Document Frequency (TFIDF)	[164, 336]
	Okapi-BM25	[419]
Gap based	Mono Mono K-Gap (monoMonoKGap)	[303]
	Mono Di K-Gap (monoDiKGap)	[303]
	Mono Tri K-Gap (monoTriKGap)	[303]
	Di Mono K-Gap (diMonoKGap)	[303]
	Di Di K-Gap (diDiKGap)	[303]
	Di Tri K-Gap (diTriKGap)	[303]
	Tri Mono K-Gap (triMonoKGap)	[303]
	Tri Di K-gap (triDiKGap)	[303]
	Composition of k-spaced Nucleic Acid Pairs (CKSNAP)	[87]
Graph Representation	Complex Network (complex_network)	[49, 50]
	Enhanced Complex Network (enhanced_complex_network)	[49, 50]
K-mer embeddings	Word2Vec	[94]
	FastText	[57, 137]
	DeepWalk	[328]
	Node2Vec	[157]
	Graph Auto Encoder (GAE)	[220]
	Graph Factorization (GF)	[153]
	Graph Representation (GraRep)	[63]
	Large scale Information Network Embedding (LINE)	[382]
	High-Order Proximity preserved Embedding (HOPE)	[476]
	Laplacian	[298]
Similarity based	Structural Deep Network Embedding (SDNE)	[401]
	Singular Value Decomposition (SVD)	[3]
	Position-specific trinucleotide propensity based on single-strand (PSTNPss)	[98, 173]
	Position-specific trinucleotide propensity based on double-strand (PSTNPds)	[98, 173]

of features. From filter based feature selection methods, proposed GFGPA framework contains 5 different methods: pearson correlation [163], mutual information [163], uni-variate [163], constant [236], quasi constant [236] and duplicate feature removal [163] encoders.

Uni-variate [163], constant [236] and quasi constant [236] feature selection methods eliminate duplicate features. Pearson correlation retains subset of features that are significantly correlated with the target but not with each other. The mutual information technique measures the

CHAPTER 2. A GENERIC FRAMEWORK FOR GENOMICS (DNA, RNA) AND PROTEOMICS (PROTEIN) SEQUENCE ANALYSIS

Table 2.2: Encoding methods for DNA and RNA sequences

Descriptor Category	DNA/RNA Encoding Method	Sequence Type	Reference
Nucleic acid composition	Nucleotide chemical property (NCP)	DNA/RNA	[79]
	Reverse compliment kmer (RCKmer)	DNA	[162, 313]
Pseudo nucleic acid composition	Pseudo dinucleotide composition (PseDNC)	DNA/RNA	[269, 270]
	Pseudo k-tupler composition (PseKNC)	DNA/RNA	[269, 270]
	Parallel correlation pseudo dinucleotide composition (PCPseDNC)	DNA/RNA	[269, 270]
	Parallel correlation pseudo trinucleotide composition (PCPseTNC)	DNA	[269, 270]
	Series correlation pseudo dinucleotide composition (SCPseDNC)	DNA/RNA	[269, 270]
	Series correlation pseudo trinucleotide composition (SCPseTNC)	DNA	[269, 270]
Autocorrelation and cross-covariance	Dinucleotide-based auto covariance (DAC)	DNA/RNA	[111, 161, 269]
	Dinucleotide-based cross covariance (DCC)	DNA/RNA	[111, 161, 269]
	Dinucleotide-based auto-cross covariance (DACC)	DNA/RNA	[111, 161, 269]
	Trinucleotide-based auto covariance (TAC)	DNA	[269]
	Trinucleotide-based cross covariance (TCC)	DNA	[269]
	Trinucleotide-based auto-cross covariance (TACC)	DNA	[269]
Nucleic acid Mapping	Electron-ion interaction pseudopotentials of Trinucleotide (PseEIIP)	DNA/RNA	[237, 307]
	MappingClass_binary_fourier	DNA/RNA	[49, 50]
	MappingClass_zcurve_fourier	DNA/RNA	[49, 50]
	MappingClass_real_fourier	DNA/RNA	[49, 50]
	MappingClass_complex_number	DNA/RNA	[49, 50]
	MappingClass_atomic_number	DNA/RNA	[49, 50]
Chaos theory	classical_chaos	DNA/RNA	[49, 50, 303]
	frequency_chaos	DNA/RNA	[49, 50, 303]
Z-curve	zCurve	DNA/RNA	[144] [303]
Nucleic acid Distribution	gcContent	DNA/RNA	[49, 50, 303]
	cumulativeSkew	DNA/RNA	[49, 50, 303]
	atgcRatio	DNA/RNA	[49, 50, 303]
	spectrum	DNA/RNA	[49, 50]
	orf	DNA/RNA	[49, 50]
	fickett_score	DNA/RNA	[49, 50]

reduction in uncertainty in one variable ‘X’ when the variable ‘Y’ is known. Mutual information assigns a score to each feature by utilizing information of input and output variables. Higher mutual information values imply that the target ‘Y’ has a low uncertainty given the predictor ‘X’. Univariate feature selection (ANOVA) selects an informative subset of features by making use of the Gaussian distribution to compute linear connections between the input and output variables.

Unlike filter based methods, wrapper based methods make use of predictors to select the important features. Although this approach is computationally expensive, however, it is considered better than filter-based feature selection methods in terms of performance. Proposed framework supports the most commonly used wrapper techniques based on sequential feature selection methods, which include forward feature selection, backward feature selection, recursive feature selection and exhaustive feature selection. These approaches iteratively choose the most informative subset of features from the feature space. Forward feature selection starts with the

Table 2.3: Encoding methods for protein sequences

Descriptor Category	Protein Encoding Method	Reference
Amino acid composition	Adaptive skip dipeptide composition (Protein_ASDC)	[416]
	Kmer dipeptides composition (Protein_DPC)	[42, 350]
	Dipeptide deviation from expected mean (Protein_DDE)	[350]
	PseAAC of distance-pairs and reduced alphabet (Protein_DistancePair)	[45, 272]
	Conjoint triad (Triad)	[362]
	Conjoint k-spaced Triad (KSCTriad)	[87, 472]
	Enhanced amino acid composition (EAAC)	[87, 472]
	Weighted Sparse Representation based Classification Global (WSRC_global)	[223]
	Weighted Sparse Representation based Classification Local (WSRC_local)	[223]
	Weighted Sparse Representation based Classification Local+Global (WSRC_local_global)	[223]
Grouped amino acid composition	Enhanced Grouped amino acid composition (EGAAC)	[87, 472]
	Grouped amino acid composition (GAAC)	[87, 472]
	Grouped tripeptide composition (GTPC)	[87, 472]
	Grouped dipeptide composition (Protein_GDPC)	[87, 472]
	Composition of k-spaced amino acid group pairs (CKSAAGP)	[87, 472]
Pseudo-amino acid composition	Pseudo-amino acid composition (PAAC)	[92, 233]
	Amphiphilic PAAC (APAAC)	[92, 233]
	Pseudo K-tuple reduced amino acids composition (PseKRAAC type 1 to type 16)	[483]
Residue composition	Protein_binary_6bit	[45, 406]
	Protein_binary_5bit_type_1	[45, 420]
	Protein_binary_5bit_type_2	[45, 420]
	Protein_binary_3bit_type_1	[416]
	Protein_binary_3bit_type_2	[416]
	Protein_binary_3bit_type_3	[416]
	Protein_binary_3bit_type_4	[416]
	Protein_binary_3bit_type_5	[416]
	Protein_binary_3bit_type_6	[416]
	Protein_binary_3bit_type_7	[416]
	Overlapping property features (Protein_OPF_10bit)	[416]
	Overlapping property features (Protein_OPF_7bit_type_1)	[416]
	Overlapping property features (Protein_OPF_7bit_type_2)	[416]
	Overlapping property features (Protein_OPF_7bit_type_3)	[416]
	Learn from alignments (Protein_AESNN3)	[45, 262]
	BLOSUM matrix	BLOSUM62 (BLOSUM62)
Z-Scale index	ZSCALE (ZSCALE)	[83]
Physicochemical property	AAINDEX (AAINDEX)	[394]
	Composition (CTDC)	[59, 60, 117, 118, 165]
	Transition (CTDT)	[59, 60, 117, 118, 165]
	Distribution (CTDD)	[59, 60, 117, 118, 165]
Quasi-sequence-order	Sequence-order-coupling number (SOCNumber)	[91, 93, 356]
	Quasi-sequence-order descriptors (QSOrder)	[91, 93, 356]
Autocorrelation	Moran	[134, 264]
	Geary	[367]
	NMBroto	[184]
	auto_covariance	[111, 161, 269]
	auto_cross_covariance	[111, 161, 269]
	bi_auto_covariance	[111, 161, 269]

feature that performs best against the target. Then we select a second feature in such a manner that when it is paired with the first, both yields the best results. This procedure is repeated until the predetermined criterion is satisfied [397].

Backward feature selection, also known as backward elimination, operates in the exact

CHAPTER 2. A GENERIC FRAMEWORK FOR GENOMICS (DNA, RNA) AND PROTEOMICS (PROTEIN) SEQUENCE ANALYSIS

Table 2.4: Feature selection and dimensionality reduction methods in the proposed GFGPA framework

Method	Algorithm	Reference
Wrapper Based Feature Selection Approaches	Forward Feature Selection	[221]
	Backward Feature selection	[221]
	Recursive feature selection	[221]
	Exhaustive Feature Selection	[221]
Embedded Feature Selection Approaches	Lasso	[236]
	Redige	[236]
Tree based Feature Selection Approaches	Random Forest based Feature Importance	[295]
	Decision Tree based Feature Importance	[295]
	Gradient Boost based Feature Importance	[295]
	Xtream Gradient Boost based Feature Importance	[295]
Filter Based Feature Selection Approaches	Mutual Information	[163, 163]
	Univariate Feature selection	[163]
	Pearson	[163]
	constant Feature	[236]
	Quasi constant Features	[236]
	Duplicate Features	[163]
Dimensionality Reduction	K-means	[167]
	T-SNE	[395]
	Principal Component Analysis (PCA)	[187]
	Kernel PCA	[358]
	Locally Linear Embedding	[352]
	Singular Value Decomposition (SVD)	[166]
	Non-Negative Matrix Factorization (NMF)	[100]
	Independent Component Analysis (ICA)	[247]
	Multi-Dimensional Scaling (MDS)	[207]
	Factor Analysis	[346]
	Feature Agglomeration	[359]
	Gaussian Random Projection	[101]
Sparse Random Projection	[254]	
Auto Encoder	Auto Encoder	[411]

opposite way as forward feature selection. This technique starts with all the input features and builds a model around them. At each iteration, it removes least performing feature from the feature set. Recursive Feature Elimination (RFE) employs a greedy search method to find the best feature subset. It builds models iteratively, identifying which features perform best or worse

in each iteration. It continues to develop models based on the features that are left in the feature space until all of them have been examined. The features are then graded according to how likely they are to be removed.

In embedded feature selection methods, the feature selection process is included in the learning or model building phase, also known as the training phase. These approaches require less time to train and are less prone to model over-fitting as compared to wrapper feature selection methods. Proposed framework supports Lasso and Ridge embedded paradigms. Lasso paradigm eliminates the feature to alleviate over-fitting in a linear classifier, whereas Ridge mainly reduces the overall impact of features which are not useful in making accurate predictions of target class labels. Besides these two paradigms, embedded feature selection paradigms based on Random Forest, Decision Tree, Gradient Boost and Extreme Gradient Boost are provided where the importance of the features is computed in terms of the purity of subset of dataset on which individual tree operates.

2.1.2.2 Dimensionality Reduction

The aim of dimensionality reduction procedures is to transform original p -dimensional feature space into the lower k -dimensional feature subspace. Proposed GFGPA framework facilitates 14 dimensionality reduction algorithms shown in Table ??, which can be segregated into two different categories: linear and nonlinear.

Linear dimensionality reduction methods including non-negative Matrix Factorization (NMF), Independent Component Analysis (ICA), Principal Component Analysis (PCA), Truncated SVD, Factor Analysis (FA) and linear discriminative analysis (LDA) transform high-dimensional feature space into a low-dimensional feature space as a linear combination of the original variables. The low-dimensional feature space retains the intrinsic structure of statistical sequence vectors such that the least numbers of parameters manage to capture the essential sequence features. NMF decomposes statistical feature space into two non-negative matrices known as NMF matrix and coefficients matrix and original statistical feature space is transformed into reduced feature space through additive combination of vectors present in underlay matrix. ICA also generates reduced feature space by separating original statistical feature space into additive components. PCA converts original statistical vectors into n principal components that represent the most relevant information. Similarly, Truncated SVD factorizes the original statistical vectors into number of columns equal to truncation to retain only a few largest singular values, Factor Analysis finds factor to describe the covariance of correlated observed variables. LDA focuses on low dimensional feature space with maximum separability between the groups.

Contrary to linear methods, nonlinear methods are applied to original statistical vectors that contain a nonlinear relationship. These methods preserve the global as well as local features of high-dimensional feature space in low-dimensional feature space. Nonlinear dimensionality reduction methods including K-means, t-SNE, Kernel PCA, Isometric Mapping (Isomap) and

multi-dimensional scaling (MDS) preserve the global features of original statistical vectors of sequences. K-means computes the cluster centers and sets the number of clusters equal to target dimensions of statistical feature space. The new statistical feature space is generated in which new features are actually the distances of each point with respect to each cluster center. The t-SNE algorithm constructs a probability distribution on the feature pairs in the higher dimensions in such a manner that similar features are assigned higher probabilities and dissimilar features are assigned lower probabilities. Kernel PCA projects the nonlinear inseparable statistical vectors onto a higher dimensional feature space where it becomes linearly separable. Isomap generates the neighborhood networks and preserves the geodesic distance in low-dimensional feature space. Similar to Isomap, MDS measures the similarities and dissimilarities between the observed variables.

Instead of preserving global features, some methods try to preserve only geometrical properties of local features of nonlinear original statistical vectors such that locally linear embedding (LLE), Hessian LLE and Laplacian eigenmap. These methods preserve the local features in low-dimension statistical vectors assuming that only the local distances are reliable in high-dimensional statistical vectors. LLE reduces the original feature space to lower embedding while preserving the embedding of original sequences. Laplacian eigenmap maps the embedding corresponding to nearest neighbor and represents the graph with its Laplacian matrix. Hessian LLE is an extension of LLE that first minimizes the curviness of high dimensional original statistical vectors and then transfers to low dimensional feature space to make low dimension feature space locally isometric. Feature Agglomeration is another nonlinear dimensionality reduction approach which groups various components that behave similarly using hierarchical clustering. This behavior can be achieved by clustering in the feature direction or clustering transposed feature space. Gaussian Random Projection reduces the dimensions of high-dimensional feature space by projecting the original input's dimensional space onto a randomly generated matrix. Sparse Random Projection transforms feature space to sparse random matrices. These matrices are the best alternative to the dense Gaussian random projection matrices as they generate similar quality feature space in less memory and allow faster computation on the new feature space.

AutoEncoder is another very efficient dimensionality reduction approach based on artificial neural network. It is based on encoder-decoder paradigm where encoder compresses the original feature space into lower dimensions using bottleneck layers and decoder produces the original feature space from compressed representation. By reducing the reconstruction loss, an effective compressed representation of statistical feature space is learned. A brief description about autoencoder base dimensionality reduction is provided in chapter 6.

2.1.3 Predictor Construction Module

This section summarizes machine and deep learning based predictors that proposed framework facilitates.

2.1.3.1 Traditional Machine Learning Classifiers and Regressors

Primarily, biomedical sequence classification approaches can be categorized into two main types: multi-class and multi-label classification. To perform binary-class or multi-class classification, proposed GFGPA frameworks facilitate 12 widely used machine learning classifiers.

Naive Bayes (NB) [413] is a probabilistic/Bayesian generative model that assigns class labels based on subsequent conditional probabilities against a certain hypothesis (presence/absence). As it is based on the Bayes theorem, it assumes independence among the features from each other, whereas the outcome of NB is the class with maximum probability. Naive Bayes by default considers Gaussian distribution, however, multinomial Naive Bayes considers multinomial distribution for the features which is more appropriate to analyze the count of features in corpus sequences. Logistic regression models the associations of features with target classes and predicts the probabilities of target classes using sigmoid function. Gaussian Process classifier can be considered a generalization of Gaussian probability distribution which predicts target class on the basis of probability values of features.

The support vector machine (SVM) is a discriminative classifier which finds the appropriate hyperplanes that isolate two classes by maximizing the margin. For nonlinear problems, it uses kernel trick that transforms feature space of nonlinear sequence samples to linearly separable feature space [58]. K-Nearest Neighbors (KNN) is the simplest distance based machine learning classifier that assumes that similar sequence samples lie in close proximity to each other [227]. For a sequence sample and k number of neighbors, distance is computed using different metrics, e.g., Euclidean distance, Hamming distance and based on the distance, a class is assigned to a sequence sample.

Decision Tree (DT) [206] classifier transforms sequence samples into tree based structure. Initially, a root node is selected on the basis of the feature having a lower Gini impurity or the maximum information gain [206, 380]. Then, a split of the sequence samples is performed based on the different categories present within that specific feature. This process is repeated until the nodes reach a stage where they only contain the sequence samples that belong to only one class. In the end, a decision is made by iterating over the nodes of the tree with specific conditions until it does not reach the terminal node. A random forest (RF) classifier incorporates the decision tree [140] as a base model built on bootstrap aggregation (bagging) and an averaged or voted decision is constructed from a forest of decision trees. Gradient boosting classifier mainly combines multiple weak classifiers to construct a strong classifier. Bagging classifier is a meta-classifier which fits the base classifiers on randomly selected subsets of sequences and afterward aggregates their predictions to formulate a final prediction. Extra trees classifier (ETC)

[149] is the extension of the ensembling method RF. In ETC, the concept is to introduce more randomness when creating subsets of sequence samples; ETC abandons bootstrapping (Bagging) and enhances the model's accuracy. Relatively, Adaptive Boosting (AdaBoost) [140] classifier is established on the idea of training multiple weak classifiers to create a strong classifier. Unlike other classifiers, AdaBoost uses all training sequence samples to train a classifier. The dataset is updated by assigning higher weights to the misclassified samples. A new classifier is trained on the updated dataset and the process is replicated N times; as a result, a strong predictor is constituted.

To support regression tasks, proposed framework facilitates 11 machine learning regressors. Working of most machine learning regressors is similar to their classifiers counterparts. Few regressors (Elastic-Net, Stochastic Gradient Descent) apply penalties and regularization strategies to obtain optimal model weights.

The multi-label classification approaches can be categorized into two main categories namely problem transformation [53] and algorithm adaptation [461]. Problem transformation based classification is a two-stage process [368, 390]. In first stage, multi-label problem is converted into a binary or multi-class problem [368, 387, 390]. Then in second stage, traditional binary or multi-class classifiers are used to perform final classification [368]. The primary objective of data transformation approaches is to transform the multi-label problem into a binary or multi-class problem without losing the label-to-label and sample-to-label relations [368]. Proposed GFGPA facilitates 3 different most widely used problem transformation methods, namely Label Powerset [391], Binary Relevance [53] and Classifier Chains [340] which have shown great effectiveness in different Natural Language Processing and Bioinformatics tasks.

With the passage of time, researchers have modified several binary or multi-class classifiers such as MLkNN [461] and BRkNN [370], both are extended from kNN classifier. Similarly, RF-Boost [5], MP-Boost [125] and MH-boost [355] are extended forms of AdaBoost [140], MLTSVM [81] is a modified form of SVM classifier [81], MLARAM [38] is an extension of Adaptive Resonance Associative Map neural-fuzzy networks. Also, several tree based, nearest neighbour based probabilistic distribution based machine learning classifiers are modified to perform multi-label classification. Proposed framework facilitates 15 most widely used algorithm adaptation approaches to support different multi-label sequence analysis tasks.

2.1.3.2 Deep Learning based Predictors

In the marathon of developing robust and precise deep learning based predictors for diverse Genomics and Proteomics sequence analysis tasks, we are witnessing the explosion of deep learning approaches , core architectures of which are mainly formed by deep feed forward neural networks [256], convolutional neural networks [256], recurrent neural networks [299] and hybrid networks that make use of both CNN and RNN layers.

Proposed GFGPA framework facilitates different deep learning based predictors for multi-

2.1. FUNCTIONAL SCOPE AND DESCRIPTION OF PROPOSED GENERIC FRAMEWORK

Table 2.5: Classifiers and regressors available in proposed GFGPA framework

Predictor category	Algorithm	Classifier	Regressor	Reference	
Machine Learning Classifiers/Regressors (Multi-class/Binary)	Random Forest	✓	✓	[140]	
	Support vector	✓	✓	[81]	
	Naïve Bayes	✓	-	[413]	
	Logistic Regression	✓	✓	[422]	
	K Neighbors	✓	✓	[227]	
	Gaussian Process	✓	-	[279]	
	Gradient Boosting	✓	✓	[279]	
	Extra Trees	✓	-	[149]	
	Decision Tree	✓	✓	[206]	
	Bagging	✓	-	[344]	
	AdaBoost	✓	-	[140]	
	Multinomial NB	✓	-	[156]	
	Elastic Net	-	✓	[156]	
	Linear Regression	-	✓	[156]	
	Huber	-	✓	[156]	
	Stochastic Gradient Descent (SGD)	-	✓	[344]	
	Extreme Gradient Boosting (XGB)	-	✓	[344]	
	Data Transformation Approaches (Multi-Label Classifiers)	Binary Relevance			[53, 53]
		Classifier Chain			[53, 340]
Label Powerset				[53, 391]	
Algorithm Adaptation Approaches (Multi-Label Classifiers)	Random Forest Classifier			[461]	
	Decision Tree Classifier			[461]	
	Extra Tree Classifier			[461]	
	K Neighbors Classifier			[461]	
	Multi-Layer perceptron (MLP) Classifier			[461]	
	Radius Neighbors Classifier			[461]	
	Logistic Regression			[461]	
	RidgeClassifierCV			[461]	
	BRkNNa Classifier			[370]	
	GridSearchCV			[58] [338]	
	GridSearchCV_MLkNN			[461]	
	Linear SVC			[58]	
	Multilabel k Nearest Neighbours (MLkNN)			[461]	
	MLARAM			[38]	
	MLTSVM			[81]	
Deep Learning Algorithms	Multi-Layer Perceptron (MLP)	✓	✓	[381, 461]	
	Dense-Net			[19]	
	Res-NET			[28]	
	CNN			[256]	
	LSTM			[299]	
	LSTM-CNN			[200]	
	CNN with attention			[200] [28]	
	LSTM with attention			[200] [28]	
	LSTM-CNN with attention			[200] [28]	

class and multi-label Genomics and Proteomics sequence classification which can be classified as modified deep learning predictors and novel deep learning predictors. The focus of modified deep learning predictors is to evaluate the efficacy of architectures that are built by inspiring from

those architectures which have shown great performance in diverse tasks of various domains. In this regard, the GFGPA framework facilitates deep predictors based on DenseNet and ResNet architectures, which utilize different strategies to more effectively propagate the error signal to earlier layers. Final classification layer provides implicit supervision to earlier layers which helps the model to converge to true parameters for accurate prediction.

Furthermore, the GFGPA framework facilitates novel deep learning predictors solely based on multi-layer perceptron, different number of Long-Short Term Memory (LSTM) layers, LSTM and attention layers, convolutional layers, convolutional and attention layers, LSTM and convolutional layers, as well as LSTM, convolutional and attention layers.

With an aim to optimize the decision making of deep learning based predictors and accelerate training, GFGPA framework facilitates multiple neural strategies such as combination of different pooling methods to retain comprehensive discriminative features, normalization to prevent exploding and vanishing gradient issues, different kinds of dropout to avoid over-fitting, learning rate decay to rapidly reduce the prediction error and converge the model parameters to true parameters. A brief description of proposed deep learning predictors is provided in chapters 4, 5, 6, 7 and 8.

2.1.4 Performance Evaluation Module

To perform a large-scale genomics (DNA, RNA) and Proteomics (protein) sequence analysis for multiple species, the proposed GFGPA framework is capable of performing various types of tasks that fall under the hood of classification and regression. However, unlike Natural Language Processing, biomedical domain tasks are extremely sensitive as false positive or false negative predictions can cost millions of lives. In order to evaluate the integrity, generalizability and applicability of any application developed through the proposed GFGPA framework, GFGPA framework facilitates comprehensive performance evaluation metrics for different classification and regression tasks, an effective visualization of which helps to present the key findings in the most efficient manner. A brief description of evaluation measures facilitated by the proposed GFGPA framework is provided in following subsections.

2.1.4.1 Multi-Class Classification Evaluation Measures

In multi-class classification tasks, each corpus sequence belongs to only one particular class label at a time. Hence, the predicted class label will fall into one of the four categories true positive, true negative, false positive and false negative that are shown in Table 2.6. True Positive illustrates the count of correctly predicted positive class values, e.g., if both the actual and predicted class labels are yes then it will be considered as true prediction of positive class label. Similarly, True Negative is accurate prediction of negative class labels. False Positive denotes the count of wrongly predicted class labels, i.e., when actual class is 'no' but model predicts 'yes'. Likewise, False Negative is wrong prediction of 'no' class when actual class is 'yes'.

Table 2.6: Confusion matrix for binary classification

Actual Class	Predicted Class		
		Class=yes	Class=no
	Class=yes	True Positive	False Negative
Class=no	False Positive	True Negative	

Accuracy (ACC) [185] determines the proportion of correct predictions with respect to total predictions. Precision (PR) measures the percentage of the complete true positive matches from all true positive matches. Specificity (SP) [185] also known as true negative rate (TNR) measures the correct predictions of negative class sequences. It is the ratio between true negative class predictions and overall predictions of negative class. Similarly, Recall/Sensitivity [185] calculates performance scores by taking into account correct predictions of positive class sequences. MCC [185] measures the correlation of the true classes with the predicted classes by taking all four true positives, false positives, true negatives and false negatives into account. F1-score measures a harmonic mean of precision and sensitivity. Mathematical expressions of the aforementioned evaluation measures are given as follows:

$$f(x) = \begin{cases} \text{Accuracy (ACC)} = (T_P + T_N)/(T_P + T_N + F_P + F_N) \\ \text{Precision (PR)} = T_P/(T_P + F_P) \\ \text{specificity (SP)} = T_N/(T_N + F_P) \\ \text{Recall/Sensitivity (SN)} = T_P/(T_P + F_N) \\ \text{False Positive Rate (FPR)} = F_P/(T_N + F_P) \\ \text{MCC} = T_P \times T_N - F_P \times F_N/Q \\ Q = \sqrt{(T_P + F_N)(T_P + F_P)(T_N + F_P)(T_N + F_N)} \\ \text{F1-score} = 2 * PR * SN/(PR + SN) \end{cases} \quad (2.1)$$

In above mathematical expressions of different evaluation measures, T_P and T_N denote the true predictions related to the positive and negative classes. While, F_P and F_N indicate the false predictions related to the positive and negative class, respectively.

Besides these, two probability curve based evaluation measures are used in proposed framework. Area under receiver operating characteristics (AUROC) [185] measures degree of separability of the model by analyzing both true positive rate (TPR) and false positive rate (FPR) at different thresholds. Area under precision recall curve (AUPRC) measures model ability to handle imbalance datasets by analyzing precision and sensitivity at different thresholds where the goal is to have higher precision and sensitivity.

2.1.4.2 Multi-Label Classification Evaluation Measures

Performance evaluation of multi-label predictors is difficult as compared to performance evaluation of binary or multi-class predictors [423]. In multi-label classification, a sequence can have two or more labels at the same time, so there is a possibility that model predicts only one label correctly, both labels correctly or both labels incorrectly [423]. Due to partial corrections, it is hard to quantify the performance of a multi-label predictor [423]. Over the time, researchers have proposed different evaluation measures to compute the performance of multi-label classifiers. Each evaluation measure has its own pros and cons. Most of the existing evaluation measures fall under the two different categories namely correct performance prediction computers and loss calculators. Both types of measures are briefly described below.

Accuracy [423] assesses the performance of classifier by computing the ratio between actual and predicted labels. Precision computes performance by closely monitoring actual true labels from the set of labels that classifier predicted as true. Recall measures how many labels are correctly predicted from actual labels. F1-score is a harmonic mean between precision and recall. The higher the value of accuracy, precision, recall and f1 the better will be the performance of a classifier. Average Precision evaluates the performance of predictor by summarizing the precision recall curve in a single value representing the average of all precisions.

Hamming loss measures how many labels are wrongly predicted and how many labels remain unpredicted. One error [423] monitors the performance of a classifier by computing number of sequences in which top-ranked labels of the classifier are different from the set of actual labels associated with those sequences. Coverage measures how many steps, on average, are needed to move down the ranked label list to cover all actual labels of a sequence. Ranking Loss measures how many times the wrong label is ranked above the actual label. Smaller values of these ranking metrics represent better performance of a classifier.

$$f(x) = \left\{ \begin{array}{l} \text{Recall} = \frac{1}{M} \sum_{i=1}^M \frac{|A_i \wedge P_i|}{|A_i|} \\ \text{Accuracy} = \frac{1}{M} \sum_{i=1}^M \frac{|A_i \wedge P_i|}{|A_i \vee P_i|} \\ \text{F1-Score} = \frac{1}{M} \sum_{i=1}^M \frac{2 * |Pre(n_i) * Rec(n_i)|}{|Pre(n_i) + Rec(n_i)|} \\ \text{Precision} = \frac{1}{M} \sum_{i=1}^M \frac{|A_i \wedge P_i|}{|P_i|} \\ \text{RankingLoss} = \frac{1}{M} * \sum_{o=1}^{M-1} \frac{1}{\|A_i\|_o * (n_{labels} - \|A_i\|_o)} \\ \text{HammingLoss} = \frac{1}{ML} \sum_{i=1}^M \sum_{j=1}^L [I(A_i^j \neq P_i^j)] \\ \text{AveragePrecision} = \frac{1}{|M|} \sum_{i=1}^{|M|} \sum_{y \in Y_i} \frac{|\{y' | f_{rank}(x_i, y') \leq f_{rank}(x_i, y), y' \in Y_i\}|}{f_{rank}(x_i, y)} \\ \text{Coverage} = \frac{1}{M} * \sum_{o=1}^{M-1} \max_{j: a_{ij}=1} \text{rank}_{ij} \\ \text{OneError} = \frac{1}{M} \sum_{i=1}^M [|\text{argmax} F(n_i) \notin A_i^+|] \\ \text{rank}_{ij} = |\{l : \hat{f}_{il} \geq \hat{f}_{ij}\}| \end{array} \right. \quad (2.2)$$

In these equations 2.2, M denotes total number of sequences, n_i represents i^{th} sequence from

m sequences, A_i represents actual class label and P_i denotes predicted label of n_i sequence, L represents length of sequence, j^{th} represents the class index, \vee represents logical OR operator and \wedge represents logical AND operator.

In addition, proposed GFGPA framework supports AUROC and AUPRC evaluation measures to quantify the performance of a multi-label classifier.

2.1.4.3 Regression Evaluation Measures

Proposed GFGPA framework contains 4 different evaluation measures (mean bias error (MBE), mean absolute error (MAE), root mean square error (RMSE) and r-squared score (R^2)) [52] to carryout performance analysis of any regression related model. To provide an intuitive understanding for readers, evaluation metrics along with mathematical expressions are briefly described below.

$$f(x) = \begin{cases} \text{Mean Absolute Error (MAE)} = I/N \sum_{i=1}^N |(y_{p,i}) - (y_{a,i})| \\ \text{Mean Bias Error (MBE)} = I/N \sum_{i=1}^N (y_{p,i}) - (y_{a,i}) \\ \text{Root Mean Squared Error (RMSE)} = \sqrt{I/N \sum_{i=1}^N ((y_{p,i}) - (y_{a,i}))^2} \\ \text{R}^2 \text{ score (R}^2\text{)} = 1 - \frac{\sum_{i=1}^N (y_{p,i} - y_{a,i})^2}{\sum_{i=1}^N (y_{a,i} - \text{avg}(y_a))^2} \end{cases} \quad (2.3)$$

Where N expresses the number of sequences, y_p is the predicted value through machine learning or deep learning regressors, y_a is the measured value through pyranometer of sequence "i" and $\text{avg}(y_a)$ is average of all pyranometer calculated values.

2.2 A Look Back & into Future: Functional Scope of the Existing and Proposed Generic Framework

To investigate the potential of proposed GFGPA framework, Table 2.7 performs a comprehensive functional scope comparison of the proposed GFGPA framework with five most recent generic sequence analysis frameworks under the hood of 17 key functionalities and criteria.

With an aim to generate statistical representations of DNA, RNA and protein sequences by extracting distribution of nucleic and amino acids, as compared to the most recent frameworks iLeanPlus [85] and MathFeature [51], proposed framework provides 71 more sequence encoding methods. Considering, that not all features contribute equally to make accurate genetic sequence analysis, another distinguishing functionality of the GFGPA framework is that it provides four times more algorithms to perform feature engineering. It enables the researchers to deeply explore 30 different algorithms related to two different paradigms namely feature selection and dimensionality reduction to optimize the feature space for different sequence analysis tasks.

CHAPTER 2. A GENERIC FRAMEWORK FOR GENOMICS (DNA, RNA) AND PROTEOMICS (PROTEIN) SEQUENCE ANALYSIS

Table 2.7: A comprehensive functional scope analysis of proposed GFGPA and existing frameworks

Category	iFeature [86]	iLearn [87]	BioSeq-Analysis2.0 [45]	MathFeature [51]	iLearnPlus iLeanPlus [85]	Proposed GFGPA Framework
Number of feature sets for DNA sequence	0	26	36	38	46	64
Number of feature sets for RNA sequence	0	18	27	38	35	58
Number of feature sets for protein sequence	53	53	53	25	66	99
Number of clustering algorithms	5	6	0	-	10	-
Number of feature selection algorithms	4	5	2	-	5	16
Number of feature normalization algorithms	0	2	0	-	2	2
Number of dimension reduction algorithms	3	3	0	-	3	14
Number of machine-learning Classifiers	0	5	5	4	21	23
Number of machine-learning Regressors	-	-	-	-	-	12
Number of machine-learning Multilabel Classifiers	-	-	-	-	-	26
Number of cross-validation methods	0	2	3	2	2	4
Number of evaluation metrics	0	8	5	4	8	14
Sequence Interaction Analysis	-	-	-	-	-	Yes
Can build machine-learning pipeline	No	Yes	No	Yes	Yes	Yes
Can perform evaluation of the feature sets/machine learning models in a batch manner	No	Yes	No	No	Yes	Yes
Case Studies and Species Evaluation	Limited	Limited	Limited	Limited	Limited	Comprehensive
Results Visualization	Limited	Limited	Limited	Limited	Limited	Comprehensive

Contrary to existing frameworks that only support two sequence analysis tasks namely clustering and binary/multi-class classification, proposed GFGPA framework supports four different sequence analysis tasks including regression, clustering, multi-class classification and multi-label classification. Furthermore, it allows researchers to assess the true performance potential of pipelines by facilitating 4 different cross-validations methods and 16 distinct evaluation metrics which are almost twice the number of evaluation methods provided by existing frameworks. The GFGPA is the only framework that provides 5 web applications based on pre-trained AI models related to multiple species and facilitates predictions on the go.

Part I

Genomics Sequence Analysis

DNA MODIFICATION PREDICTION

DNA modification is a core feature in the eukaryotic and prokaryotic genomes for the control of gene expression, chromosome inactivation, replication and cell differentiation [301]. Various DNA modifications occur due to the addition of the methyl group into two different nucleotide bases, Cytosine (C) and Adenine (A). The addition of methyl groups at different positions in cytosine produces different modifications such as N4-methylcytosine (4mc), 5-methylcytosine (5mc), 3-methylcytosine (3mc), 5-hydroxymethylcytosine (5hmc), 5-formylcytosine (5fc) and 5-carboxylcytosine (5caC) [277]. Similarly, the addition of methyl in Adenine produces N6-methyladenine (6mA) modification. Among these different modifications, 6ma, 4mc and 5hmc are considered important modifications due to their critical roles in the mammalian and prokaryotic genome [321]. The 4mc and 6ma modifications are common in bacterial genomes [73, 255, 321], while the 5hmc modification is prevalent in the eukaryotic genome. In the prokaryotic genome, 4mc and 6ma modifications regulate the correction of DNA replication errors and gene expression as well as defend the DNA from the attacks of foreign DNA of viruses and bacteria [73, 212].

The modified state of DNA (addition of methyl in cytosine or adenine) affects transcription because of its hypomethylation or hypermethylation state. Hypomethylation is an unmethylated state of DNA where the DNA is accessible and available for transcription. Whereas, in the state of methylated DNA called hypermethylation, the DNA becomes compact and inaccessible; thus, the transcription stops [301]. During the process of demethylation called hypomethylation, 5mc breakdowns and forms a critical demethylation complex (5hmc) which affects gene expression regulation. It is believed that methylated DNA (DNAm) is the core regulatory factor for aging, which highlights the preeminence of modification and demodification process. Furthermore, DNA

⁰This chapter is an adapted version of the work presented in Asim et al. "DNA-MP: A Generic DNA Modifications Predictor for Multiple Species based on Novel Sequence Encoding Method", under review in briefings in Bio-informatics 2022

modification plays a significant role in various biological processes, so it is important to explore DNA modifications in detail.

The effects of differential DNA methylation on oncogenes result in different types of cancer, such as breast, squamous cell lung cancer and glioblastoma [211]. DNA modification such as methylation plays a critical role in inducing autoimmune diseases and neurological disorders like systemic lupus erythematosus, multiple sclerosis, autism spectrum disorder (ASD) and schizophrenia [211]. Irregular expression of 6ma prompts severe consequences in prokaryotes, i.e., sensitivity to ultraviolet radiations and mitomycin C treatment in *E. coli*.

3.1 Related Work

Following the success of Artificial Intelligence (AI) in various application areas, i.e., Genomics and Proteomics [417, 481], the development of robust AI based approaches for DNA modifications prediction is an active area of research [273, 287]. To date, a number of AI-based approaches have been developed [415, 455, 482] with an aim to more precisely predict three different types of DNA modifications. Fundamentally, AI-based predictors work in a two-stage process, i.e., feature representation and classification. The first stage involves the generation of statistical representations of DNA sequences and the second stage makes use of statistical representations to extract comprehensive discriminative features using machine or deep learning classifiers.

In order to more precisely distinguish modification sites, for 4mc modification prediction, 24 different classifiers have been proposed, out of which 11 are based on machine learning (ML) [8, 80, 130, 169, 173, 257, 283, 287, 414, 431, 470] and 13 are based on deep learning (DL) [1, 6, 128, 218, 273, 342, 378, 400, 415, 435, 455, 457, 482]. Similarly, for 6ma modification prediction, out of 15 different classifiers, 8 are based on DL [2, 72, 192, 335, 377, 392, 451, 458] and 7 are based on ML [37, 61, 168, 222, 274, 333, 404]. Two generic approaches iDNA-Ms [282] and iDNA-MT [435] which are capable of predicting multiple types of DNA modifications, are based on DL [435] and ML classifiers [282], respectively. To summarize, prior mentioned deep learning classifiers make use of different neural architectures i.e., Multi-Layer Perceptron models (MLPs) [460], convolutional neural networks (CNNs) [377], recurrent neural networks (RNNs) [435], hybrid neural network (CNNs+RNNs) [72] and language models like transformers [457]. Whereas, ML based approaches make use of traditional ML classifiers i.e., random forest [8], decision trees (DT) [8], adaboost [274], random forest (RF) [169], extra tree classifier (EXT) [274], gradient boosting (GB) [404], naive bayes (NB) [130, 257], logistic regression (LR) [130] and support vector machine (SVM) [457] classifier.

At the first stage, for the conversion of DNA sequences into statistical vectors, prior mentioned deep learning models make use of one hot encoding [435], Word2vec embeddings, contextual binary encoding (C-BE) [392], contextual nucleotide chemical property and nucleotide frequency (NCPNF) based encoding [392], Bert representations [242], KNN based similarity scores [37]

and trinucleotide composition (TNC) [335] based encoding methods. In literature, diverse types of sequence encoding methods that have been utilized in combination with traditional machine learning classifiers can be categorized into 2 mathematical and physicochemical categories. Mathematical encoders include k-mer [6], adjacency dependent information [431], nucleotide positional specificity (NPS) [61], motif score matrix [61], transition probabilities [333], position-specific trinucleotide propensity (PSTNP) [173] and k-mer [414]. Physicochemical properties based encoders [6, 131, 173] include nucleotide chemical property (NCP) [6, 130], mono nucleotide binary encoding (MBE) [6, 282], dinucleotide binary encoding (DBE) [6], k-nucleotide composition (KNC) [130], electron-ion interaction pseudopotentials (EIIP) [6], pseudo dinucleotide composition (PseDNC) [415], ring function hydrogen and chemical properties [37], dinucleotide composition and dinucleotide based properties (F twist, slide, energy, enthalpy) [222], K-tuple nucleotide component [130], nucleotide property and frequency (NPF) [282] and position-specific nucleotide composition (PSNP) [404]. Furthermore, to reap the benefits of different encoding methods, few researchers have concatenated the statistical representations of multiple sequence encoders [61, 273, 287, 470].

Using a variety of standalone as well as combination of sequence encoders and machine or deep learning classifiers, a plethora of DNA modifications predictors have been proposed which can be broadly classified into two categories on the basis of their ability to predict one or multiple DNA modifications, 1) type-specific modification predictors, 2) generic modifications predictors. A critical analysis indicates that both type-specific and generic modification predictors have limited predictive performance and generalizability across benchmark datasets of multiple species. This is mainly due to the use of ineffective sequence encoding methods that lack to capture position specific distributional information of nucleotides, which is essential to most effectively characterize constant as well discriminative regions of nucleotides within DNA sequences to accurately predict different DNA modifications.

With an aim to develop an efficient large scale DNA modifications prediction landscape for multiple species, the contributions of this chapter are manifold: **(I)** It presents a novel statistical representation generation approach that makes use of position specific occurrence based on modification and non-modification class densities normalized difference to compute the score of nucleotides (POCD-ND). Unlike existing nucleotide composition, frequency and physicochemical properties based sequence encoding methods, POCD-ND captures distributional information of unique higher order nucleotides called k-mers with respect to all possible unique positions inside DNA sequences. POCD-ND method helps to encode position aware comprehensive discriminative patterns of k-mers which are extremely useful for the detection of DNA modifications **(II)** To validate the efficacy of proposed POCD-ND encoding method, we compare statistical representations generated through POCD-ND encoding method with statistical representations generated through 32 most widely used existing encoding methods. In this intrinsic evaluation, our aim is to analyze which encoder is capable of generating highly disjoint clusters for positive and

negative modification sites classes **(III)** Over 17 datasets of 12 different species, it performs a comprehensive extrinsic evaluation of proposed encoder and 32 existing encoders using 10 different machine learning classifiers for the detection of three different DNA modifications. **(IV)** To objectively evaluate the predictability and generalizability of the proposed generic DNA-MP predictor, it performs a detailed performance comparison with existing type-specific and generic modifications predictors using 17 benchmark datasets under the hood of 5-fold cross-validation and independent test sets **(V)** To enable biomedical researchers and practitioners to predict different DNA modifications on the go, we have developed a user-friendly and interactive web server, freely available at https://dna_modification_predictor.opendfki.de/.

3.2 Materials and Methods

This section illustrates details of proposed DNA sequence encoding method and DNA modification benchmark datasets. A comprehensive detail of used classifiers and evaluation measures is provided in Chapter 2.

3.2.1 Proposed DNA Sequence Encoder

Machine learning classifiers cannot directly process raw DNA sequences due to their inherent dependency on numerical values. DNA sequences are comprised of only four basic nucleotides, positions-specific distributions of which are very similar across the sequences of the same class and different across the sequences of distinct classes. To date, several encoding methods have been proposed where the aim of each newly developed method has been to capture position aware discriminative distribution of nucleotides. However, these encoders still fail to capture comprehensive position aware discriminative patterns of nucleotides in DNA sequences. To generate a more comprehensive statistical representation of DNA sequences that can capture position aware discriminative distributional information of nucleotides, we present a novel DNA sequence encoder namely position aware k-mer occur-

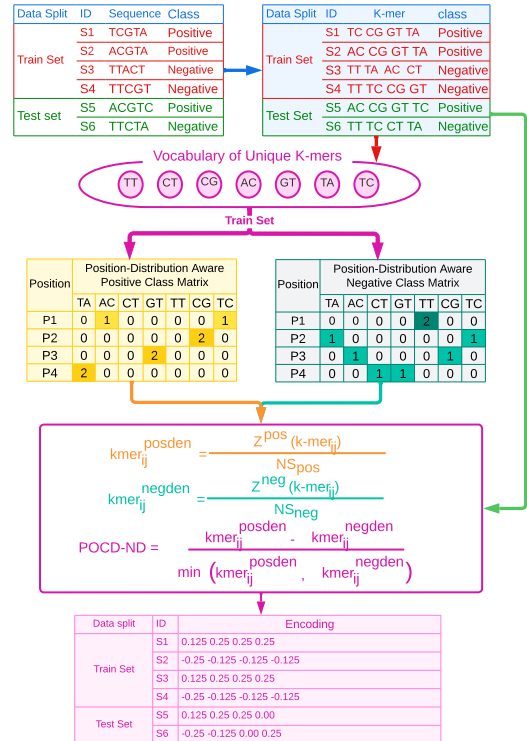


Figure 3.1: Working paradigm of proposed POCD-ND encoding method

rence based on modification and non-modification class densities normalized difference (POCD-ND).

To generate statistical representations of DNA sequences, the first step is to generate k-mers by sliding a fixed-size window with a particular stride size. In this process, DNA sequences are segregated into subsequences where each subsequence called k-mer represents a group of nucleotides. In the generated subsequences, size of k-mer or subsequence depends upon the size of a window which is slid to generate them.

Suppose we have a corpus $C = S_1, S_2, \dots, S_M$, where X number of sequences belong to training set and Y number of sequences belong to test set. In the corpus C , each S_i represents a DNA sequence that is comprised of four repeated letters including A, C, G, T . After generating k-mers of corpus sequences, each sequence S_i can be represented as $S_i = k-mer_1, k-mer_2, \dots, k-mer_k$. In each sequence, k-mers positions can be represented as $P_i = P_1, P_2, \dots, P_n$. The proposed encoder computes the vocabulary $V = v_1, v_2, \dots, v_k$ which contains unique k-mers of the corpus sequences. The size of the vocabulary depends on the size of k-mer and can be computed using 4^k where k represents the size of k-mer. For example, in the case of 1-mer, vocabulary size will be $4^1 = 4$, for 2-mers, vocabulary size will be $4^2 = 16$ and so on. With the increase in size of k-mers, the size of vocabulary also increases.

With an aim to compute position specific k-mers occurrence frequencies in modification and non-modification classes, POCD-ND encoder makes use of unique vocabulary V and k-mer sequences of training set to generate position aware distribution matrices $Z = Z^{pos}, Z^{neg}$ for modification and non-modification classes, respectively. Here, positive (pos) class represents all the sequences of training set that belong to modification site and negative (neg) class denotes all the sequences of training set that belong to non-modification site.

$$Z^{pos/neg} = \begin{bmatrix} k-mer_{1,1} & k-mer_{1,2} & \cdots & k-mer_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ k-mer_{n,1} & k-mer_{n,2} & \cdots & k-mer_{n,k} \end{bmatrix} \quad (3.1)$$

In both matrices (Z^{pos}, Z^{neg}), each entry $k-mer_{ij}$ represents i^{th} k-mer at j^{th} position occurrence frequencies in positive and negative class sequences. Specifically, Z^{pos} is populated using following mathematical expression.

$$Z^{pos} = For_{i=1}^k (For_{j=1}^n (\sum_{t=1}^{NS^{pos}} (k-mer_{ij}) Occurrence)) \quad (3.2)$$

From left to right, first loop index i is an iterator on vocabulary V of k-mers, second loop index j is an iterator on all possible positions P and last loop index t is an iterator on all sequences of positive class to compute the count of positive sequences in which i^{th} k-mer appears at j^{th} position.

Similarly, Z^{neg} is populated using the following expression, where we compute the count of

negative sequences in which i^{th} k-mer appears at j^{th} position.

$$Z^{neg} = For_{i=1}^k(For_{j=1}^n(\sum_{t=1}^{NS^{neg}} (k - mer_{ij} Occurrence))) \quad (3.3)$$

Afterward, POCD-ND method computes k-mer position specific density values in positive and negative classes. The positive density of the i^{th} k-mer at j^{th} position denoted as $kmer_{ij}^{posden}$, can be computed by normalizing the k-mer $_{ij}$ occurrence frequency value with total number of positive sequences NS^{pos} .

$$\begin{aligned} k - mer_{ij}^{posden} &= \frac{Z^{pos}(k - mer_{ij})}{NS^{pos}}, \quad 0 \leq k - mer_{ij}^{posden} \leq 1 \\ &= p(k - mer_{ij} = 1 | NS^{pos}) \end{aligned} \quad (3.4)$$

Similarly, k-mer $_{ij}$ negative density value represented as $kmer_{ij}^{negden}$ is computed by normalizing the k-mer $_{ij}$ occurrence frequency value with total number of negative sequences NS^{neg} .

$$\begin{aligned} k - mer_{ij}^{negden} &= \frac{Z^{neg}(k - mer_{ij})}{NS^{neg}}, \quad 0 \leq k - mer_{ij}^{negden} \leq 1 \\ &= p(k - mer_{ij} = 1 | NS^{neg}) \end{aligned} \quad (3.5)$$

The prime assumption behind the development of existing encoder named position-specific trinucleotide propensity based on single-stranded characteristic (PSTNPss) [85] was to generate statistical representations of DNA sequences by generating 3-mers and assigning higher scores to those 3-mers which had more discriminative class densities. PSTNPss [85] encoder utilizes only 3-mers, however, different k-mers generate different types of discriminative patterns. To address this limitation, we present a more generalized version of PSTNPss [85] encoder that can be utilized for any k-mer to capture comprehensive discriminative patterns. Using the values computed through equations 3.4 and 3.5, modified PSTNPss [85] encoder scores can be computed using equation 7.9.

$$kmer_{ij} \text{ Class Density Difference (PSTNPss)} = kmer_{ij}^{posden} - kmer_{ij}^{negden} \quad (3.6)$$

Another major downfall of PSTNPss [85] encoder is that it assigns same scores to k-mers having different level of discriminative potential. Let's briefly discuss this drawback using a contour plot [138]. Figure 3.2 illustrates contour lines with respect to positive and negative class densities. In Figure 3.2, we have shown two different k-mers along the contour lines having positive to negative class density difference values equal to 0.3. Similarly, another pair of k-mers k3 and k4 are shown where positive to negative class density difference value is equal to 0.5. Analysis of positive and negative class densities along with differences for all four k-mers (Figure 3.2) indicates that from first pair, k1 (posden=0, negden=0.3, PSTNPss=0.3) located near to y-axis is of utmost importance on the contour line. Similarly, from second pair, k3 (posden=0, negden=0.5, PSTNPss=0.5) near to y-axis has more importance on the contour line. Across both pairs, as we move along the contour line away from the origin towards the top right-corner, posden and negden values are increasing. In first pair, k2 (posden=0.2, negden=0.7, PSTNPss=0.5) and in second pair,

k4 (posden=0.8, negden=0.5, PSTNPss=0.3) are less important than k-mer k1 and k3, respectively. This is because k1 and k3 are present in only negative class and absent in positive class whereas k2 and k4 are present in both classes. PSTNPss [85] encoder assigns equal score of 0.3 to first pair of k-mers and 0.5 to second pair of k-mers, indicating that it assigns equal scores to k-mers regardless of their occurrences in positive and negative classes, which shall not be the case.

To generate more comprehensive statistical representations of DNA sequences, our proposed encoder working paradigm relies on three main assumptions. (I) Like PSTNPss assumption [85], those k-mers are discriminative which have large position aware occurrence based positive to negative class density difference, (II) Those k-mers are more discriminative whose position aware occurrence based density is high in only one particular class and close to zero in other classes, (III) If two k-mers have equal $kmer_{ij}^{posden} - kmer_{ij}^{negden}$ difference, then the k-mers having lower $\min(kmer_{ij}^{posden}, k - mer_{ij}^{negden})$ value shall be assigned higher scores. Here min denotes the minimum function which returns the minimum value by comparing modification and non-modification class densities. A comprehensive detail of these assumptions is provided in motivating example section 3.2.1.1.

To generate statistical representations of DNA sequences based on above assumptions, proposed encoder makes use of following expression to assign scores to k-mers based on their discriminative potential.

$$\begin{aligned}
 POCD - ND (k - mer_{ij}) &= \frac{PSTNPss(k - mer_{ij})}{\min(kmer_{ij}^{posden}, k - mer_{ij}^{negden})} \\
 &= \begin{cases} 0.1, & \text{if } \min(k - mer_{ij}^{posden}) == 0 \\ 0.1, & \text{if } \min(k - mer_{ij}^{negden}) == 0 \end{cases} \quad (3.7)
 \end{aligned}$$

A complete workflow of the proposed POCD-ND encoder using a hypothetical corpus of 6 sequences is illustrated in Figure 3.1. POCD-ND encoder segregates the sequences into k-mers and divides the sequences into training and test k-mer sequences sets. It computes the vocabulary of unique k-mers and utilizes vocabulary and only training sequences to precisely generate statistical representations of corpus sequences in three steps: 1) Generate $kmer_{ij}$ position aware distribution matrices for modification and non-modification classes, 2) Compute $kmer_{ij}$ densities in modification and non-modification classes, 3) Compute k-mer position aware

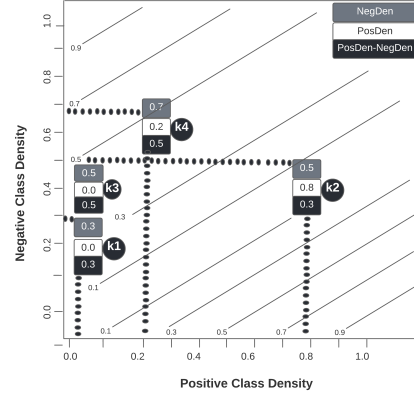


Figure 3.2: Contour plots for PSTNPss encoder where contour lines parallel to diagonal reveals PSTNPss encoder assigns same scores to k-mers that have same positive to negative class density differences

distribution based modification and non-modification class densities normalized difference. We assume that distribution of k-mers in test sequences is close to the distribution in training sequences. Hence, we utilize the k-mer position aware distribution matrices constructed using training sequences in order to generate statistical weights of k-mer test sequences by following the aforementioned steps 2 and 3.

3.2.1.1 A Motivational Example

We describe the effect of division by minimum k-mer position specific class density using a hypothetical example.

Table 3.1: A hypothetical dataset containing 15 sequences related to modification (c_1) and non-modification (c_2) classes. In the sequence samples, occurrence frequencies of a particular k-mer at 10 different positions

Sequences	Class	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀
1	c ₁	1	0	1	1	0	0	0	1	1	1
2	c ₁	0	0	1	1	1	0	0	0	0	1
3	c ₁	1	0	1	0	0	0	0	0	0	0
4	c ₁	1	1	1	1	0	0	0	0	1	1
5	c ₁	1	0	1	0	0	0	0	0	0	0
6	c ₂	0	1	0	0	0	1	0	0	0	1
7	c ₂	0	1	1	0	0	1	1	0	0	1
8	c ₂	1	1	0	0	0	1	0	0	0	1
9	c ₂	1	1	0	0	0	0	0	0	0	1
10	c ₂	0	0	0	1	0	0	0	0	0	0
11	c ₂	1	0	0	1	1	0	0	0	0	0
12	c ₂	1	0	0	0	0	1	0	0	0	0
13	c ₂	0	1	0	0	0	1	1	0	0	1
14	c ₂	0	1	1	0	0	0	0	0	0	1
15	c ₂	1	1	1	0	1	0	1	0	0	1

Table 3.1 indicates a hypothetical dataset containing 15 sequences related to two classes c_1 and c_2 . In each sequence, k-mer occurrences at 10 different positions are provided. The dataset is unbalanced because only 5 sequences belong to c_1 class and remaining 10 sequences belong to c_2 class.

Table 3.2 shows k-mer positive and negative class densities for the sample dataset and k-mer scores produced by PSTNPss and proposed POCD-ND encoder. Furthermore, we show the locations of k-mer at ten different positions in the Figure 3.3 where x-axis represents the k-mer density in c_1 class and y-axis represents the k-mer density in c_2 class.

In Figure 3.3, k-mers positions located in the top left and bottom right corners are most

Table 3.2: K-mers densities in, modification ($kmer_{ij}^{posden}$) and non-modification ($kmer_{ij}^{negden}$) classes, based on the k-mer densities, scores and ranks assigned by the existing PSTNPss [85] and proposed POCD-ND encoders to a particular k-mer present at 10 different positions

k-mer	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀
$kmer_{ij}^{negden}$	0.5	0.7	0.3	0.2	0.2	0.5	0.3	0	0	0.7
$kmer_{ij}^{posden}$	0.8	0.2	1	0.6	0.2	0	0	0.2	0.4	0.6
PSTNPss [85] Score	0.3	0.5	0.7	0.4	0	0.5	0.3	0.2	0.4	0.1
POCD-ND Score	0.6	2.5	2.23	2	0	5	3	2	4	0.17
PSTNPss [85] Rank	7	2	1	4	10	3	6	8	5	9
POCD-ND Rank	8	4	5	6	10	1	3	7	2	9

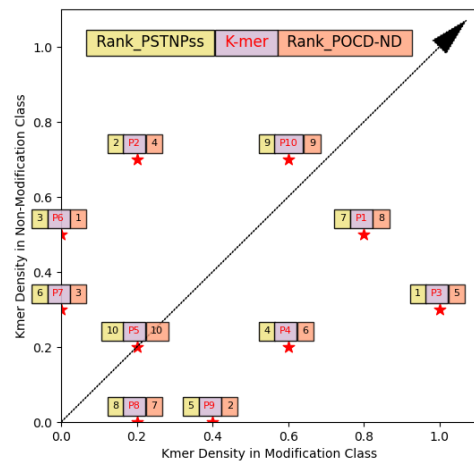


Figure 3.3: Graphical representation of a particular k-mer at 10 different positions

discriminative. K-mers located along the diagonal are the least discriminative as their occurrences in both classes are equal. The k-mers located on axes, except the k-mer closer to the origin are the most discriminative because they occur in only one class. A good encoding method shall assign higher scores to the k-mers located in top left and bottom right corners. We discuss why POCD-ND scores and rankings for most k-mer positions are different than PSTNPss [85] encoder scores and rankings.

- The k-mer at 1st and 7th positions has same score of 0.3 for PSTNPss [85] encoder. POCD-ND encoder assigns higher score to k-mer at 7th position and lower score to k-mer at 1st position. It is evident in the Figure 3.3 that k-mer at 7th position lies on y-axis as compared to k-mer at 1st position which lies on slight distance to diagonal at middle of lower and upper right corner. So intuitively, k-mer at 7th position is more discriminative and shall be assigned a higher score, as done by the POCD-ND encoder.
- The k-mer at 4th and 9th positions has equal PSTNPss [85] scores. We can see in the Figure 3.3 that k-mer at 9th position is far more important than k-mer at 4th position as it is very close to x-axis, hence POCD-ND encoder assigns higher score to k-mer at 9th position and lower score to k-mer at 4th position.
- The k-mer at 3rd position has the highest weight and rank for PSTNPss [85] encoder among the ten k-mer positions. POCD-ND places k-mer at 3rd position at fifth rank because of normalization with 0.3 that lowers its score as compared to k-mer at 2nd, 6th, 7th and 9th positions. It is clear from the Figure 3.3 that k-mer at 6th position lies on y-axis. Furthermore, this k-mer position is the nearest to top left or bottom right corner, hence this k-mer position is the most discriminative among all k-mer positions.
- The k-mer at 5th position is assigned the lowest score and rank by both PSTNPss [85] and POCD-ND as it lies on diagonal and has equal positive and negative class densities. Both encoders assign this k-mer position zero score.

As a whole, proposed POCD-ND encoder assigns better scores and ranks to k-mers at different positions by correctly quantifying their discriminative potential.

3.2.2 Benchmark Datasets

To evaluate the integrity of proposed DNA modification predictors, researchers have developed several benchmark datasets for different types of modifications and species [282, 342]. Recently, Lv. et al. [282] developed 12 different species related benchmark datasets and independent test sets for three different types of modifications namely 4mc, 5hmc and 6ma. These datasets are being used to evaluate the performance of newly developed predictors [392, 435]. Few of these datasets are utilized by Yang et al. [435] to evaluate the performance of their proposed

modification predictor and Tsukiyama et al. [392], also utilized 11 different datasets related to 6ma modification prediction.

The datasets related to 4mc modification prediction contain DNA modification sequences of 4 different species i.e., *Casuarina equisetifolia*, *Fragaria vesca*, *Saccharomyces cerevisiae* and *Ts. SUP5-1*. These datasets have been developed by collecting DNA modification sequences from the MDR database [275]. 5hmc modification has 2 benchmark datasets belonging to *Homo sapiens* and *Mus musculus* species. These datasets have been developed by collecting DNA modification sequences from the NCBI GEO database [188]. Whereas, 6ma modification datasets contain DNA modification sequences related to 11 different species i.e., *A. thaliana*, *C. elegans*, *C. equisetifolia*, *D. melanogaster*, *Homo sapiens*, *S. cerevisiae*, *Xoc.BLS256*, *T. thermophile*, *R. chinensis*, *F. vesca* and *Ts. SUP5-1*. These datasets have been collected from different sources i.e., MDR database [275], MethSMRT database [441] and NCBI GEO database [188]. For all DNA modification prediction datasets, negative samples are collected by satisfying the requirement that the 41 nucleotides long sequences with Cytosine/Adenine in the center are proved not to be modified by experiments [282]. A comprehensive detail of 17 different datasets is provided in Figure 3.4.

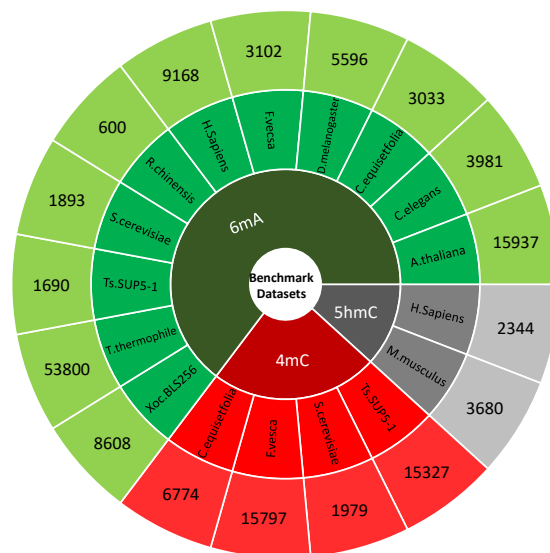


Figure 3.4: Distribution of sequences in 17 benchmark datasets related to 4mc, 5hmc and 6ma modifications

To evaluate the integrity of proposed approach, we have used these datasets for multiple reasons. First, three most recent DNA modifications predictors have reported their performance values over these datasets, which has facilitated us with a way to directly compare our model performance with existing predictors. Second, the datasets contain a significant number of sequences for the training and testing of the models. Third, the datasets are well distributed across different DNA modifications with respect to different species which makes the selection of these datasets a versatile choice to test any statistical representation generation method or classification model. Lastly, the datasets contain sequences that are rich in terms of motifs, whether the sequences are in the same species or in different species, which can aid in cross-species validation of the model as explained in the study [282]. The negative samples are random biological sequences not having similar motifs like positive samples, i.e., not having cytosine in the middle of the sequences [282].

3.3 Evaluation Criteria

Following evaluation criteria of existing DNA modifications predictors [8, 257, 342], to evaluate the integrity of proposed DNA modifications predictor by making a fair performance comparison with existing DNA modifications predictors, we assess the performance of proposed predictor in terms of 6 different evaluation measures i.e., accuracy (ACC), specificity (SP), sensitivity (SN), Matthews correlation coefficient (MCC), and area under the receiver operating characteristic (AU-ROC).

3.4 Results and Discussions

This section briefly illustrates the performance produced by the proposed DNA sequence encoder at different k-mers using a random forest (RF) classifier. It comprehensively illustrates the efficacy of the proposed DNA sequence encoder with 10 different machine learning classifiers for 17 datasets of 12 different species related to three different DNA modifications (4mc, 5hmc and 6ma) in two different settings, k-fold cross-validation and independent test sets based evaluation. Furthermore, it compares the performance of proposed DNA sequence encoder with existing 32 different encoders using 10 different classifiers for the prediction of 3 distinct DNA modifications under 2 different paradigms, extrinsic evaluation and intrinsic evaluation. Finally, it compares the performance of proposed generic DNA modifications predictor with existing generic i.e., iDNA-MS [282], iDNA-MT [435] and single type DNA modification predictors i.e., DCNN-4mc [342] and Bert6ma [392].

3.4.1 Performance Analysis of Proposed DNA Sequence Encoder at Different k-mers

The k-mer size directly impacts the feature representation by yielding either common or rare patterns. Lower-order k-mers (1-mer and 2-mer) based features are known as frequent features, as the occurrence frequency of these k-mers among sequences belonging to different classes would be approximately same. However, their occurrence distribution with respect to positions may vary. On the other hand, higher-order k-mers (3-mer, 4-mer and 5-mer) are less frequent features, since their occurrence frequencies vary significantly among sequences of different classes. While generating statistical representation of raw sequences, the proposed encoder makes use of the position specific occurrence information of k-mers. Hence, to analyze whether lower or higher-order k-mers generate better representations, we perform the performance analysis on different size k-mers with respect to different datasets. Figure 3.5 illustrates the accuracy produced by the RF classifier using statistical vectors generated by the proposed encoder “POCD-ND” with 5 different k-mers ($k=(1, \dots, 5)$) under the hood of 5-fold cross validation.

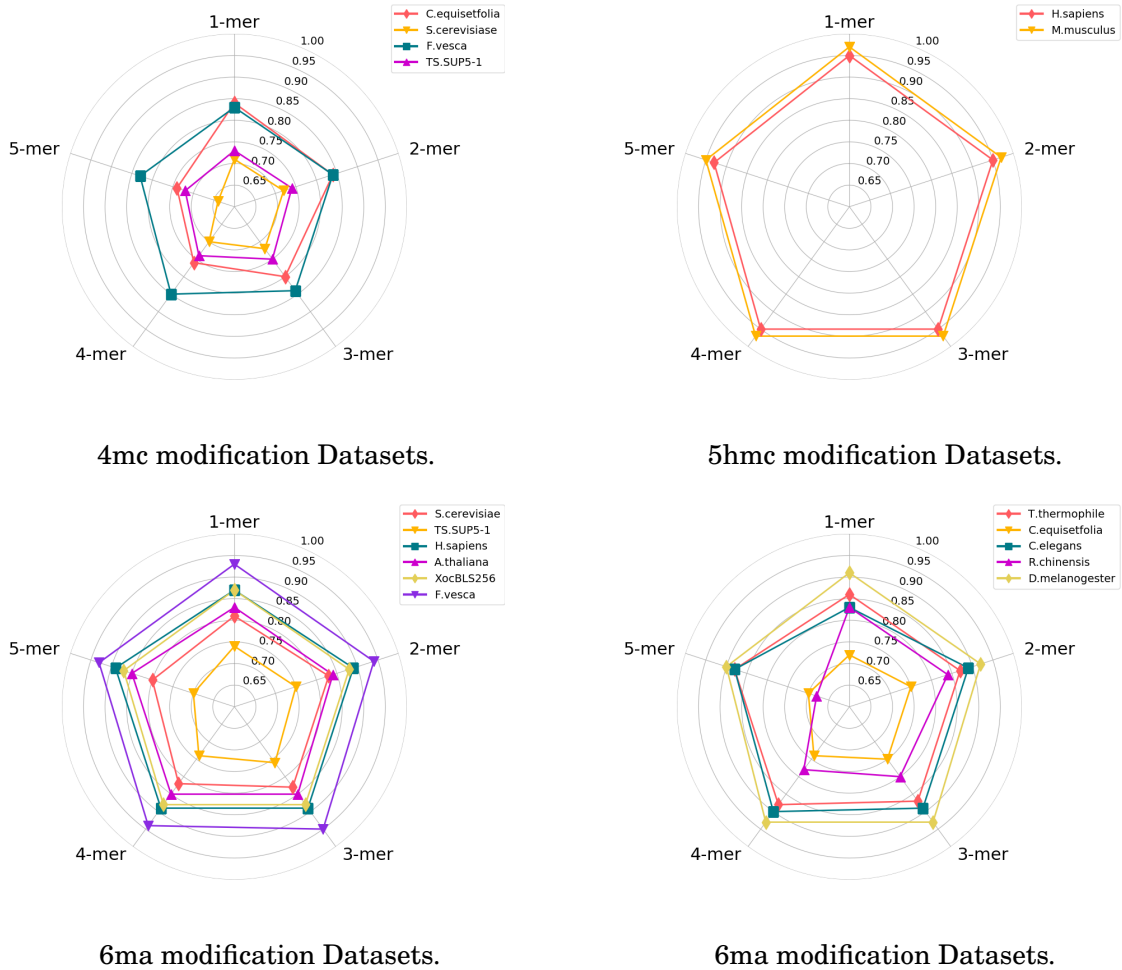


Figure 3.5: Random Forest classifier based extrinsic performance analysis of statistical representations generated at different k-mers using proposed encoder for different modification datasets.

For the task of 4mc modification prediction, accuracy analysis of 4 different species benchmark datasets indicates that, on two datasets namely *TS.SUP5-1* and *F.vesca*, performance of RF classifier slightly increases until 4-mer and 3-mer, respectively, however, drops afterwards. On *C.equisetifolia* dataset, performance of RF classifier drops with the increase of k-mer and RF marks lowest performance on 5-mer. On *S.cerevisiae* dataset, performance increase to 72% and remains same until 3-mer however, drops to 64% at 5-mer.

For 5hmc modification prediction datasets, on *H.sapiens* dataset, 95% performance of RF classifier remains same until 4-mer but drops by 2% at 5-mer. A similar performance trend is evident on *M.musculus* dataset where the RF classifier performance of 97% remains same until 4-mer but drops to 95% at 5-mer.

For 6ma modification prediction datasets, on three datasets namely *S.cerevisiae*, *C.equisetifolia* and *R.chinensis*, accuracy of RF classifier fluctuates at different test k-mers. On *t.thermophile*,

accuracy is slightly improved at each k-mer starting from 1-mer to 5-mer. On *Tolypocladium* and *H.sapiens*, accuracy of RF is slightly improved with the increase of k-mer size from 1-mer to 3-mer, afterward, it slightly drops on remaining k-mers. On all other 5 6ma datasets, the performance of RF classifier improves until 4-mers before dropping at 5-mer.

In a nutshell, the most dominant trend in 4mc and 6ma modifications prediction datasets is that the performance of RF slightly improves with the increase of k-mer up to 3-mer or 4-mer as compared to 5hmc datasets where performance remains the same with few k-mers and drops on 5-mer. Across most datasets of 3 different DNA modifications predictions, RF classifier achieves lowest performance with 5-mer. A variety of trends analyzed on different DNA modifications prediction datasets reiterate the importance of selecting the most appropriate window size while performing any DNA sequence classification task.

3.4.2 Performance Impact of Proposed DNA Sequence Encoder on Different Classifiers

This section performs 5-fold cross-validation based performance comparison of 10 different classifiers using the statistical representations generated by the proposed DNA sequence encoder “POCD-ND”.

Accuracy achieved by different classifiers under the hood of 5-fold cross-validation is shown in the Figure 3.6. It is evident that, using novel sequence encoder POCD-ND statistical representations, across all 17 benchmark datasets related to three different DNA modifications, from all machine learning classifiers, tree based classifiers produce better accuracy. Furthermore, from tree based classifiers, Random Forest (RF), GradientBoost (GB) and Extra tree (ET) based classifiers achieve better performance as compared to adaboost (AB) and decision tree (DT) classifiers, achieving the top accuracy around 85%, 84%, 75% and 73% on 4mc datasets such as *F.vesca*, *C.equisetofolia*, *TS.SUP5-1* and *S.cerevisiae*, respectively.

These classifiers achieve the accuracy of 95% and 97% on 2 benchmark 5hmc modification datasets related to two species *Homo sapiens* and *M.musculus*, respectively. Furthermore, these classifiers achieve best performance on all 11 6ma benchmark datasets where the accuracy falls in range of 73% to 95%. The primary reason behind the dominance of tree based machine learning classifiers such as RF is its ability to operate on random subset of features using multiple individual tree and combine the output of individual decision trees to generate the final output.

After tree based classifiers, Multi-Layer Perceptron (MLP) and K-nearest neighbour (KNN) classifiers achieve decent performance across most DNA modifications prediction datasets followed by Logistic Regression (LR). Generative classifier such as Naive Bayes (NB) performs better than its counterpart discriminative classifier Support Vector Machine (SVM) across all datasets except two 5hmc modification prediction dataset namely *Homo sapiens* and *M.musculus* where SVM performs better. Overall decision tree and SVM achieve lower accuracies on most benchmark DNA modifications prediction datasets.

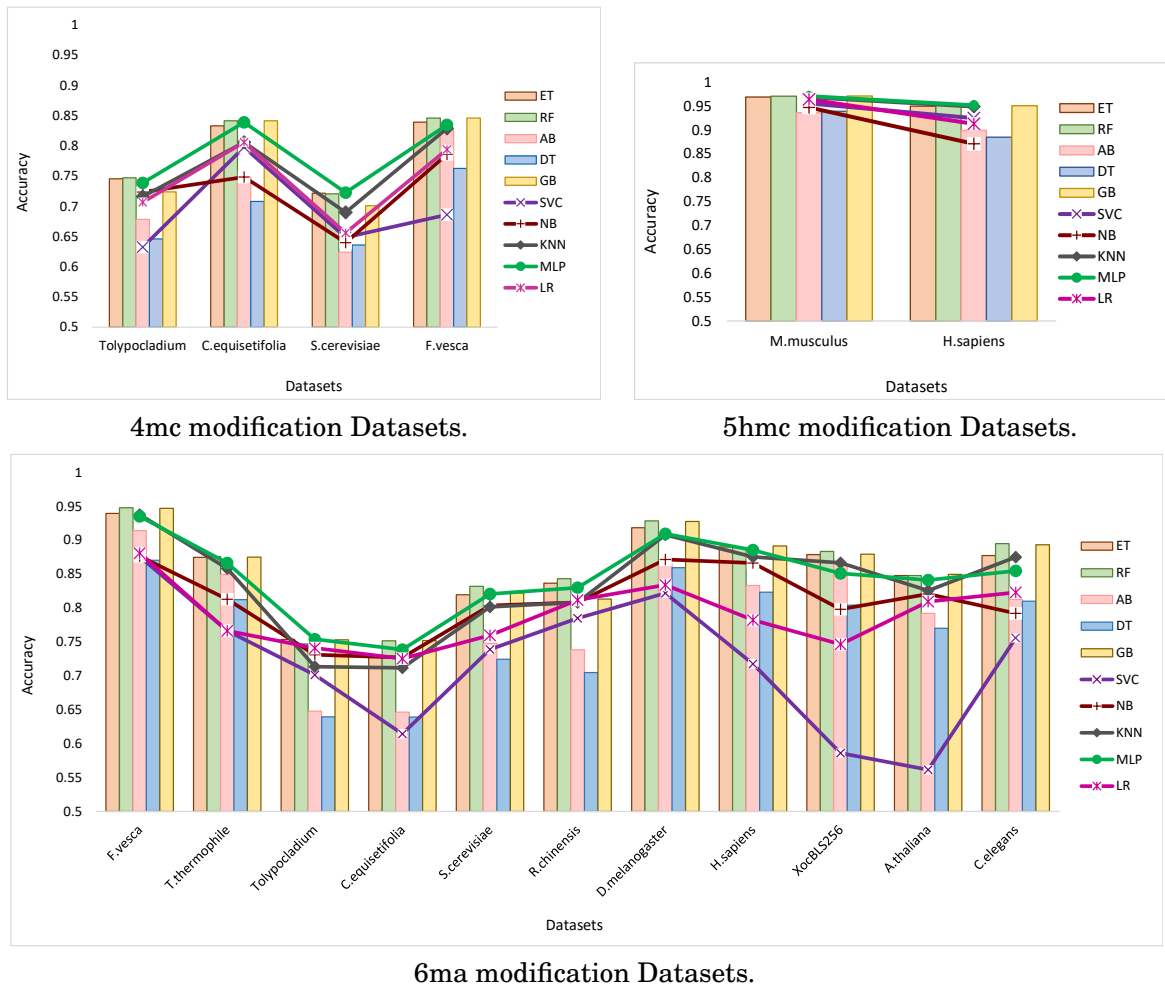


Figure 3.6: 5-fold cross-validation based performance comparison of 10 classifiers using 17 modification datasets related to (a) **4mc modification datasets**, (b) **5hmc modification datasets** and (c) **6ma modification datasets**.

In a nutshell, detailed analysis of performance potential of proposed novel sequence encoder indicates that proposed encoder most effectively characterizes DNA sequences and generate comprehensive discriminative patterns which enable even simple machine learning classifiers to achieve good DNA modifications performance across multiple datasets of distinct species.

As discussed earlier, size of k-mers largely impacts the performance of machine learning classifiers. We perform 5-fold cross-validation based evaluation of ten different classifiers using statistical representations generated by novel sequence encoder with different k-mers, falling in range of 1-to-5. Most machine learning classifiers perform better with lower sized k-mers falling in range of 1-to-3 for most DNA modifications prediction datasets. This is primarily due to the fact that with lower size k-mer, size of vocabulary is limited and statistical representation is generated by focusing on comprehensive discriminative position specific distributional patterns of k-mers. However, with the increase of k-mer size, vocabulary also increases which brings a lot

of rare k-mers as well which have limited discriminative position specific distributional patterns in sequences.

Extrinsic Performance Comparison of the Proposed DNA Sequence Encoder with Different Existing Encoders

The efficiency of the proposed POCD-ND encoding approach is further proven by performing a comprehensive extrinsic performance comparison of POCD-ND with 32 different existing encoding methods, out of which 11 are physicochemical properties based encoders, 12 are mathematical encoders and 8 are gap-based encoders. We feed the statistical representations generated by different sequence encoders to 10 different machine learning classifiers to predict three different DNA modifications across 17 different datasets. Accuracy values produced by different sequence encoders in combination with distinct best performing machine learning classifiers for three different DNA modifications prediction are mentioned in Table 3.3.

For 4mc, 5hmc and 6ma modifications prediction, from the category of physicochemical properties based encoders, SCPSeDNC, PCPSeDNC, PSEIIP and PseDNC encoders achieve better performance using RF and ET classifiers. Whereas TAC and TCC encoders mark lower performance for different DNA modifications prediction. Top four better performing encoder achieve an average performance of 94%, 84.5%, 80.7%, 79.5% on *C.equisetifolia*, *F.vesca*, *S.cerevisiae* and *TS.SUP5-1* datasets and overall average performance of 85% for 4mc modification prediction. These encoders achieve an average performance of 74.5%, 88.3% on *H.sapiens* and *M.musculus* datasets and an overall average performance of 81% for 5hmc modification prediction. Likewise, these encoders obtain an overall average performance of 84% for 6ma modification prediction.

Across different DNA modifications prediction, from the category of mathematical encoders, PSTNPss [85] achieve best performance with tree based machine learning classifier achieving an average performance of 89%, 95%, 93% on benchmark 4mc, 5hmc and 6ma modifications prediction datasets. Furthermore, sequence encoders namely pseudoKNC, kmer, RCKMER and spectrum achieve top four performance values using tree based machine learning classifiers. Whereas three sequence encoders namely orf, gc content and atcg ratio achieve least performance using three different classifiers SVM, LR and DT classifiers for 4mc modification prediction. For 5hmc modification prediction fickett score, cumulative skew and orf mark lower performance and for 6ma modification prediction, gcontent and orf sequence encoders mark lower performance across most benchmark datasets using different classifiers. Better performing mathematical encoders in combination with tree based machine learning classifier achieve an overall average performance of 85%, 82% and 86% on benchmark 4mc, 5hmc and 6ma modifications prediction datasets, respectively.

Furthermore, from the category of gap-based encoders, CKSNAP, monoDiKgap, diDiKgap and diMonoKGap achieve better performance using tree based machine learning classifiers for 4mc modification prediction. For 5hmc modification prediction, top five better performing se-

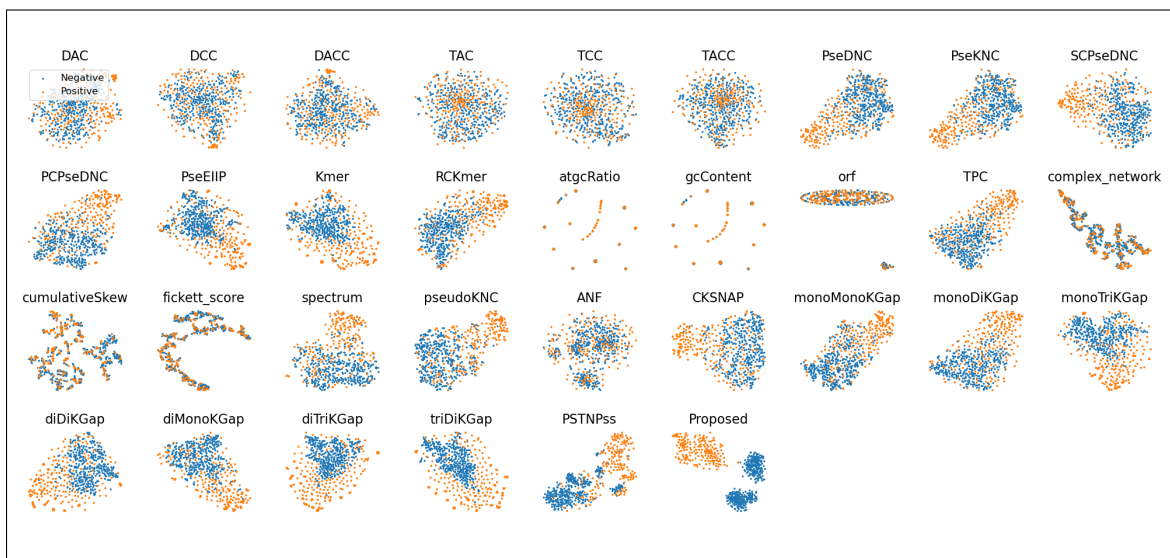


Figure 3.7: Intrinsic performance analysis of proposed POCD-ND and 32 existing encoders using 4mc modification prediction dataset *C.equisetifolia*.

Overall, for 4mc modification prediction, mathematical encoders achieve better performance. For 5hmc modification prediction, gap based sequence encoders mark better performance. For 6ma modification prediction, gap based and physicochemical properties based sequence encoders achieve better performance.

Proposed POCD-ND encoding method outperforms all 32 different sequence encoders by a decent margin across all three different DNA modifications prediction using tree based machine learning classifier. On 4 benchmark datasets related to 4mc modification prediction, it outperforms better performing mathematical encoders by an overall average accuracy of 5% and existing best performing encoder PSTNPss [85] by an average accuracy of 2% using tree based classifier. Furthermore, it beats the performance of gap-based and physicochemical properties based sequence encoders by an overall average accuracy of 14% and 9% on benchmark 5hmc and 6ma modifications prediction datasets, respectively.

The primary factor behind the dominant performance of proposed POCD-ND encoding method is its ability to capture discriminative position specific distributions of lower and higher order nucleotides called k-mers. PSTNPss [85] fails to assign appropriate scores to different level discriminative k-mers as briefly described in section 3.2.1. In addition, few physicochemical properties based sequence encoders solely focus on positional information of k-mers, other encoders pay more attention to distributional information of k-mers, indicating existing encoders fail to capture position aware discriminative distribution of k-mers. Furthermore, in order to encode k-mers, selecting the most appropriate physical or chemical properties from a huge set of properties is difficult. Researchers usually employ expert knowledge or extensive hit-and-trial experimentation methods to find appropriate k-mers properties, however, both methods

lack generalizability which is why such approaches fail to mark consistent performance across different datasets of the same or distinct species. In addition, because of unique functional paradigm, K-gap based sequence encoders generate large vocabulary which make length of statistical vectors very large and negatively impact the predictive performance.

In a nutshell, unlike proposed novel sequence encoding approach, all categories of sequence encoders including physicochemical properties, mathematical and k-gap based sequence encoders lack to capture comprehensive position aware discriminative distribution of nucleotides which are crucial to distinguish different class DNA sequences that commonly have many regions based on repetitive entries of four basic nucleotides.

3.4.3 Intrinsic Performance Comparison of the Proposed DNA Sequence Encoder with Different Existing Encoders

To perform an intrinsic evaluation of proposed encoder and existing encoders, we have randomly selected the dataset of *Casuarina equisetifolia* from 4mc type DNA modification. After generating statistical representations using all encoders, we have utilized t-distributed stochastic neighbor embedding (TSNE) approach to reduce statistical vectors to two dimensions that are graphically illustrated in Figure 3.7.

Overall, mathematical encoders like complex network, orf, fickett score, cumulative skew, atgskew, gccontent and physicochemical properties based encoders like DAC, DCC, DACC, TAC and TACC, lead to the formation of poor clusters which suggests the poor quality of the encodings generated by these encoders. Comparatively, physicochemical encoders like PseDNC, PseKNC, SCPseDNC, PCPseDNC, PseEIIP and mathematical encoders like k-mer, RCKMER, TPC and most of the gap based encoders lead to the formation of unique yet dependent clusters. Among all existing encoders, mathematical encoder based on simple k-mer class dependent densities difference namely PSTNPss [85] generates less overlapping clusters. Although, these encodings can be used for classification purposes, but there is a strong possibility that the classification results may not be optimal. This is because clusters generated by all existing sequence encoders are not highly disjoint, indicating existing sequence encoders fail to extract and encode discriminative patterns of nucleotides while generating statistical representations of DNA sequences. In comparison to existing encoders, the visualization of the encodings generated by the proposed encoder reveals a clear difference in the formation of unique and independent clusters, which proves the efficiency and strength of the proposed encoder for capturing discriminative features from DNA sequences.

In a nutshell, among all existing methods, only PSTNPss [85] manages to generate partially disjoint clusters for modification and non-modification classes. However, there still exist a significant number of sequences which belong to modification class cluster but falsely placed in non-modification class cluster and vice versa. In contrast, proposed POCD-ND encoder generates fully disjoint clusters for modification and non-modification classes. Although the working

paradigm of PSTNPss [85] is almost similar to POCD-ND but PSTNPss [85] only generates DNA sequence encoding of 3-mers. However, the unique distribution of nucleotides at different sizes k-mers offers more comprehensive discriminative patterns of nucleotides. Furthermore, PSTNPss [85] only computes simple class densities difference. In this process, k-mers which occur more sequences of one class and very less sequences of other class get higher scores, however, a similar score is also assigned to other k-mers which occur in almost all sequences of one class but do not occur at all in other classes. In particular, those k-mers are more discriminative which occur in more sequences of one class and do not occur at all in second class at a particular position, which are rightly characterized by proposed POCD-ND encoder by normalizing the PSTNPss score with minimum of modification and non-modification class densities. This is why proposed POCD-ND manages to generate highly disjoint clusters for modification and non-modification classes as compared to existing PSTNPss [85] encoder.

3.4.4 Performance Comparison of Proposed DNA-MP Predictor with State-of-the-art Predictors

Table 3.4 and Table 3.5 compare the performance figures of proposed DNA-MP predictor with four most recent predictors namely Bert6ma [392], iDNA-MT [435], DCNN-4mc [342] and iDNA-MS [282] in terms of accuracy, sensitivity, specificity, MCC and AUROC, for three different types of DNA modifications prediction over two different settings 5-fold and independent test sets.

Table 3.4: 5-fold cross-validation based performance comparison of the proposed DNA modification predictor with existing generic and type specific predictors i.e., iDNA-MS [282], DCNN-4mc [342] and Bert-6ma [392], across 17 different benchmark datasets in terms of 5 evaluation measures.

Measures	Predictor	4mc				5hmc		6ma										
		C.equisetifolia	F.vesca	S. cerevisiae	Ts. SUP5-1	H.sapiens	M.musculus	A.thaliana	C.elegans	C.equisetifolia	D.melanogaster	F.vesca	H.sapiens	R.chinensis	S.cerevisiae	Ts.SUP5-1	T.thermophile	Xoc.BL5256
Sensitivity	iDNA-MS [282]	0.727	0.846	0.705	0.725	0.974	0.963	0.82	0.852	0.707	0.906	0.942	0.852	0.84	0.771	0.749	0.959	0.859
	DCNN-4mc [342]	0.765	0.81	0.731	0.773													
	Proposed DNA-MP	0.841	0.834	0.72	0.747	0.951	0.97	0.848	0.895	0.751	0.928	0.947	0.889	0.843	0.832	0.747	0.876	0.883
Specificity	iDNA-MS [282]	0.724	0.807	0.732	0.711	0.929	0.978	0.843	0.839	0.72	0.899	0.925	0.895	79.33	0.823	0.726	0.757	0.861
	DCNN-4mc [342]	0.786	0.82	0.754	0.778													
	Proposed DNA-MP	0.841	0.846	0.72	0.747	0.951	0.97	0.848	0.895	0.751	0.928	0.947	0.889	0.843	0.832	0.747	0.876	0.883
Accuracy	iDNA-MS [282]	0.726	0.826	0.718	0.718	0.951	0.97	0.831	0.846	0.714	0.902	0.933	0.873	81.67	0.797	0.737	0.858	0.86
	DCNN-4mc [342]	0.772	0.82	0.675	0.725													
	Proposed DNA-MP	0.841	0.846	0.72	0.747	0.951	0.97	0.848	0.895	0.751	0.928	0.947	0.889	0.843	0.832	0.747	0.876	0.883
MCC	iDNA-MS [282]	0.452	0.654	0.438	0.437	0.905	0.941	0.664	0.692	0.429	0.805	0.86	0.748	0.634	0.596	0.476	0.733	0.721
	DCNN-4mc [342]	0.517	0.692	0.547	0.501													
	Proposed DNA-MP	0.685	0.669	0.442	0.494	0.904	0.941	0.696	0.79	0.504	0.857	0.895	0.779	0.686	0.665	0.4947	0.756	0.766
AUROC	iDNA-MS [282]	0.79	0.905	0.791	0.788	0.966	0.987	0.906	0.922	0.786	0.962	0.977	0.944	0.902	0.883	0.803	0.925	0.932
	Bert-6ma [392]							0.928	0.962	0.802	0.97	0.978	0.963	0.897	0.892	0.939	0.838	0.95
	DCNN-4mc [342]	0.887	0.911	0.78	0.801													
Proposed DNA-MP	0.889	0.921	0.804	0.82	0.962	0.986	0.916	0.953	0.822	0.971	0.979	0.952	0.909	0.903	0.82	0.936	0.949	

Table 3.4 compares the accuracy of proposed DNA-MP predictor with most recent DNA modification predictors over 17 different benchmark datasets using 5-fold cross validation. From existing 4mc modification predictors, 4mc type-specific predictor DCNN-4mc [342] achieves better

performance on two datasets as compared to generic iDNA-MS [282] predictor as it achieves an increment of 4% for *C.equisetofolia* and 1% for TS. SUP5-1 datasets. On *F.vesca* and *S.cervisiae* datasets, iDNA-MS [282] predictor marks better performance than DCNN-4mc [342] where it achieves the increment of 1% and 4%, respectively. Analysis of the performance in terms of other evaluation metrics indicates that DCNN-4mc [342] predictor achieves better MCC, specificity, AUROC across all four datasets, as well as achieves better sensitivity across most 4mc prediction datasets. The primary reason behind the limited performance of generic iDNA-MS [282] predictor is the ineffective statistical representation generated by inherently used different sequence encoders that fail to extract position aware discriminative distributions of k-mers.

Proposed DNA-MP predictor outperforms both iDNA-MS [282] and DCNN-4mc [342] predictors across all four 4mc prediction datasets in terms of five distinct evaluation metrics. It achieves an accuracy increment of 7%, 2%, 0.2% and 2% on *C.equisetofolia*, *F.vesca*, *S. cerevisiae* and *Ts. SUP5-1* datasets, respectively. Furthermore, it achieves a sensitivity increment of 7% on *C.equisetofolia* dataset, specificity increment of 5% and 3% on *C.equisetofolia* and *F.vesca* datasets, MCC increment of 17%, 4%, 11% and 6% and AUROC increment of 1%, 1%, 2%, 2% on *C.equisetofolia* *F.vesca*, *S.cerevisiae*, *TS.SUP5-1* datasets. Proposed DNA-MP predictor outperforms best performing existing 4mc predictor namely DCNN-4mc [342] across all four different datasets because it performs characterization of DNA sequences using k-mers position aware distribution based class densities difference as compared to DCNN-4mc [342] predictor which makes use of traditional one-hot encoding that lacks to capture contextual information of nucleotides while generating statistical representation of DNA sequences.

As shown by Table 3.4, for 5hmc modification prediction, proposed DNA-MP predictor equalizes the accuracies of iDNA-MS [282] predictor on *H.sapiens* and *M.musculus* datasets. In addition, it equalizes the performance on *H.sapiens* and *M.musculus* datasets across most of the other evaluation metrics including MCC, AUROC, sensitivity and achieves slightly better specificity on *H.sapiens* dataset.

It is evident in the Table 3.4 that, for 6ma modification prediction, proposed DNA-MP predictor outperforms existing iDNA-MS [282] predictor on all 11 different datasets by a comparable margin in terms of most evaluation metrics, specifically an average accuracy of 3%, average sensitivity of 2%, average specificity of 3% and average MCC of 5%.

Furthermore, on 4 independent test sets of 4mc modification prediction (Table 3.5, from existing predictors, DCNN-4mc [342] predictor achieves better performance in terms of most evaluation metrics. DCNN-4mc [342] outperforms other existing predictors by the accuracy of 8%, 14%, 7% and 12% on *F.vesca*, *S.cerevisiae*, *C.equisetofolia* and *TS.SUP5-1* species datasets, respectively. The ineffective statistical representations generated by one-hot encoding method and limited discriminative features extracted by bidirectional gated recurrent units hinder generic iDNA-MT [435] to achieve good performance across multiple datasets. Using effective statistical representations generated by novel encoder, proposed DNA-MP outperforms all existing

Table 3.5: Performance comparison of the proposed DNA modification predictor with existing generic and type specific predictors i.e., iDNA-MS [282], DCNN-4mc [342], iDNA-MT [435] and Bert-6ma [392], based on independent test sets in terms of 5 evaluation measures for 17 different DNA modification datasets.

Measures	Predictor	4mc				5hmc		6ma										
		C.equisetifolia	F.vesca	S.cerevisiae	Ts.SUP5-1	H.sapiens	M.musculus	A.thaliana	C.elegans	C.equisetifolia	D.melanogaster	F.vesca	H.sapiens	R.chinensis	S.cerevisiae	Ts.SUP5-1	T.thermophile	Xoc.BLS%6
Sensitivity	iDNA-MS	0.71	0.829	0.701	0.715	0.977	0.968	0.824	0.867	0.71	0.889	0.939	0.863	0.879	0.753	0.742	0.957	0.825
	iDNA-MT	0.83	0.826	0.692	0.72				0.873	0.714				0.796		0.742		
	DCNN-4mc	0.912	0.934	0.876	0.849			0.846	0.908	0.707	0.913	0.925	0.891	0.743	0.801	0.772	0.925	0.848
	Bert-6ma																	
	Proposed DNA-MP	0.931	0.918	0.886	0.873	0.947	0.968	0.94	0.955	0.885	0.974	0.974	0.959	0.926	0.918	0.891	0.909	0.945
Specificity	iDNA-MS	0.7	0.81	0.7	0.7	0.91	0.96	0.851	0.843	0.704	0.902	0.905	0.905	0.829	0.817	0.725	0.754	0.865
	iDNA-MT	0.83	0.79	0.728	0.731				0.857	0.745				0.826		0.767		
	DCNN-4mc	0.931	0.923	0.895	0.858			0.859	0.895	0.736	0.917	0.926	0.901	0.938	0.825	0.732	0.823	0.878
	Bert-6ma																	
	Proposed DNA-MP	0.931	0.918	0.886	0.873	0.947	0.968	0.94	0.955	0.885	0.974	0.974	0.959	0.926	0.918	0.891	0.909	0.945
Accuracy	iDNA-MS	0.71	0.823	0.7	0.71	0.947	0.96	0.837	0.855	0.711	0.896	0.922	0.884	0.854	0.785	0.734	0.856	0.845
	iDNA-MT	0.83	0.81	0.71	0.72				0.865	0.72				0.826		0.754		
	DCNN-4mc	0.902	0.905	0.845	0.835			0.853	0.902	0.721	0.915	0.926	0.896	0.781	0.813	0.752	0.874	0.863
	Bert-6ma																	
	Proposed DNA-MP	0.931	0.918	0.886	0.873	0.947	0.968	0.94	0.955	0.885	0.974	0.974	0.959	0.926	0.918	0.892	0.909	0.945
MCC	iDNA-MS	0.422	0.648	0.408	0.423	0.947	0.936	0.676	0.712	0.423	0.792	0.846	0.769	0.71	0.57	0.468	0.728	0.691
	iDNA-MT	0.666	0.635	0.413	0.448				0.731	0.438				0.913		0.511		
	DCNN-4mc	0.848	0.858	0.773	0.708			0.705	0.803	0.443	0.83	0.851	0.792	0.564	0.627	0.505	0.752	0.726
	Bert-6ma																	
	Proposed DNA-MP	0.871	0.848	0.793	0.772	0.896	0.936	0.887	0.914	0.791	0.949	0.95	0.921	0.862	0.849	0.802	0.832	0.896
AUROC	iDNA-MS	0.78	0.9	0.771	0.78	0.96	0.984	0.911	0.935	0.779	0.956	0.977	0.95	0.924	0.868	0.813	0.922	0.921
	iDNA-MT	0.904	0.896	0.776	0.798				0.937	0.792					0.822			
	DCNN-4mc							0.927	0.962	0.799	0.967	0.976	0.962	0.865	0.89	0.834	0.938	0.936
	Bert-6ma	0.97	0.97	0.946	0.914													
	Proposed DNA-MP	0.99	0.987	0.995	0.983	0.962	0.985	0.989	0.995	0.986	0.997	0.997	0.995	0.996	0.993	0.991	0.983	0.99

approaches by achieving an average sensitivity increment of 1%, average specificity increment of 0.5%, average accuracy increment of 3%, average MCC increment of 2% and average AUROC increment of 4%.

As shown in Table 3.5, on 2 independent test sets (H.sapiens, M.musculus) of 5hmc modification prediction, proposed DNA-MP predictor equalizes the accuracies of iDNA-MS [282] predictor. Furthermore, it achieves sensitivity and MCC values which are close to performance figures of iDNA-MS [282] predictor and achieve an average specificity increment of 3% and AUROC increment of 0.5%.

On 11 independent test sets of 6ma modification prediction (Table3.5), from existing predictors, Bert-6ma [392] predictor achieves better performance followed by two generic iDNA-MS [282] and iDNA-MT [435] predictors. Proposed DNA-MP predictor outperforms existing predictors by an average sensitivity, specificity and AUROC of 7% and MCC of 16%. Proposed DNA-MP predictor is better than existing predictors especially Bert-6ma [392] in multiple ways. Despite the utilization of 7 different sequence encoders (i.e., nucleotide chemical property, nucleotide frequency, binary encoding, Word2vec) and 8 different neural networks (CNN, LSTM, Hybrid, Bert), Bert-6ma [392] predictor lacks to capture position aware discriminative distribution of k-mers. Proposed DNA-MP predictor only makes use of a single novel sequence encoder to encode comprehensive position aware k-mer discriminative patterns and a simple RF classifier to most

accurately predict 6ma as well as 5hmc and 4mc modifications prediction.

To summarize, a comprehensive k-fold and independent test sets based performance comparison of proposed DNA-MP predictor with existing predictors over 17 benchmark datasets indicates that novel k-mer position aware distribution based class densities normalized difference encoder captures comprehensive position specific discriminative k-mers patterns which helps the Random forest classifier to most accurately predict different DNA modifications.

3.5 Conclusion

The contributions of this study are manifold. It presents position specific k-mer occurrence based class densities difference regularized through their minimum value to assign effective scores to k-mers that lead to better characterization of DNA sequences. Experimental results reveal that position aware higher order k-mers based statistical representations generate rare features which negatively impact the performance of machine learning classifiers. Intrinsic analysis indicates that using discriminative distributions of k-mers in which k-mers present in large number of sequences in one class at certain positions and either do not present at all or present in low number of sequences in other class at same positions, highly disjoint clusters for modification and non-modification classes are obtained as compared to 32 existing sequence encoders. Furthermore, these features also significantly enhance the generalizability of 10 different machine learning classifiers. Using novel sequence encoder, Random forest classifier manages to beat state-of-the-art type specific and generic DNA modifications predictors on 17 benchmark datasets of 12 different species. A compelling future line of current work would be to investigate the effectiveness of novel sequences encoder for other Genomics and Proteomics sequence classification tasks.

HISTONE OCCUPANCY/MODIFICATIONS PREDICTION AND ENHANCER IDENTIFICATION/STRENGTH PREDICTION

The way different living organisms grow, survive, develop and reproduce is regulated by an instruction manual called Deoxyribonucleic Acid (DNA) or genetic code [174, 281]. The genetic code is organized into chromatin in a series of nucleosomes, where in each nucleosome, DNA is wrapped around histone octamers that are made up of four pairs of histone proteins (H2A, H2B, H3 and H4). A graphical representation of nucleosome construction with Histone Octamer and DNA binding is illustrated in Figure 4.1. DNA consists of three main components namely: genes, non-coding DNA and regulatory elements. Mainly, genes produce proteins while other two components control the production type and number of proteins. To produce proteins, genes go through two different stages namely transcription and translation. Specifically, at transcription stage, negatively charged DNA sequence gets unwrapped from positively charged histone octamers in order to allow the regulatory network read the target instructions and produce the most appropriate proteins through the expression of certain genes. After the expression of genes, DNA sequence gets tightly wrapped once again around histone octamers because of the dynamics of opposite charges. Histone octamers get saturated with acetylation and methylation modifications that induce high negative and positive charges, respectively on histone octamers. Histone acetylation modification promotes gene expression, whereas histone methylation modifications hinder gene expression by making unwrapping of DNA sequences difficult. Therefore, without altering the DNA sequence, histone modifications influence the remodeling of chromatin, which eventually

⁰This chapter is an adapted version of the work presented in Asim et al., "Histone-Net: A Multi-Paradigm Computational Framework for Histone Occupancy and Modification Prediction", In *Complex & Intelligent Systems (2022)* [22] and Asim et al., "Enhancer-DSNet: A Supervisedly Prepared Enriched Sequence Representation for the Identification of Enhancers and Their Strength.", In *27th International Conference on Neural Information Processing, (ICONIP-2020)* [20]

impacts the gene expression. Cells maintain a balance between acetylation and methylation levels to express more genes or repress certain sets of genes to produce right amount of proteins essential for normal functioning of different organs. Irregularities in acetylation or methylation cause over-expression or under-expression of genes that initiate and propagate several diseases such as cancer, autoimmune diseases, mental disorders and diabetes.

Methylation of histone proteins H3 and H4 mainly regulates the core activity of DNA replication [113] and acetylation of different histone proteins affects the structure of the chromatin as well as gene transcription [99, 208]. Histone modifications are responsible to regulate multifarious biological processes including chromosome wrapping [34, 54], transcriptional activation and de-activation [46, 90, 225], damaging and repairing of DNA [228, 310]. To acquire a deeper comprehension of epigenetic regulation at cellular level and to pave way for the development of drugs specifically cancer treatment and histone altering enzymes [322], histone modification detection is essentially required [10]. A thorough analysis of histone acetylation and methylation areas in histone sequences can decipher the association of histone modification with metabolism that mediates diverse epigenetic abnormalities in multifarious pathological conditions [440].

Furthermore, to study the state of the nucleosome array before gene expression and nucleosome array recreation after gene expression, it is important to analyze histone occupancy. Determining whether DNA around histone octamer is tightly wrapped or loosely wrapped, a genetic task known as histone occupancy determination has profound importance in genetic research [235, 240]. Histone occupancy significantly influences epigenetic silencing [261], cell replication [343], differentiation [418] and re-programming [418]. Accurate determination of histone occupancy can facilitate a deeper understanding of DNA accessibility to proteins, chromatin functions and occupancy correlation with promoter strength [41].

In addition to histone modifications, regulatory elements primarily enhancers also impact the production of proteins. Specifically, genes are expressed when DNA sequence gets unwrapped from histone octamers and the strength of enhancers decides the duration for which a gene will remain active. If a gene unexpectedly remains deactivated for a longer time period, then there might be some issue with its associated enhancer. Enhancers impact cell growth, cell differentiation, cell carcinogenesis, virus activity and tissue specificity through enhancing genes transcriptions [209]. Discriminating enhancers from regulatory elements, estimating their location and overall

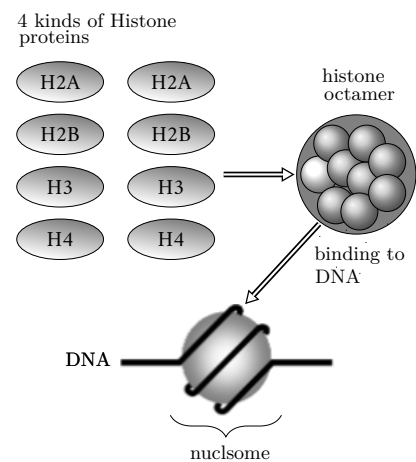


Figure 4.1: Histone octamer and nucleosome formation

strength are few most promising tasks which can facilitate deeper comprehension of eukaryotic spatio-temporal gene regulation and evolution of diseases [266].

4.1 Related work

This section summarizes existing computational approaches that have been developed for both application areas, namely enhancers and histone occupancy/modification prediction.

4.1.1 Histone Occupancy and Modification Prediction

Developing a robust computational approach for accurate histone occupancy and modification prediction has been an active area of research since the public availability of 10 benchmark datasets developed by Phaml et al. [332]. They proposed the very first computational approach that utilized occurrence frequency of k-mers to generate statistical representations of histone sequences and SVM classifier for histone modification prediction. Tran et al. [389] generated a Boolean representation of DNA sequences based on the presence and absence of k-mers. They utilized conditional random fields (CRFs) to infer nucleosome acetylation and methylation levels in DNA sequences. Considering the significance of both position and frequency of k-mers, Pham et al. [331] combined k-mer frequency based encoding with positional information based encoding to reap the benefits of both kinds of representation learning approaches. Aggregated representations were fed to SVM classifier for accurate prediction of histone occupancy, methylation and acetylation levels. Higashihara et al. [177] utilized the filter-based feature selection algorithm "Gini Index" to obtain highly discriminative k-mers and used their occurrence frequency to generate statistical representation that was passed to the SVM classifier. Benveniste et al. [39] developed a logistic regression based approach to infer histone modifications by solely utilizing transcription related factor-binding profiles to generate statistical representation.

On the other hand, taking into account the wide success of deep learning approaches, Nguyen et al. [312] proposed a deep learning based methodology by utilizing k-mer one-hot vector encoding and convolutional neural network (CNN). Likewise, Yin et al. [443] proposed another deep learning approach in which they transformed one-hot encoded vector of k-mer sequences into image-like tensors by making use of Hilbert curves. Image-based representations of histone sequences were passed to a CNN model for the extraction of discriminative features and the final classification. Aforementioned approaches were developed under the paradigm of binary classification where each sequence of every task is either associated with positive class or negative class.

A critical analysis of existing computational approaches [177, 312, 332, 443] indicates that these approaches lack in generating comprehensive statistical representations of histone sequences. To generate statistical representations of histone sequences, few approaches utilized bag-of-words based approaches [177, 332] which only manage to capture k-mer frequency and

neglect rich semantic information. Whereas, others have utilized one-hot encoding scheme that lacks the ability to capture comprehensive contextual information and correlations of k-mers [33, 251, 312, 444]. Although, image based representation manages to find discriminative k-mers, however, it failed to handle positional information of k-mers and transnational invariance of k-mers mainly due to the supreme attention towards local residue context.

Furthermore, existing computational predictors have not been evaluated in cross-domain setting [43, 183, 217, 471], where the key idea is to train model on one type of histone marker and predict histone occupancy and modifications on other type of histone marker. Also, a closer look at existing computational approaches reveals that in all existing approaches, 10 different model checkpoints are deployed by rigorously training the single model separately over 10 benchmark datasets to provide practical application for the prediction of histone occupancy, methylation and acetylation areas in histone sequences. In this strategy, one needs to know the target histone marker beforehand to select appropriate model checkpoint amongst all available model checkpoints while making prediction over unseen histone sequences. More recently, Yin et al. [444] developed a deep learning approach “DeepHistone” that can simultaneously predict different histone markers associated with particular sequence. Yin et al. [444] work motivates us to develop a single multi-label predictor that can simultaneously predict histone occupancy, acetylation and methylation levels associated to different histone markers.

4.1.2 Enhancer Identification and Strength Prediction

There exists a plethora of computational approaches which can discriminate enhancers from other regulatory elements, however, few 2-layer predictors have been proposed which can predict enhancers as well as their strength. This section summarizes enhancer identification and strength prediction approaches.

Liu et al. [266] presented the very first 2-layered computational predictor namely IENHANCER-2L that can discriminate enhancers from other regulatory elements as well as estimate their strength. They leveraged pseudo k-tuple nucleotide composition encoding method to transform DNA sequences and SVM classifier.

To improve the performance of IENHANCER-2L predictor, Jia et al. [209] developed Enhancer-Pred. In order to learn optimal representations of DNA sequences, they utilized 3 different feature encoding schemes including: Bi-profile Bayes, nucleotide composition and pseudo-nucleotide composition. They fed the optimized representations to SVM classifier. Liu et al. [265] developed another 2-layer computational framework namely iEnhancer-PsedeKNC that used pseudo k-mer nucleotide composition encoding method and SVM classifier. He et al. [172] proposed Enhancer-Pred2.0 that utilized 2 different physicochemical property based encoding methods electron-ion interaction potential and position-specific trinucleotide propensity. They utilized wrapper based feature selection and SVM classifier.

Liu et al. [268] presented another 2-layer predictor namely iEnhancer-EL which fused six

classifiers probabilities for layer-1 and 10 classifiers probabilities for layer-2 prediction. They performed sequence encoding using 3 different methods including k-mer, subsequence profile and pseudo K-tuple nucleotide composition. Tan et al. [379] proposed an ensemble of convolutional neural network and recurrent neural network for efficient identification of enhancers and their strength prediction. For generating sequence representations, they employed 6 different kinds of dinucleotide physicochemical properties. Le et al. [243] presented iEnhancer-5Step predictor based on neural k-mer embeddings and SVM classifier.

Nevertheless, still there is a lot of room for the improvement in the predictive performance especially in distinguishing strong enhancers from weak enhancers. To develop an optimal machine learning model for enhancer identification and strength prediction task, the most crucial step is to encode biomedical sequences into fixed-size low-dimensional vectors. In this context, few sequence encoding methods including: Local Descriptor, Conjoint Triad (CT), Auto Covariance (AC) and PSE-KNC [268] have been utilized. However, such methods fail to take semantic information of residues into account (such as residues order). To overcome these shortcomings up to certain extent, Le et al. [243] recently employed neural word embeddings prepared in an unsupervised manner. Although unsupervised k-mer embeddings capture semantic information of k-mers, however, they still lack to associate inherent k-mer relationships with sequence type while learning low-dimensional vector space.

Considering the downfalls of existing computational approaches in both application areas and following the success of FastText approach in diverse NLP tasks, this research presents a novel deep learning approach, working paradigm of which is similar to FastText model which incorporates class label information while learning discriminative k-mers weights of histone sequences. Furthermore, to assess the true generalization aptitude of proposed approach across 10 distinct histone markers belonging to histone occupancy, acetylation and methylation, proposed approach is evaluated in cross-domain setting, where aim is to evaluate whether proposed predictor can perform accurate predictions over new histone markers.

To develop a more comprehensive landscape for histone sequence analysis, we develop a multi-label classification dataset which will further assist researchers to develop a unified model for multiple histone markers related to three different tasks namely histone occupancy, acetylation and methylation level prediction. In multi-label classification paradigm, performance of proposed predictor is evaluated in terms of its ability to simultaneously predict histone type, occupancy, acetylation and methylation levels.

To explore the potential of proposed predictor for other Genomics sequence analysis tasks, we utilize same model for enhancers identification and their strength prediction tasks.

4.2 Materials and Methods

This section describes the details of proposed predictor and benchmark datasets related to two different tasks namely Histone occupancy/modification prediction and enhancer identification/strength prediction.

4.2.1 Proposed Methodology

Histone sequences are comprised of 4 repetitive nucleotides adenine (A), guanine (G), cytosine (C) and thymine (T). To develop a deep learning based predictor for analysis task of any biological sequence, first step is to generate k-mers by sliding a fixed-size window over raw sequences [109, 203]. However, while generating k-mers, it is important to decide the size of k-mers. Because performance of classifiers relies on the number of discriminative k-mers among different classes. To analyze which window size generates the most discriminative k-mers for histone occupancy, acetylation and methylation prediction tasks. we generate overlapping k-mers with 10 different degrees ranging from 2-to-12. We find that among positive and negative classes, there do not exist any discriminative k-mers until 7-mers, however, afterward with the increase of k-mers size, the number of discriminative k-mers also get increased which are present in one class and absent in other class.

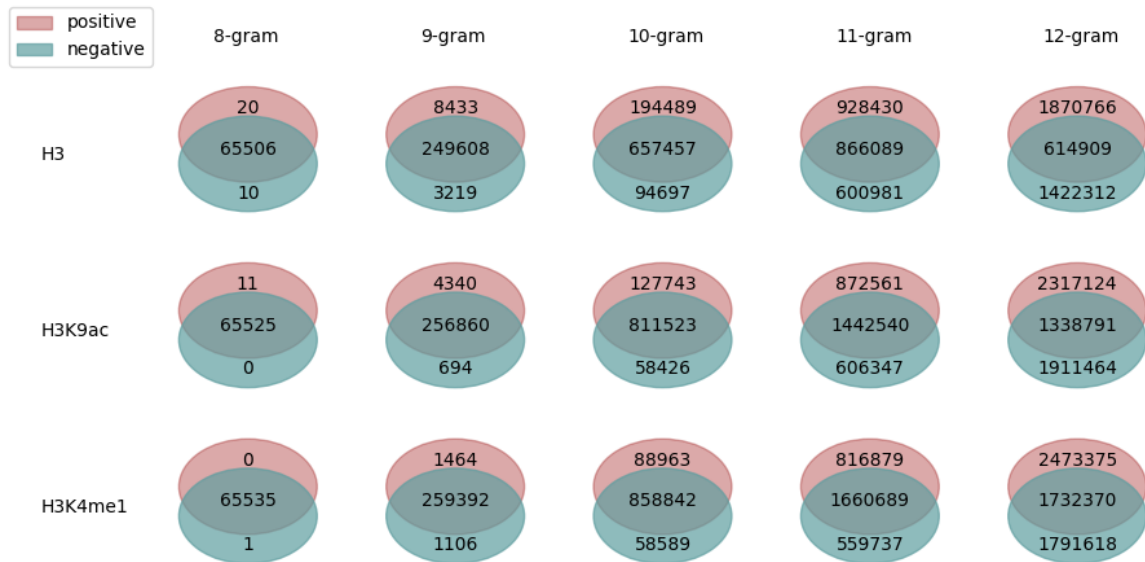


Figure 4.2: Discriminative and overlapping k-mers in positive and negative classes

In particular, to provide a bird's eye view of discriminative potential of different size k-mers across positive and negative classes, we randomly select a dataset from each histone sequence analysis task and reveal the discriminative k-mers for each class for 5 different higher order

residues (8-mers to 12-mers) in the form of Venn diagrams (Figure 4.2). As shown by the Figure 4.2, for histone occupancy H3 dataset, in case of 8-mers, positive class has only 20 unique k-mers which are not present in the negative class while negative class has 10 unique k-mers which are not present in positive class. Whereas, 65506 k-mers are present in both classes. With the increase in degree of k-mers, discriminative as well as overlapping k-mers also increase. Histone acetylation (H3K9ac) and methylation (H3K4me1) datasets also follow the distributional trend of H3 dataset.

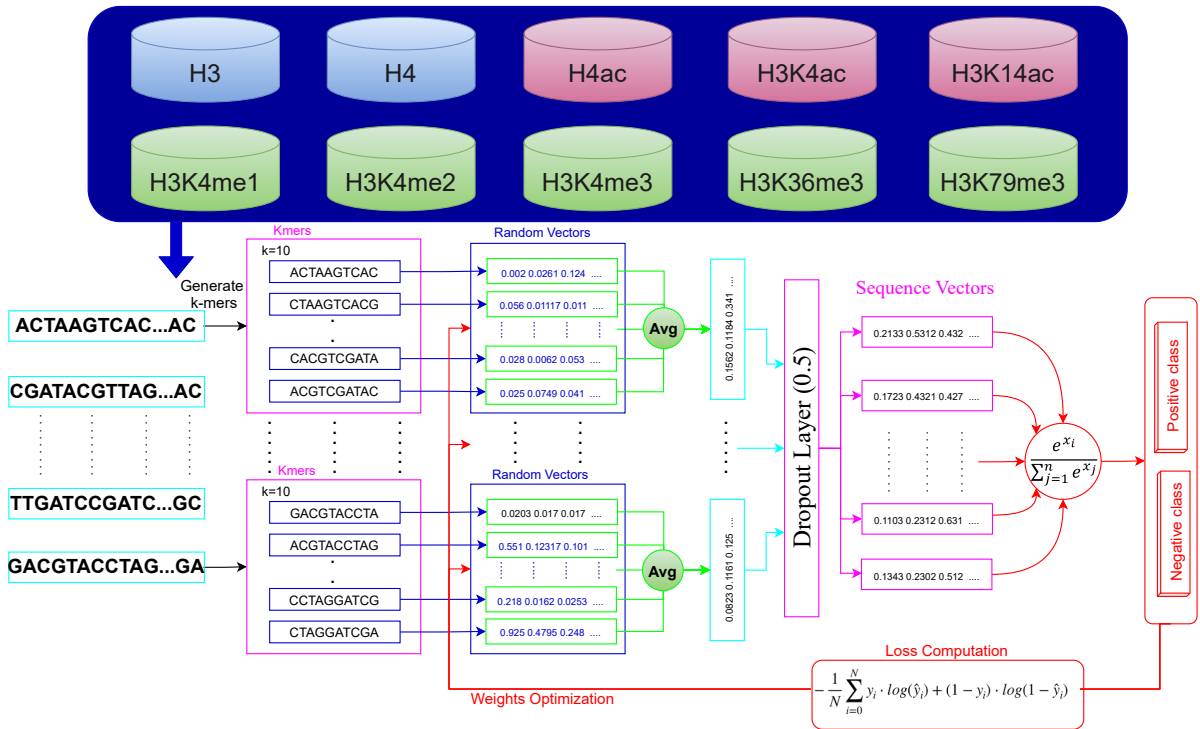


Figure 4.3: Workflow of Histone-Net [22] approach

Considering above discussed distribution of k-mers in both classes, the key idea behind the proposed approach is an assumption that traditional deep learning classifiers that make use of CNN, RNN or hybrid layers (CNN + RNN) cannot perform better for classification tasks where vocabulary of unique k-mers is very large. In particular, such humongous vocabulary contains several rarely occurring k-mers that confuse deep learning models while extracting more comprehensive features. To handle this problem, we propose a deep learning predictor competent in learning representations of k-mers by incorporating class label information. We believe in this particular scenario, rarely occurring k-mers will get less weights while discriminative k-mers will get more weights.

A graphical illustration of proposed predictor is provided in Figure 4.3, where for binary classification problems related to 3 different tasks of histone sequence analysis and two different tasks of enhancer sequence analysis, the process of generating task aware distributed representa-

tion of histone sequences starts by dividing the sequences into higher order residues (K-mers, e.g., 10-mers). Then, distributed representations of sequences are learned by averaging the randomly initialized 100-dimensional vectors of k-mers present in sequences, regularization of which is performed by applying the dropout rate of 0.5. The training objective of embedding generation model is to optimize k-mer embedding matrix by monitoring the cross entropy loss produced while estimating the probability of target class (overall classes) through softmax classifier. Through iterative learning, proposed predictor manages to encapsulate histone occupancy, acetylation and methylation information in embedding matrix.

While in case of multi-label classification paradigm, we expose 3 histone sequence analysis tasks related to meta-data at once to generate more generalized 64-dimensional k-mer embeddings followed by sigmoid classifier to simultaneously get the probabilities with respect to histone marker type, occupancy, acetylation and methylation levels. Histone-Net [22] multi-label classification paradigm effectively handles the overhead of training 10 different binary classifiers for 10 histone markers belonging to histone occupancy, acetylation and methylation.

As proposed model is evaluated in two different scenarios, in case of binary classification paradigm it is evaluated for two different applications areas histone sequence analysis and enhancer sequence analysis. To better elaborate the performance, in case of histone sequence analysis, we name this approach Histone-Net [22], while for enhancer sequence analysis, we name this approach as Enhancer-DSNet [20]. In multi-label scenario, we name this approach as Histone-Net multi-label.

4.2.2 Benchmark Datasets

This section describes benchmark binary classification datasets for both application areas. It also summarizes the process used to develop a multi-label classification dataset for histone occupancy and modification prediction.

4.2.2.1 Benchmark Binary Classification Datasets for Histone Occupancy and Modification Prediction

This section illustrates the details of 10 public benchmark histone occupancy and modifications (acetylation and methylation) prediction datasets [443] used to evaluate the performance of proposed multi-paradigm computational framework Histone-Net [22]. A comprehensive detail of experimental process used to prepare 10 benchmark datasets is described in [331], here we only summarize the statistics of 10 benchmark datasets. Table 4.1 describes sequence-to-label distribution of 2 histone occupancy (H3, H4), 5 methylation (H3K4me1, H3K4me2, H3K4me3, H3K36me3, H3K79me3) and 3 acetylation datasets (H3K9ac, H3K14ac, H4ac). For acetylation and methylation level prediction datasets, K with its leading number represents the K^{th} amino acid which has to be modified with mono, di, or tri acetyl (“ac”) and methyl (“me”) modifications.

For example, in H3K4me1 dataset, 4th amino acid of H3 protein is modified with a mono methyl group.

For each benchmark dataset, histone sequences having relative occupancy, methylation and acetylation values greater than 1.2 belong to positive class and lower than 0.8 belong to negative class.

Table 4.1: Statistical summary of 10 benchmark datasets including 2 datasets for histone occupancy detection, 3 datasets for acetylation and 5 datasets for methylation level prediction.

Dataset Name	Description	Positive Samples	Negative Samples
H3	H3 occupancy	7667	7298
H4	H4 occupancy	6480	8121
H3K4me1	H3K4 mono-methylation relative	17266	14411
H3K4me2	H3K4me2 H3K4 di-methylation relative to H3	18143	12540
H3K4me3	H3K4me3 H3K4 tri-methylation relative to H3	19604	17195
H3K36me3	H3K36me3 H3K36 tri-methylation relative to H3	18892	15988
H3K79me3	H3K79me3 H3K79 tri-methylation relative to H3	15337	13500
H3K9ac	H3K9 acetylation relative to H3	15415	12367
H3K14ac	H3K14 acetylation relative to H3	18771	14277
H4ac	H4 acetylation relative to H4	18410	15686

4.2.2.2 Multi-label Classification Dataset for Histone Occupancy and Modification Prediction

This section describes the process used to develop multi-label classification dataset for histone occupancy and modification prediction.

Figure 4.4 illustrates the complete workflow used to develop imbalanced and balanced version of multi-label histone sequence analysis dataset by utilizing 10 benchmark datasets given by Pham et al. [331]. All 10 benchmark datasets have total of 2.74 million sequences where each sequence is annotated with either 0 or 1. A closer look at sequence ids provided by Pham et al. [331] reveals that a significant number of sequence ids appear in multiple histone markers datasets. For instance, consider a sequence id "iTELL-Chr1_61" which is annotated as 1 in H3 histone marker dataset indicates that the sequence has histone occupancy the more than 1.2, same sequence id is annotated as 0 in H3k4me1 histone marker dataset indicating that the sequence methylation level is less than 0.8 and same sequence id is annotated as 1 in H4kme2 histone marker dataset indicating that the sequence methylation level is more than 1.2.

This analysis serves as a basis to formulate multi-label dataset where each sequence id may have 20 labels at max instead of one label (0 or 1). From 20 labels, 10 labels represent the association of sequence with positive class distribution of 10 benchmark histone markers whereas

CHAPTER 4. HISTONE OCCUPANCY/MODIFICATIONS PREDICTION AND ENHANCER IDENTIFICATION/STRENGTH PREDICTION

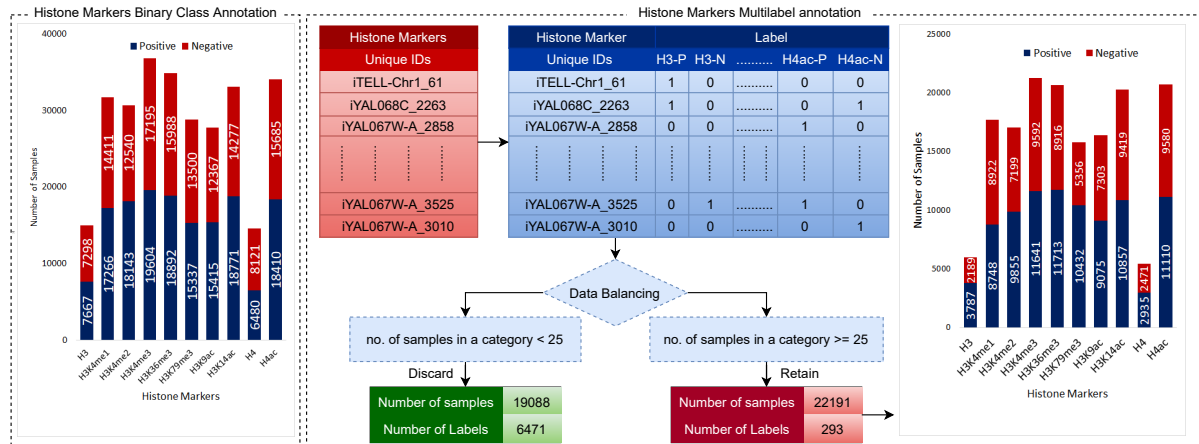


Figure 4.4: Development workflow of imbalanced and balanced multi-label classification datasets for histone occupancy and modification prediction

other 10 labels show the association of sequence with negative class distribution of 10 benchmark histone markers. In this manner, each sequence target label is represented as a 20-dimensional binary vector where 1 is assigned for the association of sequence with positive class of particular histone marker and 0 is assigned for the association of sequence with negative class of particular histone marker. Considering, whether multi-label dataset is imbalanced or balanced largely impact the predictor performance, 2 different versions of multi-label histone analysis dataset is prepared to assess the behavior of Histone-Net on account of imbalance and balance dataset.

A balanced version of multi-label dataset is prepared by eliminating all label cardinalities which have less than 25 sequences, whereas imbalanced version of dataset is obtained by retaining all label cardinalities. In this manner, balanced multi-label dataset of 19,088 and imbalanced dataset of 22,191 sequences are obtained.

4.2.2.3 Enhancer Identification and Strength Prediction Benchmark Datasets

To assess the performance of enhancer identification and strength predictors, researchers [172, 209, 266, 268, 379] have utilized two benchmark core datasets and independent test sets. These datasets are comprised of chromatin states of nine distinct cells involving GM12878, HepG2, H1ES, K562, HUVEC, HSMM, NHLF, HMEC and NHEK [266] where each enhancer sequence has same 200 bPs size. Statistics of both core datasets and independent test sets are shown in Table 4.2.

Table 4.2: Statistics of enhancer identification and strength prediction datasets

Class	enhancer	non-enhancer	weak enhancer	strong enhancer
Core Data set	1484	1484	742	742
Independent Test Set	200	200	100	100

4.3 Evaluation Criteria

In binary classification paradigm for histone sequence analysis, following Yin et al. [443] criteria, 80% of sequences of each dataset are used for training and 10% of sequences are used as validation set to find best parameters of the model. Whereas, 10% sequences are used to evaluate the performance of Histone-Net [22] predictor. In multi-label classification paradigm, we utilize 70% data for training, 10% data for validation and 20% data for testing. Similarly, in enhancer identification and strength prediction, following evaluation criteria of existing predictors [266, 268, 379], we perform 5-fold cross-validation over core datasets and independent test sets based evaluation. In binary classification, Histone-Net [22] and Enhancer-DSNet predictors make use of categorical cross-entropy loss and softmax classifier. In multi-label classification Histone-Net uses binary cross-entropy loss function and sigmoid classifier. Furthermore, in both classification scenarios, predictors are trained using an Adam optimizer with a decay rate of 0.95, epsilon of 1e-08, learning rate of 0.008 and 64 batch size.

4.4 Results and Discussions

This section illustrates performance of proposed Histone-Net [22] approach at different size k-mers using 10 different datasets for histone occupancy, methylation and acetylation prediction tasks. It compares the performance of proposed approach with adapted DeepHistone approach [444] and state-of-the-art image representation based predictor ‘HCNN’ [443]. Furthermore, it describes the generalization potential of proposed Histone-Net predictor by performing cross-domain evaluation. To evaluate the aptitude of Histone-Net predictor for simultaneously identifying histone marker, occupancy, acetylation and methylation areas in histone sequences, performance of Histone-Net predictor is analyzed in the paradigm of multi-label classification. Finally, it summarizes the performance statistics of the proposed Enhancer-DSNet and existing approaches for the task of identifying enhancers and predicting their strength.

4.4.1 Evaluation of Histone-Net in Intra-Domain Setting using Binary Classification Paradigm

Table 4.3 reveals the performance figures produced by proposed predictor at different k-mers in terms of accuracy. As clearly inferred by Table 4.3, the idea of using a traditional softmax classification layer to fuse label information into sequence vectors and perform sequence classification proves extremely effective.

Among different k-mers, at 9-mers and 10-mers, proposed Histone-Net predictor manages to achieve the top performance of 89% and 90% over two datasets of histone occupancy (H4, H3), respectively. On the other hand, at 11-mers it proves versatile enough for overall methylation and acetylation prediction datasets by achieving the performance figures of around 90%.

Table 4.3: Performance statistics of proposed Histone-Net [22] predictor over 10 benchmark datasets using 5 different k-mers

K-mers	Occupancy		Methylation					Acetylation		
	H3	H4	H3K4me1	H3K4me2	H3K4me3	H3K79me3	H3K36me3	H4ac	H3K14ac	H3K9ac
7	0.863	0.873	0.685	0.686	0.673	0.809	0.747	0.706	0.721	0.746
8	0.885	0.884	0.730	0.741	0.753	0.838	0.792	0.776	0.787	0.778
9	0.895	0.891	0.789	0.797	0.833	0.875	0.845	0.849	0.855	0.818
10	0.899	0.886	0.820	0.828	0.873	0.899	0.874	0.876	0.887	0.838
11	0.878	0.872	0.825	0.831	0.873	0.902	0.876	0.875	0.896	0.836
12	0.857	0.861	0.815	0.823	0.867	0.891	0.871	0.868	0.882	0.822

4.4.2 Performance Comparison of Histone-Net Predictor with Adapted and State-of-the-art Histone Occupancy and Modification Predictors

We perform a fair performance comparison of proposed Histone-Net predictor with image representation based on state-of-the-art histone occupancy and modification predictor namely "HCNN" [443] and adapted convolutional neural network based approach DeepHistone [444].

Table 4.4: Accuracy comparison of proposed Histone-Net [22] approach with state-of-the-art HCNN [443] and adapted DeepHistone [444] approach. Accuracy values of DeepHistone are obtained by processing raw histone sequences of various histone markers using convolutional neural network model presented by the authors [444] and accuracy values of HCNN are taken from Table 4.3 of Yin et al. [443] study.

Method	Occupancy		Methylation					Acetylation		
	H3	H4	H3k4me1	H3k4me2	H3k4me3	H3k79me3	H3k36me3	H4ac	H3K14ac	H3K9ac
HCNN [443]	0.8734	0.8733	0.7321	0.7427	0.7445	0.8163	0.7703	0.867	0.7479	0.7919
Deep-Histone [444]	0.8697	0.8979	0.6944	0.6496	0.6533	0.811	0.7609	0.7152	0.7334	0.7433
Proposed Histone-Net [22]	0.8951	0.8911	0.8251	0.8312	0.8726	0.9022	0.8756	0.8747	0.8962	0.8384

Table 4.4 reports the performance of Histone-Net predictor, state-of-the-art HCNN [443] approach and adapted DeepHistone [444] approach over 10 different histone occupancy, methylation and acetylation prediction datasets in terms of accuracy. As illustrated in Table 4.4, for both histone occupancy prediction datasets (H3, H4), on average, HCNN achieves the performance figures of around 87%. For most methylation prediction datasets, HCNN average performance falls around 75% except H3k79me3 dataset where it achieves 80% performance. Similarly, for acetylation prediction datasets, it manages to mark the performance of nearly 80%.

To perform a rich performance assessment of proposed Histone-Net predictor, we adopt a convolutional neural network based approach DeepHistone proposed by Yin et al. [444]. As shown in Table 4.4, DeepHistone only manages to achieve over 85% accuracy on 2 histone occupancy prediction datasets, over 80% accuracy on only 1 histone methylation prediction dataset (H3K79me3) from 5 histone methylation prediction datasets and over 70% accuracy on 3 histone acetylation prediction datasets. The reasons behind the limited performance of DeepHistone [444] in comparison to state-of-the-art HCNN [443] is the use of suboptimal statistical representation learning scheme which lacks to capture translational invariance of residues.

Proposed Histone-Net predictor produces superior performance than HCNN [443] across 10

different benchmark datasets. While for histone occupancy and acetylation prediction datasets (H3, H4), on average, Histone-Net produces performance figures around 90% and 86%, respectively. Whereas, for most methylation prediction datasets, its performance crosses the landmark of 85% and on H3k79me3 dataset it manages to achieve the top performance figure of 90%. Likewise, Histone-Net significantly outperforms adapted DeepHistone [444] approach across all 10 benchmark histone markers datasets for 3 different histone sequence analysis tasks. For histone occupancy prediction, on average, Histone-Net achieves an increment of 3%, for histone methylation prediction, it attains an increment of 15% and for histone acetylation prediction, it achieves an increment of 14%.

4.4.3 Evaluation of Histone-Net in Cross-Domain Setting

In biomedical sequence analysis, generally, cross-domain evaluation is used to examine the practical significance of predictive approaches in terms of their ability to perform accurate predictions over new histone markers. In cross-domain setting, for histone occupancy sequence analysis task, Histone-Net predictor is trained over the sequences of different histone markers belonging to Histone occupancy and tested on one of the test sets of particular histone marker that was not added in training set. This process is repeated to ensure that Histone-Net predictor is evaluated on the test set of each histone marker belonging to histone occupancy. A similar process is repeated for histone acetylation and methylation prediction tasks to ensure that Histone-Net is not biased towards specific histone marker data. In this manner, cross-domain performance of Histone-Net predictor over test sets of 10 benchmark datasets belonging to 3 distinct histone sequence analysis tasks is computed.

Table 4.5: Performance of proposed Histone-Net [22] predictor in cross-domain setting using different degree higher order residue based sequence representation.

Histone Marker Test Set	K-mers	Accuracy	Precision	Recall	F1-score
H3	7	0.7112	0.7214	0.7112	0.7069
H4	7	0.6753	0.6894	0.6753	0.6742
H3K14ac	10	0.8493	0.8461	0.8493	0.8495
H3K9ac	10	0.8297	0.8321	0.8297	0.8285
H4ac	11	0.8618	0.8625	0.8618	0.8615
H3K4me1	7	0.5792	0.5724	0.5792	0.565
H3K4me2	7	0.5913	0.5717	0.5913	0.5869
H3K4me3	7	0.4511	0.4433	0.4511	0.4471
H3K79me3	7	0.7153	0.7146	0.7153	0.7146
H3K36me3	7	0.6127	0.6087	0.6127	0.6103

Like intra-domain setting, in cross-domain setting, performance of Histone-Net predictor is assessed using 5 different higher order residues (7-to-11). Table 4.5 summarizes the peak performance achieved by Histone-Net predictor under different higher order residues over the test sets of 10 different benchmark datasets belonging to histone occupancy, acetylation and methylation prediction. As indicated in Table 4.5, just like intra-domain setting, Histone-Net

achieves top performance of around 86% in terms of 4 different evaluation metrics using upper degree higher order residues (11-mers) based sequence representations in cross-domain setting for the task of histone acetylation prediction. Whereas, for 2 other histone sequence analysis tasks including Histone Occupancy and Histone Methylation prediction, unlike intra-domain setting, here Histone-Net marks better performance with medium degree higher order residue (7-mers) based sequence representations. For histone occupancy, Histone-Net achieves best performance of 71% on test set of H3 histone marker as compared to H4 across all 4 evaluation metrics. For histone acetylation prediction, Histone-Net achieves better performance of 86% on test set of H4ac followed by H3K14ac and H3K9ac. Whereas, for histone methylation prediction, Histone-Net attains the best performance of 72% on the test set of H3K79me3 dataset.

Empirical evaluation on the test sets of 10 benchmark datasets belonging to 3 distinct histone sequence analysis tasks indicates that Histone-Net manages to attain the average performance of more than 80% for histone acetylation (H3K14ac, H3K9ac, H4ac), 70% for histone occupancy and 60% for histone methylation prediction. Across 10 benchmark datasets, compared to average performance of 87% and peak performance of 90% achieved by Histone-Net in intra-domain setting, Histone-Net manages to attain an average performance of 70% with the peak performance of 86% in cross-domain setting. Usually, the performance of computational approaches drops up to great extent when evaluated using cross-domain paradigm, however, Histone-Net predictor shows decent generalization potential across a variety of datasets belonging to 3 distinct histone sequence analysis tasks.

4.4.4 Intrinsic Evaluation of Histone-Net Predictor

Intrinsic evaluation assesses the quality of internal sequence representations of proposed Histone-Net predictor with an aim to evaluate the clusters of both classes in terms of cohesiveness and coupling levels using Principal Component Analysis (PCA) and T-distributed Stochastic Neighbor Embedding (t-SNE).

In particular, to provide a bird's eye view of discriminative potential of learned statistical vectors in positive and negative classes, we randomly select one dataset from each histone sequence analysis task. By training the Histone-Net predictor on training sets of all three datasets, we extract internal representations of all sequences belonging to test sets of all three datasets.

Extracted vectors are passed to PCA approach that reduces the dimensions from 100 to 25. These 25-dimensional statistical vectors are passed to T-SNE visualizer that further reduces the dimensions and creates mappings in two-dimensional space. T-SNE graphs for 1 randomly selected histone occupancy, methylation and acetylation dataset using 11-mer are shown here.

As depicted by the embedding chart 4.5, clusters for both positive and negative class for all three selected datasets are far less overlapping. Also, it is quite evident that Histone-Net predictor better captures the local and global semantic composition of k-mers which eventually

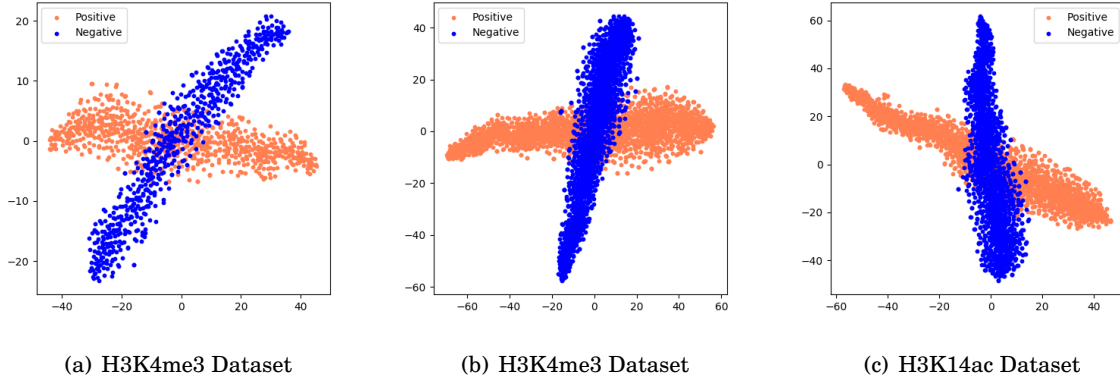


Figure 4.5: Internal representation of Histone-Net predictor [22] for three distinct datasets at 11-mers

assists to develop optimal sequence clusters.

4.4.5 Evaluation of Histone-Net in Multi-label Classification Paradigm

This section briefly describes the performance of the Histone-Net multi-label classification paradigm using balanced and imbalanced versions of multi-label datasets in terms of 11 different evaluation metrics. To better quantify the effectiveness of the Histone-Net multi-label classification paradigm, it compares the performance of proposed Histone-Net predictor with an adapted convolutional neural network based DeepHistone approach [444].

Table 4.6 reports the performance produced by proposed Histone-Net predictor over imbalanced and balanced versions of multi-label histone occupancy and modification prediction datasets in terms of 9 distinct multi-label evaluation metrics. Table 4.6 illustrates that across both versions of multi-label datasets, Histone-Net performance almost gradually improves by increasing the degree of higher order residues, indicating Histone-Net achieves the best performance with upper degree higher order residues (10-mers, 11-mers) across all evaluation metrics. Furthermore, across all different higher-order residues, Histone-Net achieves slightly better performance on balanced versions of multi-label dataset as compared to imbalanced version in terms of the most evaluation metrics. On balanced version of multi-label dataset, Histone-Net achieves the F1-score of 72% and hamming loss of 0.20 which surpasses the Histone-Net performance achieved on imbalanced version by the figure of 3% and 1%.

Table 4.7 compares the performance of Histone-Net predictor with adapted DeepHistone approach using imbalanced and balanced versions of multi-label histone occupancy and modification prediction dataset. It is evident from the Table 4.7 that Histone-Net significantly outperforms adapted DeepHistone predictor across all 11 evaluation metrics. On imbalanced version of multi-label histone occupancy and modification prediction dataset, Histone-Net achieves the accuracy increment of 14%, precision increment of 8%, recall increment of 21%, F1-score increment of

Table 4.6: Performance statistics of Histone-Net predictor [22] using different size k-mers over imbalanced and balanced version of multilabel dataset

Multi-Label DNA Sequence Analysis Datasets	Performance Measures	K-mers					
		7	8	9	10	11	12
Imbalance Dataset	Accuracy	0.475	0.5129	0.5535	0.5714	0.57	0.559
	F1	0.6146	0.6451	0.6787	0.693	0.6921	0.684
	Average Precision	0.529	0.529	0.529	0.529	0.529	0.529
Balanced Dataset	Accuracy	0.5486	0.5783	0.6099	0.6171	0.6078	0.6051
	F1	0.676	0.6973	0.7217	0.7299	0.7187	0.709
	Average Precision	0.5712	0.5712	0.5712	0.5712	0.5712	0.5712

15%, average precision increment of 13%, AUPRC increment of 18%, AUROC increment of 12%, hamming loss improvement of 7% and coverage improvement of 2%. On the balanced version of multi-label histone occupancy and modification prediction dataset, Histone-Net achieves the increment of 10%, 4%, 15%, 10%, 15%, 17%, 10%, 5% and 2% in terms of aforementioned distinct evaluation metrics. On average, Histone-Net supersedes the performance of adapted DeepHistone by the figure 10% and 8% on imbalanced and balanced version of multi-label histone occupancy and modification prediction datasets, respectively.

Table 4.7: Performance statistics of proposed Histone-Net [22] and adapted DeepHistone predictors using optimal size k-mer, over imbalanced and balanced versions of multilabel dataset in terms of 11 distinct evaluation metrics

Datasets	Dataset	Accuracy	Precision	Recall	F1	Average Precision	AUPRC	AUROC	Ranking Loss	OneError	Hamming Loss	Coverage
Imbalanced Dataset	Histone-net [22]	0.5714	0.6882	0.7257	0.693	0.529	0.788	0.868	0.3812	0.8556	0.2049	15.8996
	DeepHistone [444]	0.4269	0.6127	0.5245	0.5467	0.4014	0.6114	0.7502	0.2295	0.8143	0.2717	17.7378
Balanced Dataset	Histone-net [22]	0.6171	0.7113	0.7652	0.7299	0.5712	0.8244	0.8821	0.3412	0.8324	0.1977	15.7452
	DeepHistone [444]	0.5235	0.6693	0.6164	0.6296	0.4088	0.6464	0.7842	0.1958	0.8297	0.2439	18.082

Furthermore, in order to analyze the effectiveness of proposed Histone-Net predictor for accurately predicting histone occupancy and modifications, we utilize one-versus-all strategy to generate 20 binary confusion matrices for 10 histone markers for imbalanced (Figure 4.6) and balanced versions of multi-label datasets (Figure 4.6). In one-versus-all strategy, false positives, false negatives, true negatives and true positives are computed by treating one particular histone marker class as positive and all other histone markers classes belonging to same histone sequence analysis task as negative irrespective of the multi-label problem. More specifically, we evaluate the behavior of Histone-Net when there is a decent gap between the total number of positive and negative sequences.

A critical analysis of 20 confusion matrices (Figure 4.6) produced by Histone-Net over imbalanced dataset shows that overall 64% positive histone marker appearances and 82% negative histone marker appearances (represented as rest) are correctly predicted by Histone-Net. Top true positive figure of 89% is achieved on H3 histone marker, whereas, top true negative figure of 94% is achieved on H3ac histone marker. Among histone markers related to occupancy, a higher

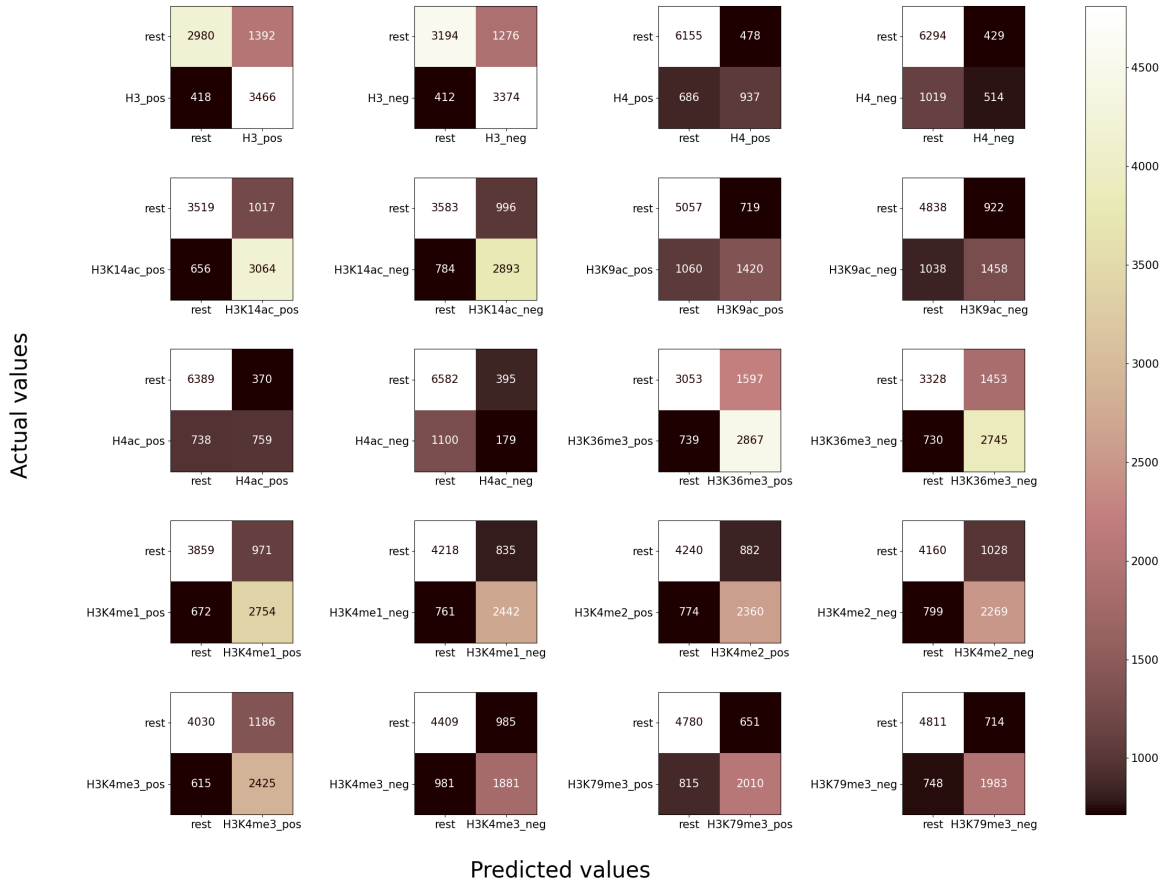


Figure 4.6: Confusion metrics of Histone-Net [22] predictor for unbalanced version of multi-label classification dataset, where each confusion matrix illustrates correct and wrong predictions of a particular class.

number of positive and negative sequences are correctly classified on H3 makers as compared to H4. Among histone markers related to acetylation, most number of positive class sequences are accurately classified in H3K14ac histone marker with the performance around 82% whereas the most numbers of negative class sequences are correctly predicted in H34ac histone maker with the performance around 95%. Turning towards the performance of methylation related histone markers, 80% of positive class sequences are correctly classified in 3 histone markers (H3K36me3, H3K4me1, H3K4me3) whereas 88% of negative class sequences are correctly predicted by Histone-Net on H3K79me3 histone marker.

In one-versus-all setting as negative class gets more number of samples which is why usually there exist a huge gap between the performance of positive and negative class, however, here the

CHAPTER 4. HISTONE OCCUPANCY/MODIFICATIONS PREDICTION AND ENHANCER IDENTIFICATION/STRENGTH PREDICTION

gap is not large at all due to the robustness of Histone-Net predictor towards imbalance class distribution.

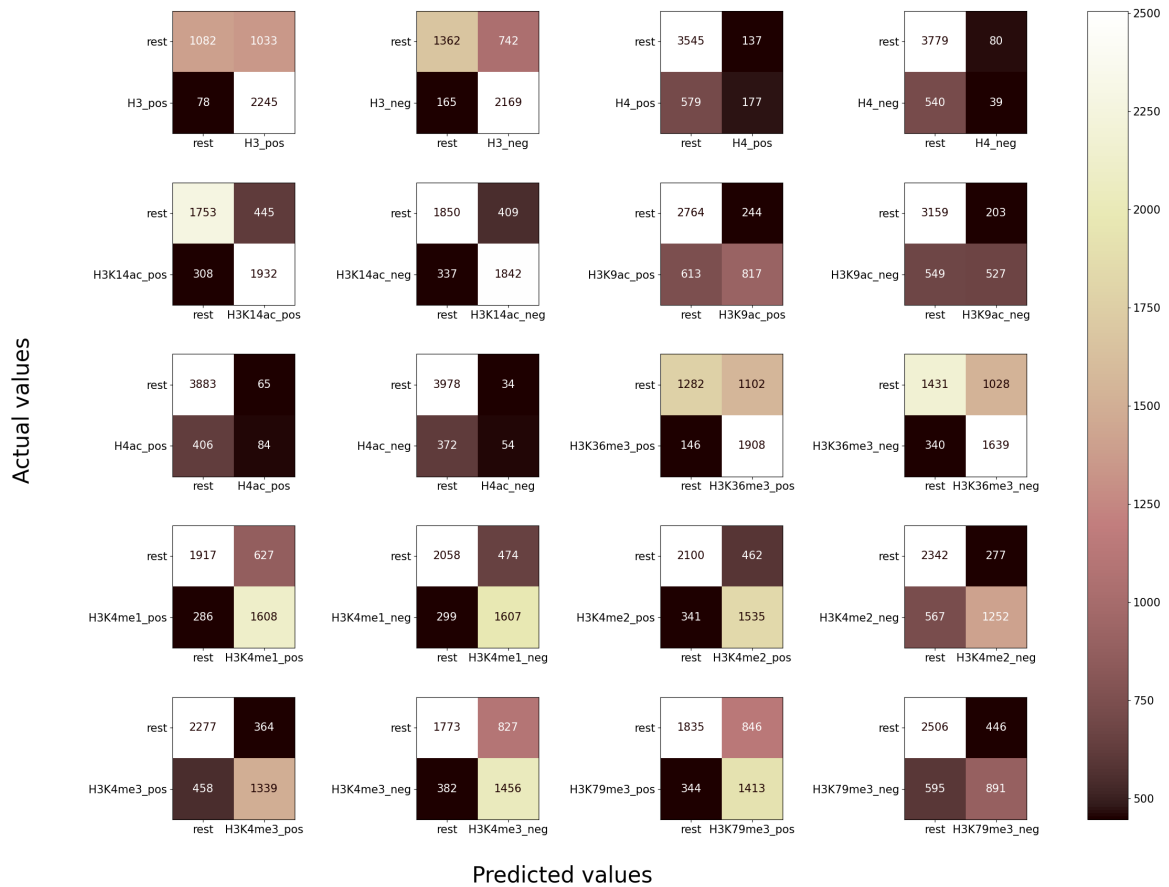


Figure 4.7: Confusion metrics of Histone-Net predictor [22] for balanced version of multi-label classification dataset, where each confusion matrix illustrates correct and wrong predictions of a particular class.

In imbalanced version of multi-label dataset, there exist only 5 uni-label and 24 bi-label sequences which are too less for effective model training. Considering uni-label and bi-label sequences act as a noise and derail the generalizability of classifier, we perform experimentation on balanced version of multi-label dataset prepared after eliminating all uni-label and bi-label sequences. Performance analysis on 20 confusion matrices produced by Histone-Net on balanced version of multi-label dataset reveals that overall 66% positive and 80% negative histone marker appearances are predicted accurately.

Across different histone markers, overall Histone-Net marks better performance on balanced

versions of multi-label datasets as compared to imbalanced version of multi-label datasets. Highest true positive figure of 97% and true negative figure of 99% is achieved on H3 and H4ac histone markers, respectively, achieving an increment of 8% and 5% as compared to the peak performance achieved by Histone-Net on imbalanced version of dataset. From histone markers related to occupancy, while the most positive sequences are correctly predicted in H3 histone marker, the higher number of negative sequences are correctly classified in H4 histone marker. Among histone markers related to acetylation, greater number of positive class sequences are accurately classified in H3K14ac histone marker with the performance of around 86% whereas the most number of negative class sequences are correctly predicted in H34ac histone marker with the performance of around 99%, outperforming the performance attained on imbalanced version by 4%. Concerning the performance of methylation related histone markers, 93% of positive class sequences are correctly classified in H3K36me3 histone marker whereas 89% of negative class sequences are correctly predicted by Histone-Net on H3K4me2 histone marker, achieving an increment of 13% and 1%, respectively when compared with top performance attained by Histone-Net on methylation histone marker of imbalanced dataset.

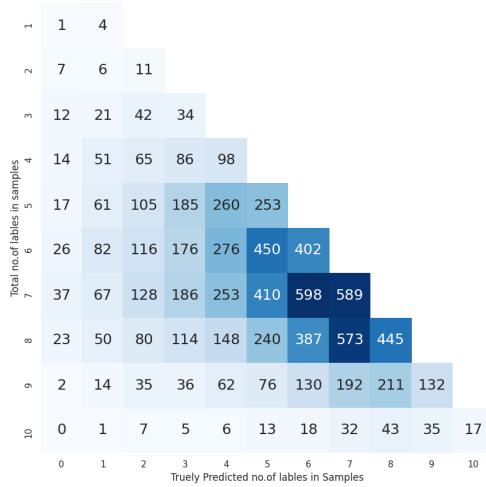
In a nutshell, across different evaluation metrics, although average performance figures attained by Histone-Net on imbalanced and balanced versions of multi-label dataset are comparable. However, a close look indicates that across most histone markers, Histone-Net achieves better performance on balanced version of multi-label dataset as compared to imbalanced version.

To identify up to what degree Histone-Net manages to simultaneously predict histone-occupancy, acetylation and methylation areas in unseen histone sequences, performance of Histone-Net is analyzed over imbalanced and balanced version of multi-label dataset in terms of multi-label confusion matrices corresponding to unique sequence-label distributions. In both versions of multi-label datasets, number of correctly predicted histone markers out of all actual histone markers are highlighted in confusion matrices (Figure 4.8).

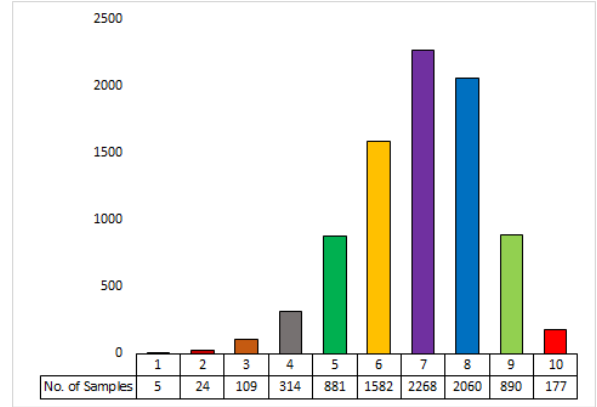
For imbalanced version of multi-label dataset, a closer look at sequence-label distribution (Figure 4.8) and confusion matrix (Figure 4.8) reveals that, Histone-Net manages to make accurate predictions for 90% uni-label sequences as only 1 sequence is misclassified out of 5 sequences. For bi-label sequences, it correctly predicts 46% sequences because 11 bi-label sequences are correctly classified out of 24 sequences. For tri-label sequences, Histone-Net performance drops as it only manages to identify the target histone markers of 31% sequences. For tetra-label sequences, Histone-Net achieves best performance of around 83% as it makes correct predictions for 260 sequences out of 314 sequences. However, afterward, with the increase of histone marker combinations, Histone-Net best performance of 83% keeps on declining with great margin, dropping to 51%, 25%, 26% and 22%, for penta, hexa, hepta and octa-label sequences, respectively, achieving lowest performance 15%, 9% on highest label cardinalities including nona-label and deca-label sequences.

On the other hand, for balanced version of multi-label dataset, uni and bi-label samples are

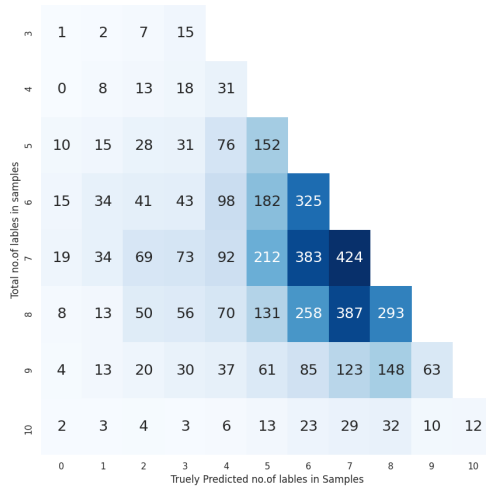
CHAPTER 4. HISTONE OCCUPANCY/MODIFICATIONS PREDICTION AND ENHANCER IDENTIFICATION/STRENGTH PREDICTION



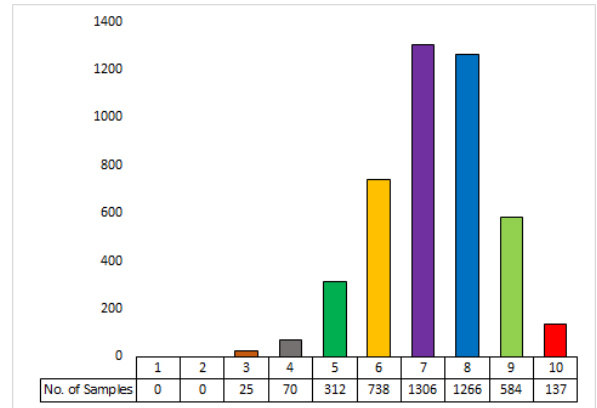
A. Confusion matrix for imbalanced dataset



B. Sequence-Label distribution for imbalanced dataset



C. Confusion matrix for balanced dataset



D. Sequence-Label distribution for balanced dataset

Figure 4.8: Performance analysis of proposed Histone-Net predictor [22] in terms of sequence label distributions using imbalance and balanced versions of multilabel classification datasets

removed from the dataset and label cardinalities which have 25 or more samples are kept. For tri-label and tetra-label sequences, Histone-Net manages to correctly predict 60% and 26% sequences, respectively. Afterward, with the increase of label cardinality, unlike imbalanced dataset, here Histone-Net performance decreases with less margin, it manages to correctly predict penta, hexa, hepta and octa-label sequences with 49%, 44%, 32% and 23% accuracy. However, for highest label

cardinalities like nona-label and deca-label sequences, Histone-Net once again only manages to correctly predict 11% and 9% sequences. Overall, Histone-Net achieves better performance on balanced version of multi-label dataset. For balanced version of dataset, Histone-Net achieves better performance with medium to higher level of histone marker combinations. Whereas, for imbalanced version, Histone-Net achieves better performance with low to medium level of histone marker combinations. However, for highest histone marker combinations (9 and 10), Histone-Net makes correct predictions for only 10% of sequences.

To summarize, a comprehensive evaluation of Histone-Net in multi-label sequence classification paradigm using imbalanced and balanced datasets proves the capability of Histone-Net for simultaneously predicting histone type, occupancy, acetylation and methylation areas in histone sequences. Furthermore, Histone-Net achieves decent performance on both imbalanced and balanced versions of multi-label datasets, showing its robustness to handle diverse data and sample-to-label distributions.

4.4.6 Performance Comparison of Proposed Enhancer-DSNet Approach with Existing Enhancer Identification and Strength Prediction Approaches

Here, we briefly describe and compare the performance of proposed Enhancer-DSNet approach with state-of-the-art Enhancer determinant and strength prediction approaches by performing cross-validation on core benchmark datasets and independent test sets based evaluation.

4.4.6.1 Evaluation on Benchmark Core Dataset

Following Tan et al. [379] work, in our experimentation, we used 5-fold cross-validation on benchmark core datasets.

Table 4.8 reports the average performance figures produced using 5-fold cross-validation at layers 1 and 2 in terms of accuracy, specificity, sensitivity and Matthews Correlation Coefficient (MCC). As indicated in Table 4.8, for enhancer/non-enhancer prediction task (layer-1), proposed Enhancer-DSNet approach outperforms Tan et al. Enhancer [379] approach by the sensitivity figure of 3% and MCC figure of 2%. However, for strong/weak enhancer prediction task (layer-2), proposed Enhancer-DSNet outperforms Tan et al. Enhancer approach [379] with a significant margin across 4 different evaluation metrics. Enhancer-DSNet significantly superior performance overshadows the most recent Tan et al. approach [379] performance by the figure of 17% in terms of sensitivity, 29% in terms of specificity, 4% in terms of accuracy and 6% in terms of MCC.

4.4.6.2 Performance Comparison of Proposed Enhancer-DSNet with Existing Predictors over Independent Test Set

Table 4.9 compares the performance of proposed Enhancer-DSNet and existing predictors over independent test sets for enhancer/non-enhancer and strong/weak enhancer prediction tasks

Table 4.8: 5-Fold cross-validation based performance comparison of proposed Enhancer-DSNet [20] and latest existing predictor [379] for enhancer/non-enhancer and strong/weak enhancer prediction.

Classifiers	Sensitivity	Specificity	Accuracy	MCC
1st Layer (Enhancer/Non-Enhancer)				
Proposed Enhancer-DSNet [20]	0.76	0.76	0.76	0.52
Tan et al. Enhancer [379]	0.73	0.76	0.74	0.50
2nd Layer (Strong Enhancer/Weak Enhancer)				
Proposed Enhancer-DSNet [20]	0.63	0.67	0.63	0.26
Tan et al. Enhancer [379]	0.80	0.38	0.59	0.20

in terms of accuracy, specificity, sensitivity and MCC. According to Table 4.9, at layer-1, among all existing predictors excluding the most recent Tan et al. approach [379], -iEnhancer-EL [268] marks better performance across most evaluation metrics. Here, proposed Enhancer-DSNet outperforms the most recent Tan et al. approach [379] by the figure of 2%, 1%, 2% and 5% in terms of sensitivity, specificity, accuracy and MCC and second best performing -iEnhancer-EL [268] by the figure of 7%, 3% and 6% in terms of sensitivity, accuracy and MCC, respectively. Whereas, at layer-2, once again proposed Enhancer-DSNet outshines the most recent Tan et al. approach [379] by the promising figure of 21% in terms of specificity, 15% in terms of accuracy and 39% in terms of MCC and second best performing predictor -iEnhancer-EL [268] by the figure of 29% in terms of sensitivity, 22% in terms of accuracy and 48% in terms of MCC.

Table 4.9: Performance comparison of proposed Enhancer-DSNet [20] with existing Enhancer/Non-Enhancer and Strong/Weak Enhancer predictors over independent test sets

Classifiers	Sensitivity	Specificity	Accuracy	MCC
1st Layer (Enhancer/Non-Enhancer)				
Proposed Enhancer-DSNet [20]	0.78	0.77	0.78	0.56
Tan et al. Enhancer [379]	0.76	0.76	0.76	0.51
iEnhancer-EL [268]	0.71	0.79	0.75	0.50
iEnhancer-2L [266]	0.71	0.75	0.73	0.46
EnhancerPred [209]	0.74	0.75	0.74	0.48
2nd Layer (Strong Enhancer/Weak Enhancer)				
Proposed Enhancer-DSNet [20]	0.83	0.67	0.83	0.70
Tan et al. Enhancer [379]	0.83	0.46	68.49	0.31
iEnhancer-EL [268]	0.54	0.68	0.61	0.22
iEnhancer-2L [266]	0.47	0.74	0.61	0.22
EnhancerPred [209]	0.45	0.65	0.55	0.10

4.5 Conclusion

Histone markers and Enhancers are considered core areas of epigenetic sequence analysis that pave way for the development of drugs and identification of diseases. To supplement epigenetic sequence analysis, contributions of this chapter are manifold: it presents a generic classifier that can accurately predict occupancy, acetylation and methylation levels in histone markers. It provides a unique dataset for Histone markers sequence analysis. This dataset will facilitate

researchers to develop and evaluate novel methods that can simultaneously predict Histone markers type, occupancy, acetylation and methylation levels. Furthermore, proposed classifier managed to outperform existing histone sequence analysis predictors over 10 public benchmark datasets. Experimental results in cross-domain setting reveal proposed classifier potential to perform accurate analysis over unseen histone markers. While, simultaneously predicting histone type, occupancy, acetylation and methylation levels, it also produced comprehensive performance over newly developed dataset. Furthermore, effectiveness of proposed classifier over state-of-the-art enhancer identification and strength predictors, proves its practical significance and usability for other Genomics sequence analysis tasks.

SMALL NON-CODING RNA CLASSIFICATION

Ribonucleic acid (RNA) is an essential molecule for living entities which is involved in multifarious biological processes, such as translation, sponging, gene regulation and splicing [180, 292]. Four basic nucleotides, namely guanine (G), uracil (U), adenine (A) and cytosine (C) define the basic structure of RNA molecules [345], where structure means to have knowledge about its biological properties. Involvement of RNA molecules in different biological functions and their importance in different diseases attracts many researchers to analyze RNA molecules in more detail to find their new functions and roles in biological processes [339, 341, 438]. Primarily, RNA molecules are categorized into coding and non-coding RNA classes, where about 3% of total RNA is coding that produce proteins (so called messenger RNA = mRNA) and the remaining 97% is known as non-coding (ncRNA) or functional RNA [198]. While the function of mRNAs is well-known and has been studied extensively, non-coding RNAs were considered junk code and thought not to participate in the process of developing proteins [324, 412, 438]. At the beginning of 21st century, analysis of mouse [96] and human [239] Genomes and later in 2005 findings of human Genome project revealed that majority of ncRNAs are involved in many essential biological processes such as dosage compensation, genomic imprinting and cell differentiation [13, 129]. After these findings, in-depth analysis of ncRNAs became even more interesting because of their importance in understanding the phenomena behind human health and diseases [13].

Most recent literature reveals that ncRNAs not only participate in the development of proteins, but also control the process in which proteins are produced. They act as key players in the development and progression of complex diseases [182] and are involved in several

⁰This chapter is an adapted version of the work presented in Asim et al., "A Robust and Precise ConvNet for small non-coding RNA classification (RPC-snRC)." , In *IEEE Access* 9 (2020) [28] and Asim et al., "Advances in Computational Methodologies for Classification and Subcellular Locality Prediction of Non-Coding RNAs", In *International Journal of Molecular Sciences* (2021) [19]

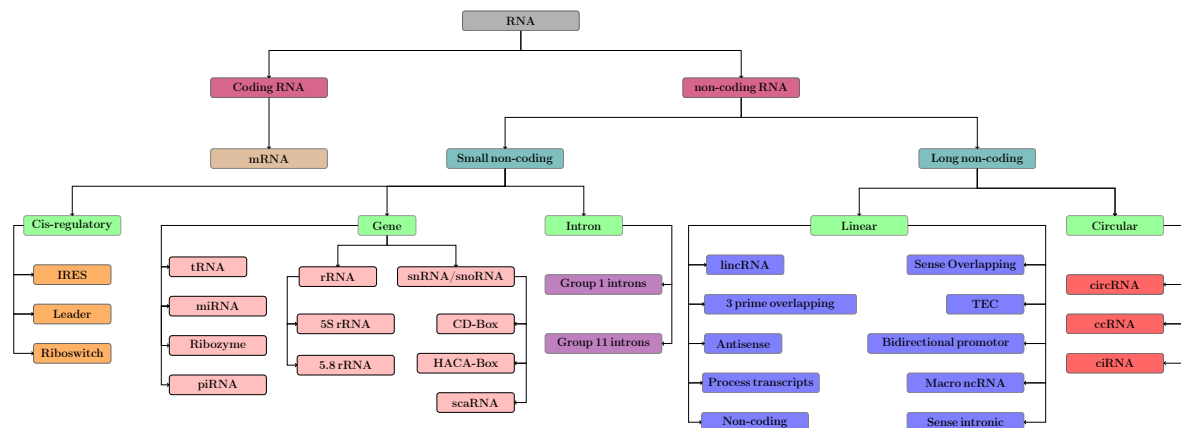


Figure 5.1: A comprehensive taxonomy of RNA families

biological and physiological processes [182] such as gene expression regulation [302] and RNA maturation [232]. Contribution of ncRNAs in vital oncogenic processes such as differentiation, proliferation, migration, angiogenesis and apoptosis has gained much attention as potential diagnostic and prognostic biomarkers in leukemia [44]. Furthermore, ncRNAs are discovered as tumor suppressors, oncogenic drivers in different cancer types [325] and are strongly linked to the development of Alzheimer's and cardiovascular diseases [124, 363].

Based on cellular functionality, variation in sequence length, unique structure, physical and chemical properties [13], ncRNAs can be segregated into different subclasses, a taxonomy of which is depicted in Figure 5.1. ncRNAs are categorized into small non-coding RNAs (sncRNA) and long non-coding RNAs (lncRNA). The lncRNAs are further categorized into linear and circular RNAs. Linear RNAs play diverse roles in intracellular processes such as gene transcription and translation [452]. Circular RNAs are involved in gene regulation, where irregularities cause complex diseases like lung cancer and tumor [67, 285].

Primarily, small ncRNAs are classified into 13 subclasses where each subclass has distinct medical and biological significance. For instance, scaRNAs, most of which are functionally and structurally identical to snoRNAs can guide modifications in pseudo uridylation and methylation. miRNAs are involved in various complex human diseases such as cancer, autoimmune, cardiovascular and neurodegenerative diseases [124]. Similarly, Ribosomal RNA (rRNA) plays an essential role in protein synthesis and its characteristics are considered very valuable for the development of antibiotics. 5.8S ribosomal RNAs actively participate in protein translation [122] and facilitate to understand other rRNA pathways and processes in the cell [15]. Although the complete functionalities of 5 S ribosomal RNA have not been discovered yet, it has been shown that its deletion substantially reduces protein synthesis that creates harmful effects on cell fitness [311].

Accurate discrimination of ncRNAs from coding RNAs and identification of their subtypes can lay the foundation for demystifying the core functions and biological roles of different subclasses of

ncRNAs, their involvement to suppress the mechanism [253] underlying complex human diseases [31, 179] or to develop effective treatments and optimize therapeutics [291, 409]. Classification of small non-coding RNAs (sncRNAs) is of high importance due to unique role of each subclass into molecular processes and in the development of diseases. It can support biologists and clinicians to get a better understanding for the role of sncRNAs in biological processes such as classification of sncRNAs is important in developing procedures for cancer therapeutics [13].

5.1 Related Work

The interest to develop sophisticated computational methods for ncRNA classification has rocketed over the period since knowing the family of ncRNA is substantial for drug targeting and understanding growth of various complex diseases. Non-coding RNA classification is a vast domain where classification at different levels of ncRNA (shown in figure 5.1) has been performed. Mainly, researchers have been focusing on 1) distinguish non-coding RNA from coding RNA, 2) categorize ncRNA into long and small non-coding RNA, 3) segregate long non-coding RNA into its subtypes such as circular RNA and 4) classify small non-coding RNA into its 13 subclasses. Classification of ncRNAs at each level facilitates distinct biological advantages.

To date, several computational approaches have been proposed for non-coding RNA classification at different stages which are comprehensively summarized in our paper [19]. As compared to other types of ncRNA classification, small non-coding RNA classification lacks AI based approaches.

Antonino Fiannaca et al. [136] proposed first computational predictor named nRC that can classify small non-coding RNAs using only sequence information. nRC extracts secondary structures of RNA sequences and feeds them to Convolutional Neural Network (CNN) that discriminates small non-coding RNAs into 13 subclasses. More recently, Emanuele Rossi [345] proposed another predictor that also extracts secondary structural features and uses Graph convolutional neural network for further feature extraction and classification.

As described above, both small non-coding RNA classification approaches use secondary structure of RNA sequences as input and extract discriminative features by utilizing convolution layers or graph based methodologies. Secondary structure extraction methods only consider global characteristics of nucleic acids and ignore their local characteristics [142]. Furthermore, transformation of raw RNA sequences to secondary structural features creates high-dimensional feature space which is computationally inefficient [142].

Instead of extracting secondary structures of RNA sequences, we explore the potential of three different RNA discretization strategies namely: one-hot vector encoding, random and pretrained embeddings. Following the success of DenseNet architecture for diverse types of classification tasks in computer vision domain [229, 365]. We propose a robust and precise classifier based on DenseNet architecture where key idea is to provide a proper gradient flow

path among CNN layers which could learn more discriminative features. However, while using deep learning based classifier, it is unexplored whether such architectures performs better with one-hot vector encoding or random embeddings or pretrained embeddings. Furthermore, in all three types of representations, we examine whether classifier performs better while training character level feature of RNA sequence or k-mer level feature. To answer above questions, we have performed detailed experimentation on small non-coding RNA classification dataset with the proposed RPC-snRC classifier using all three types of representations at character level and k-mer level. Moreover, to further analyze the idea of utilizing primary RNA sequences, we performed experiments with two adapted ResNet architectures which vary in terms of depth and hyper-parameters.

5.2 Materials and Methods

This section illustrates the proposed RPC-snRC predictor and details of ResNet-based architectures along with benchmark dataset.

5.2.1 Proposed RPC-snRC Methodology

This section briefly describes the proposed RPC-snRC methodology for the classification of small non-coding RNA. We develop a deep classifier in which a phenomenon similar to DenseNet is used to enable proper flow of gradient between the layers. RPC-snRC utilizes a set of convolutional layers for extraction of discriminative features from the primary sequences of small non-coding RNA. Discriminative features are then fed to dense layers for classification of sequences into a set of predefined classes.

Figure 5.2 illustrates the architecture of the proposed methodology along with noteworthy model parameters. The proposed RPC-snRC methodology is based on three dense modules. Each dense module contains the same number of layers; however, output units get doubled in each successive dense module. Each dense module first performs batch normalization on the given input and then applies ReLu activation to introduce nonlinearity followed by convolution operation to extract discriminative features. Finally, it repeats the discussed operations one more time in order to better learn hierarchical representation of data. Each dense module is followed by a transition layer that performs batch normalization, ReLu activation, convolution with the filter size 1×1 and max pooling with the size of 4 to retain discriminative features and discard useless ones. Dense architecture was proposed by Gao Huang et al. [191] and has been widely utilized for various applications of computer vision. We utilize this architecture for sequence data which is one-dimensional and entirely different from visual data. Integral components of the proposed methodology such as DenseNet, Dense connectivity, Composite function, Pooling layers, Growth rate and Bottleneck layers which are adapted to cope one-dimensional data, are discussed below.

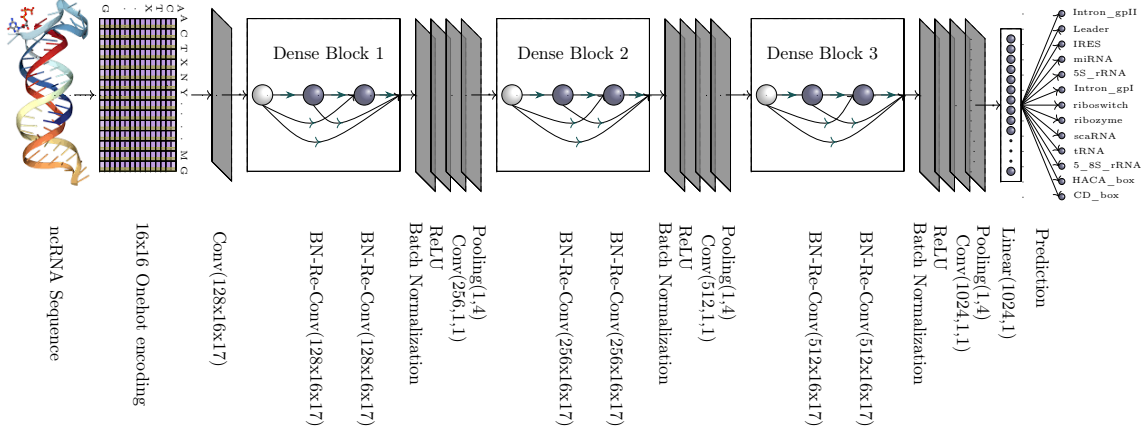


Figure 5.2: Proposed RPC-snRC methodology for small non-coding RNA classification. In figure, (128,16,18) indicates there are 128 kernels, each of width 16 and length 18 in a convolutional layer and (1,4) indicates kernel width and length are set to 1 and 4, respectively in a pooling layer. Remaining layers of network also follow same dimensionality pattern.

5.2.1.1 DenseNet

Consider a small non-coding RNA sample S_0 that is passed through a convolutional network. The network consists of L layers, each of which performs a non-linear conversion $H_L(\cdot)$, where L indicates the layer. $H_L(\cdot)$ may be a composite function for operations like batch normalization [202], rectified linear units (RELU) [150], Pooling [244], or Convolution (Conv). We refer to the L^{th} layer output as x_L .

Dense connectivity: State-of-the-art feed-forward convolutional networks attach the L^{th} layer output as an input to the $(L+1)^{th}$ layer, which produces the following transition layer $x_L = H_L(x_{L-1})$ [229]. ResNets [171] along with skip connection strategy use an identity function to bypass nonlinear transformations shown in equation 5.1

$$X_L = H_L(X_{L-1}) + x_{L-1} \quad (5.1)$$

ResNets benefit is that the gradient can flow straight from subsequent layers to previous layers through the identity function. However, the identity function and output of H_L are mixed by summation which can hinder the flow of data in the network.

We utilize Densenet a distinct connectivity model to further enhance the information flow between layers. In this model L^{th} layer gets all previous layer's feature maps, $x_0, \dots; x_{L-1}$, as

input.

$$X_L = H_L([x_0, x_1, \dots; x_{L-1}]) \quad (5.2)$$

In equation 5.2, $x_0, \dots; x_{L-1}$ relates to the concatenation of the feature maps in the $0, \dots, L-1$ layers

Composite function: Following He et al. [171], we define $H_L(\cdot)$ as a composite function of three successive operations: Batch Normalization (BN) [202], Activation function named as rectified linear unit (ReLU) [150] and a convolution (Conv) layer.

Transition layers: We refer to the layers between blocks that perform convolution and pooling operations as transition layers. The procedure of concatenation used in equation 5.2 is not applicable if size of feature maps is variable. In our architecture, we split the network into various tightly linked dense blocks to generate feature maps of same size. Down sampling is performed through transition layers which consist of a batch normalization layer and a convolution layer of kernel size 1, followed by an average pooling layer of kernel size 4.

Table 5.1: Architecture summary of Res18-nRC and Res50-nRC

Layer_Name	Res18_nRC		Res50_nRC	
	Output Size	Parameters detail	Parameters detail	Output Size
Conv-1	64×1182	$(64, 3), s=1, p=1$		64×1182
Conv-2	64×1182	$\begin{bmatrix} (64, 17) \\ (64, 17) \end{bmatrix} \times 2, p=8$	$\begin{bmatrix} (64, 1) , p=0 \\ (64, 17) , p=8 \\ (256, 1) , p=0 \end{bmatrix} \times 3$	256×1182
Pool-1	64×591	$(2, 2)$		256×591
Conv-3	128×296	$\begin{bmatrix} (128, 17) \\ (128, 17) \end{bmatrix} \times 2, s=2, p=8$	$\begin{bmatrix} (128, 1) , p=0 \\ (128, 17) , p=8 \\ (512, 1) , p=0 \end{bmatrix} \times 4, s=2$	512×296
Pool-2	128×148	$(2, 2)$		512×148
Conv-4	256×74	$\begin{bmatrix} (256, 17) \\ (256, 17) \end{bmatrix} \times 2, s=2, p=8$	$\begin{bmatrix} (256, 1) , p=0 \\ (256, 17) , p=8 \\ (1024, 1) , p=0 \end{bmatrix} \times 6, s=2$	1024×74
Pool-3	256×37	$(2, 2)$		1024×37
Conv-5	512×19	$\begin{bmatrix} (512, 17) \\ (512, 17) \end{bmatrix} \times 2, s=2, p=8$	$\begin{bmatrix} (512, 1) , p=0 \\ (512, 17) , p=8 \\ (2048, 1) , p=0 \end{bmatrix} \times 3, s=2$	2048×19
Pool-4	512×9	$(2, 2)$		2048×9
Output	13	Flatten-4608	Flatten-18432	13

Growth rate: If each composite function $H_L(\cdot)$ produces N feature maps, then L^{th} layer will have $N_0 + N \times (L - 1)$ input feature-maps, where N_0 denotes number of channels in the input layer. We refer to the N hyperparameter as the network’s growth rate.

5.2.1.2 Adapted ResNet Architectures

To make sure whether, deeper classifiers with proper gradient flow among layers can produce better performance, we adapted two predefined ResNet architectures Res18-nRC and Res50-nRC from the domain of computer vision. Table 5.1 illustrates parameter details of the adapted ResNet architectures. In both architectures, ncRNA samples are passed through convolutional layers before feeding to ResNet modules. Both architectures have 4 ResNet modules, while each module of Res18-nRC has 2 basic blocks, where each basic block has two convolutional layers but Res50-nRC architecture has variable bottleneck blocks in each ResNet module which are mentioned by a number outside the matrix brackets, i.e., first ResNet module has 3 bottleneck blocks and second has 4. In the first matrix (64,17) 64 represents number of feature maps and 17 shows the kernel size.

5.2.2 Benchmark Dataset

We perform experimentation on a small non-coding RNA classification dataset provided by Antonino et al. [136]. This is the only benchmark dataset which is publicly available. It contains 8920 samples that belong to 13 different ncRNA classes and each class has 700 samples except IRES class which contains 520 samples. Table 5.2 illustrates samples distribution and sequence length variation in each class.

Table 5.2: Characteristics of non-coding RNA classification dataset, where Max-seq length and Min-seq length illustrate maximum and minimum length of nucleotides in each class.

Classes	No.of Samples	Max-seq length	Min-seq length
IRES	520	630	53
Intron_gpI	700	1182	133
leader	700	237	38
scaRNA	700	445	78
S5_rRNA	700	199	61
miRNA	700	631	52
tRNA	700	177	47
riboswitch	700	399	44
ribozyme	700	1136	41
S8_rRNA	700	290	50
CD-box	700	404	54
HACA-box	700	508	59
Intron_gpII	700	241	48

The dataset has benchmark defined split with 6320 training and 2600 test samples. In the test set, each class has 200 samples, whereas in training set, each class has 500 samples except the IRES class which has 320 samples available for training.

5.3 Evaluation Criteria

A detailed parametric description of adapted ResNet based methodologies is summarized in Table 5.1. We use cross entropy loss function and Adam [219] optimizer with learning rate 0.001. In order to alleviate training time, an early stopping approach is used. Proposed classifier performance is evaluated using four different evaluation metrics namely: Accuracy, Precision, Recall and F_1 measure. Following existing studies [345], [296] leave one out cross-validation is used to perform experimentation.

5.4 Results and Discussions

This section briefly describes the performance of the proposed RPC-snRC classification system and two adapted ResNet architectures (ResNet 18 layers, ResNet 50 layers) for the task of ncRNA classification. It shows the impact of three sequence representation schemes while treating RNA sequence as a set of characters and k-mers based words for both proposed and adapted methodologies. In the benchmark dataset maximum length of the sequence is 1180, so to make the length of sequences equal, we apply paddings for the sequences which have length less than 1180. Experimentation is performed in two different ways: First, RNA sequence is taken as a set of characters with two different representation schemes namely one-hot vector encoding and random embedding initialization, which are separately fed to the proposed RPC-snRC system. Second, we generate 3-mers of the sequence by sliding a window of size three on the sequence. K-mers based sequence representations along with one-hot vector encoding, random embedding initialization and pretrained word embeddings provided by Asgari et al. [18] are fed to the proposed RPC-snRC system.

Performance Measures	Proposed RPC-snRC [28]				Res18-nRC				Res50-nRC			State-of-the-art	
	Character one-hot	3-mers one-hot	3-mers random embeddings	3-mers prot2vec embeddings	Character one-hot	3-mers one-hot	3-mers random embeddings	3-mers prot2vec embeddings	Character one-hot	3-mers random embeddings	3-mers prot2vec embeddings	nRC [296]	RNAGCN [345]
Accuracy	0.9538	0.9285	0.9327	0.9326	0.9169	0.8842	0.8880	0.9000	0.8680	0.8365	0.8915	0.7838	0.8573
Precision	0.9539	0.9312	0.9344	0.9322	0.9185	0.8859	0.8929	0.9000	0.8701	0.8377	0.8941	0.7780	-
Recall	0.9538	0.9285	0.9326	0.9326	0.9169	0.8842	0.8880	0.9000	0.8680	0.8365	0.8915	0.7830	-
F1-Score	0.9536	0.9286	0.9328	0.9319	0.9174	0.8842	0.8880	0.8987	0.8680	0.8357	0.8921	0.7790	0.8561

Table 5.3: Performance statistics of the proposed RPC-snRC, adapted (Res18-nRC, Res50-nRC) and state-of-the-art (nRC [296] and RNAGCN [345]) methodologies on the benchmark small non-coding RNA dataset.

Table 5.3 compares the performance of state-of-the-art and adapted resnet based methodologies with the proposed RPC-snRC methodology for the task of small non-coding RNA classification. It also illustrates the performance of the proposed RPC-snRC methodology when RNA sequence is treated as a set of characters, 3-mers based features with random and pre-trained neural word embeddings. As depicted by the Table 5.3 renowned methodology proposed by Antonio Fiannaca et al. [136] managed to achieve the performance figures of 78%, 77%, 78% and 77% in terms of accuracy, precision, recall and F1 measure, respectively. This performance is outperformed by a

recent Graph Convolutional Neural architecture based methodology given by Emanuele RossiGet et al. [345] as it marked state-of-the-art performance for small non-coding RNA classification with 85.7% accuracy. However, the adapted ResNet-18 and ResNet-50 manage to produce the peak performance of 91% and 89% by representing RNA sequences as character with one-hot encoding and as 3-mers features with pre-trained prot2vec embedding, respectively. On the other hand, the proposed RPC-snRC classification system has significantly outperformed the state-of-the-art methodology, as well as, the two adapted ResNet architectures in all settings. While, RPC-snRC with 3-mers random embedding initialization and pre-trained neural word embeddings schemes has raised state-of-the-art performance almost by the figure of 8% in terms of F1 measure, the RPC-snRC with character level features and one-hot encoding manages to mark the peak performance at 95% thereby clearly outperforming all the other systems (previously existing systems and ResNet based systems adapted in this research).

In a nutshell, convolutional neural network based deep architectures have the ability to extract discriminative features directly from primary sequences of small non-coding RNA. This is depicted by the results where performances of the proposed and adapted methodologies are significantly higher than the state-of-the-art methodologies which take secondary structural features as input. Moreover, performance of ResNet based architectures is lower than the performance of the proposed RPC-snRC methodology because in ResNet models gradient does not flow properly from subsequent layers to previous layers [191]. It can also be inferred that ResNet model with 50 layers extracted some irrelevant and redundant features which slightly reduced its performance as compared to the performance of ResNet 18 layers model.

5.4.1 Class Level Performance Comparison of Proposed RPC-snRC and State-of-the-art nRC Methodologies

In order to further compare the performance of the proposed RPC-snRC and the adapted ResNet based methodologies with the state-of-the-art methods, a class level performance comparison is performed in terms of accuracy confusion matrix. Accuracy confusion matrices of RPC-snRC, ResNet-18 and nRC methodologies on the test set of nRC dataset are shown in the Figure 5.3. RNAGCN [345] is the most recently reported method for small non-coding RNA classification, however, the authors have not provided class level results of their method. Therefore, we performed class level performance comparison of the proposed RPC-snRC and adapted methodologies with nRC classification methodology. The proposed RPC-snRC and the adapted Res18snRC based methodologies produce the highest performance with character level and one-hot vector representation. So here we take confusion matrices of both methodologies with the highest performance values. As depicted in Figure 5.3, RPC-snRC methodology correctly classifies all 200 samples of two classes namely Intron gpII and tRNA as compared to the state-of-the-art nRC methodology which manages to correctly classify only 180 samples of tRNA and 196 samples of Intron gpII class. Performance of Res18-snRC remains in between the performance of nRC and RPC-snRC

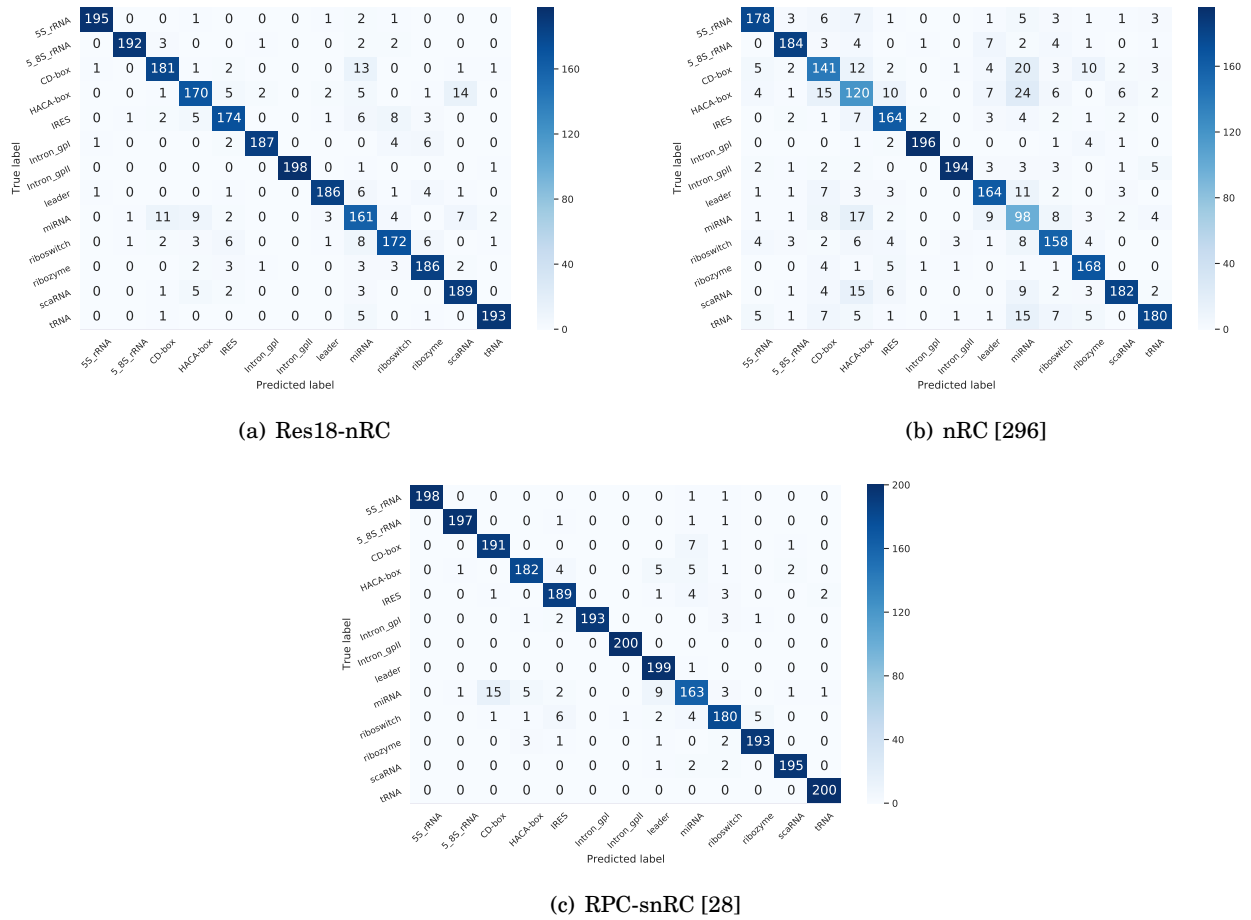


Figure 5.3: Accuracy confusion matrix of the proposed RPC-snRC [28], adapted Res18-nRC and state-of-the-art nRC [296] classification methodologies.

methodologies as it correctly predicted 198 samples of Intron gpII and 193 samples of tRNA class. In addition, state-of-the-art nRC methodology fails to mark prominent performance as significant samples of almost every class are mistakenly classified in miRNA, HACA-box, CD-box and IRES classes, while, only a few samples of each class are misclassified in the proposed RPC-snRC methodology.

Although, miRNA has shown the lowest performance among all classes in both methodologies, the proposed RPC-snRC still correctly classifies 163 samples out of the maximum possible 200 as compared to state-of-the-art nRC methodology which only manages to correctly classify only 98 samples. Also, the proposed RPC-snRC methodology successfully classifies more than 190 samples in each of the nine classes, i.e., intron_gpII, tRNA, 5S_rRNA, 5_8S_rRNA, leader, scRNA, ribozyme, intron_gpI and CD-box. Whereas, the other classes achieve counts of 180s and 160s as shown in Figure 5.3. In contrast to the state-of-the-art nRC methodology, only two classes intron_gpI and intron_gpII correctly classify more than 190 samples. Similarly, the adapted Res18-nRC methodology was able to correctly predict more than 190 samples for 4 classes, namely

5S_rRNA, 5_8S_rRNA, intron_gpII and tRNA.

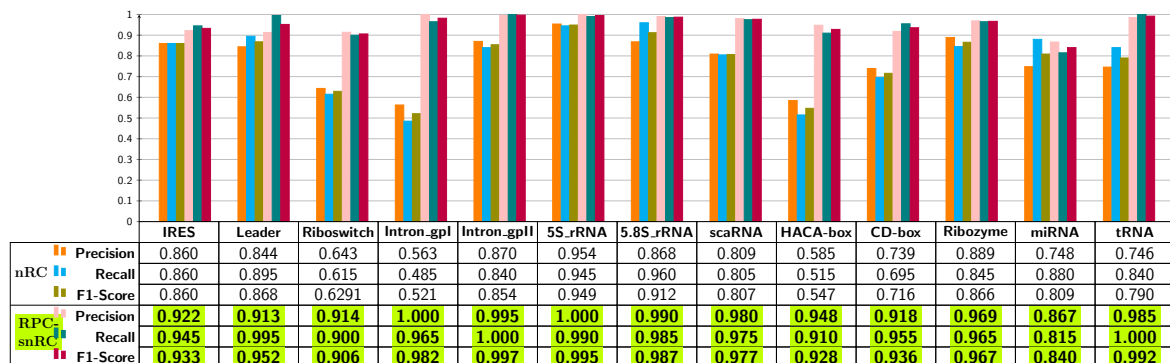


Figure 5.4: Class level performance comparison of proposed RPC-snRC [28] approach and state-of-the-art nRC [296] approach using small ncRNA classification dataset.

Figure 5.4 shows individual class level performances of RPC-snRC and nRC classification methodologies over small ncRNA classification dataset in terms of precision, recall and F_1 measure. Overall, for all classes, RPC-snRC methodology significantly outperforms the state-of-the-art nRC methodology in all three performance metrics with exception of the miRNA class, where nRC methodology manages to deliver better recall figure. Moreover, among all performance metrics, nRC classification methodology manages to sustain performance values of precision, recall and F_1 measure only for three classes (IRES, 5.8S rRNA, scaRNA). On the other hand, the performance of RPC-snRC classification methodology remains consistent for 6 classes namely: ribozymes, 5_8S_rRNA, tRNA, scaRNA, Intron_gpII and riboswitch. This unique behavior of RPC-snRC methodology shows that it suffers less from type I and type II errors as compared to nRC methodology-performance which seems less stable at class level.

5.5 Conclusion

This chapter presents a novel RPC-snRC methodology, which classifies small non-coding RNA sequences into their relevant families by utilizing positional and occurrence information of various nucleotides. Experimental results reveal that the proposed RPC-snRC methodology is highly robust as it is neither biased towards false positive nor towards false negative predictions. Adapted Res18-snRC and Res50-snRC methodologies perform better than the state-of-the-art small non-coding RNA classification methodologies. However, their performance is less than the performance of proposed RPC-snRC methodology because in ResNet architectures gradient cannot flow properly from subsequent layers to previous layers. The proposed RPC-snRC methodology marks the highest F1-score of 95% by representing character based features through one-hot encoding, while state-of-the-art ncRNA, RNAGCN and adapted Res18-nRC, Res50-nRC classification methodologies manage to produce the performance figures of 77%, 85%, 91% and 89%, respectively. Moreover, in our experimentation, almost all methodologies perform better with

one-hot vector encoding than randomly initialized or pretrained word embeddings. From these results, it can be concluded that character or atom level feature generates better performance as compared to k-mers based features.

CIRCULAR RNA IDENTIFICATION

Among different types of non-coding RNAs, following recent findings about different roles of circular RNA in diverse biological processes such as disease prediction and their use in therapies; in depth exploration of circular RNAs become more attractive area of research [194]. Moreover, identification of suitable drugs targeting the regulatory circuits of functional RNAs requires information about the subtype of non-coding RNAs family also known as RNA sequence classification [28].

In order to perform circular RNA classification, it is necessary to understand the formation of circular RNAs [372]. In circular RNA formation, firstly DNA is transcribed into a precursor messenger RNA (pre-mRNA) [194] that consists of introns and exons regions [388]. Through the process of splicing from pre-mRNA, intron regions are removed and mature messenger RNA is produced from exons regions [388]. The process of pre-RNA splicing, also produces circular RNAs [241]. A slight change in the process of circular RNA formation leads towards failure of various biological processes that initiate and propagate diverse types of diseases, such as cancer, Alzheimer and Parkinson [31, 246, 467]. Although, circular RNA is found to be involved in various biological processes, however, its complete functionality still remains unexplored [467]. Precise identification of circular RNAs facilitates in depth exploration of their biological roles [13, 136, 412, 466]. Circular RNA classification is different from small non-coding RNA classification, as distribution of nucleotides in circular RNA is different from small non-coding RNAs. Furthermore, length of circular RNA sequences is much longer than small non-coding RNA sequences [31, 179, 253].

⁰This chapter is an adapted version of the work presented in Asim et al., "CircNet: an encoder–decoder-based convolution neural network (CNN) for circular RNA identification", In *Neural Computing and Applications (2021)* [372] and Asim et al., "Advances in Computational Methodologies for Classification and Subcellular Locality Prediction of Non-Coding RNAs", In *International Journal of Molecular Sciences (2021)* [19]

6.1 Related Work

One way to classify or identify circular RNAs is to perform laboratory experiments, as done by Zaghlool et al. [453] and Zirkel et al. [479]. Unfortunately, performing classification through such experimental methods suffers from multiple drawbacks such as Zirkel et al. [479], experimental method requires chemical materials that are costly and experimentation process is time consuming [479]. Laboratorios experimentation is error-prone, as in Zaghlool et al. [453] work a low reproducibility rate of different experimental methods is reported. Furthermore, relatively low appearance rate of circular RNAs compared to other RNAs and circular RNAs sequence similarity with non-linear RNAs make their classification difficult.

Thanks to high-throughput technologies which produce large amount of nucleotide sequencing data [194, 425] and provide another way to perform RNA classification by utilizing machine learning approaches. To the best of our knowledge, there are currently three computational approaches that can discriminates circular RNAs from other long non-coding RNAs.

The first approach PredcircRNA proposed by Pan et al. [324] generates statistical representation of raw RNA sequences by extracting seven different features including graph features, sequence composition, conservation information, tandem repeat, ALU, ORF features and SNP density. Based on generated statistical representation, multi kernel learning classifier acquire a linear weight combination of multiple kernels in which every kernel transforms the hands-on representation into a higher-dimensional space where data becomes linearly separable. Finally, SVM classifier makes use of high-dimensional feature space to make final predictions. Using a similar set of features, Chen et al. [75] developed H-ELM predictor, which additionally utilizes minimum redundancy maximum relevance (mRMR) as well as iterative features selection approach with an aim to retain discriminative set of features. By using the most informative features hierarchical extreme learning classifier discriminates circRNAs from other lncRNAs. CircDeep [68] predictor transforms raw sequences to statistical vectors by utilizing three different encoding methods. Conservation scoring method extracts motif specific information while other two encoders word2vec embedding generation model and Reverse Complement Matching (RMC) method capture context of nucleotides. Furthermore, three different types of representations are passed to hybrid model based on Convolution Neural Network (CNN) and Bidirectional Long Short-Term Memory (BLSTM) layers that extracts discriminative features and perform classification.

Existing predictors utilize hand crafted features, while recent research about genomics analysis has proved that deep learning based methodologies perform better when they are fed with raw DNA or RNA sequences as compared to their performance when they are fed with manually extracted features. We [28] proposed an end-to-end deep learning based approach for small non-coding RNA classification. Based on the experimental results we concluded that when deep learning predictors are fed with hand crafted features, their performance decreased because during the process of feature extraction important information about occurrences and positions

of nucleotides may get lost. Our experimental results proved that deep learning methodologies perform better by extracting more discriminative features from raw sequences based on the position and occurrences of basic nucleotides. Moreover, to perform different sequence analysis tasks such as classification of long non-coding RNAs and RNAs subcellular location prediction, several deep learning predictors have produced state-of-the-art performance values by utilizing raw RNA sequences.

In order to improve the performance of circular RNA identification, we propose a two stage classification methodology where at first stage we utilize an encoder decoder approach for the extraction of latent space and at second stage, by utilizing learned representation, a convolutional neural network is used for the extraction of discriminative features. Discriminative features are fed to a fully connected layer that discriminate circular RNAs from other long non-coding RNAs. Lastly, in order to explore different regions of genome which contain more important information about the identification of circular RNAs, we performed extensive experimentation by taking different combinations of sequence lengths, scaling methods and number of added adjacent nucleotides.

6.2 Materials and Methods

This section describes the details of proposed predictor and benchmark dataset.

6.2.1 Proposed Methodology

We propose a two stage classification methodology, where at first stage we learn discriminative features by utilizing an encoder-decoder architecture and at second stage the learned features are passed to a convolutional neural network for the extraction of more discriminative features and to perform classification between circRNAs and other long non-coding RNAs (lncRNAs). The encoder utilizes convolution and pooling, while decoder makes use of deconvolution and un-pooling/up-sampling to reconstruct the original raw sequence. The key idea is to apply encoder based convolutional operations to learn sequence representation in less space while the up-sampling of decoder network makes sure whether the sequence can be reconstructed from the learned space. This architecture substantially reduces the number of trainable parameters of classifier. A brief description of encoder-decoder architecture is given in section 6.2.1.2 and deep learning classifier is described in section 6.2.1.3. In order to understand the area of genome that contains more important information about the identification of circRNAs, we take different segments of the genome which are briefly described in the preprocessing stage, section 6.2.1.1.

6.2.1.1 Preprocessing

The dataset provided by Chaabane et al. [68] contains circular RNA and lncRNA sequences along with their positional information in the human genome, i.e., in the genome sequence

start and end locations of nucleotides on the basis of which authors extracted non-coding RNA sequences. Utilizing start and end locations, we extracted adjacent nucleotides from genome sequence and embed them with non-coding RNA sequences, since these regions might contain valuable information about circular RNA classification [205, 405]. Adjacent nucleotides appear directly after or before the start and end locations of non-coding RNA sequences. The concept of adjacent nucleotides is illustrated in Figure 6.1, where let's we have a genome sequence in which start and end positions denote the sequence of circular RNA. To more precisely illustrate the concept of adjacent nucleotides, in the genome sequence CAG nucleotides are before the start position and ATC nucleotides are after end position of circular RNA sequence.

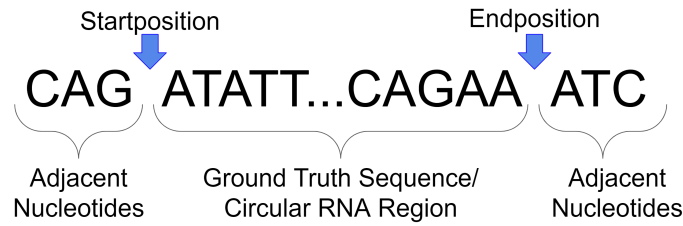


Figure 6.1: Circular RNA sequence extension by adding adjacent nucleotides

Furthermore, in the benchmark dataset, length of non-coding RNA sequences varies from 201 to 3050672 nucleotides, however, deep learning predictors require same length of sequences. In order to fix the length of sequences, rather than taking maximum length of 3050672 nucleotides and apply zero padding to shorter sequences, we set maximum length to a predefined number M , then zero padding is applied in the sequences which are smaller than M and truncates the nucleotides from sequences which are longer than M . We take 3 different values 200, 500 and 1000 for M . The size of sequences equal to a predefined length M is made by applying three padding approaches, denoted as *post*, *pre* and *middle*. In *post* padding, from sequences which are longer than predefined number M , we remove all nucleotides appearing after the M^{th} nucleotide. If the sequences are shorter than M , an additional *zero Z* symbol is added at the end of the sequence as many times as needed to achieve the predefined length M . On the other hand. *pre* padding removes or adds nucleotides from the beginning of the sequences. In *middle* approach, the first and last $M/2$ nucleotides of sequences are kept, while removing or adding nucleotides in between. An illustration of said approaches can be seen in Table 6.1.

Table 6.1: Scaling of two different sequences ATAG and ATATGUAT to length of 6 by either addition or removal with three different methods namely Pre, Middle and Post.

	Pre	Middle	Post
Addition	ZZATAG	ATZZAG	ATAGZZ
Removal	ATATGUAT	ATATGUAT	ATATGUAT

As deep learning methodologies require data in real number format, we transform each sequence

in one-hot encoded representation. Furthermore, we extract the sequence from the genome dataset based on positional information, which also includes the letter N in addition to the four nucleotides A, C, G and U. Note that due to ambiguity between nucleotides, an exact identification is not always possible. Therefore, the additional symbol N represents either A, U, C or G. Because we are interested in the positions of nucleotides rather than removing them, we give them a one-hot vector representations. In one-hot encoding every nucleotide is represented by a vector of five bits, where four bits are 0 and one bit is 1. The position of the 1 bit is always the same for a specific nucleotide. Using this methodology adenine is represented as $A = [1,0,0,0,0]$, Cytosine $C = [0,0,0,1,0]$, Guanine $G = [0,0,1,0,0]$, Uracil $U = [0,1,0,0,0]$, $N = [0,0,0,0,1]$ and zero symbol $Z = [0,0,0,0,0]$.

6.2.1.2 Latent Space Extraction using Autoencoder

We utilize raw ncRNA sequences for the extraction of latent space features, where each RNA sequence has four basic nucleotides: adenine (A), cytosine (C), guanine (G) and uracil (U). Furthermore, each nucleotide is encoded using one hot vector encoding, as described in section 6.2.1.1. A graphical representation of proposed autoencoder used for latent space learning is shown in Figure 6.2.

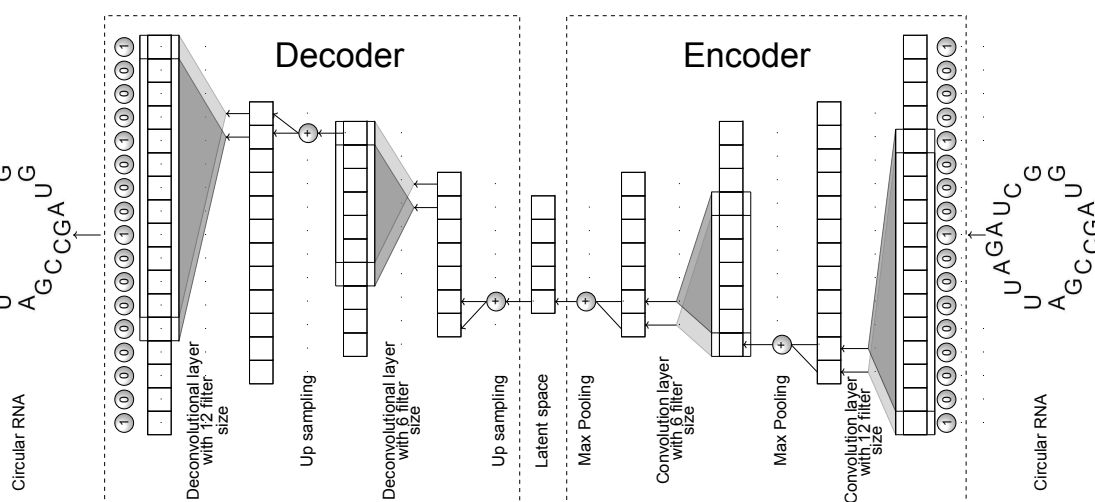


Figure 6.2: Graphical representation of the employed autoencoder.

We use 1d convolutional layers with 128 filters, kernel size 12 and stride size 1. This layer extract discriminative features based on the nucleotide's occurrences and positions. In order to reduce the dimensions of extracted feature space, we employ a max pooling layer with kernel size 2. Another convolutional layer, with 128 filters, kernel size 6, stride size 1 and max pooling layer with kernel size 2 is used to extract more discriminative features. Reconstruction of initial sequence from the latent space verifies the extraction of comprehensive features. For this purpose, we use the same number of layers in reverse order. The output of each convolutional layer is calculated by:

$$c_{xyf} = \sum_{i=1}^k \sum_{j=1}^5 n_{ksf,ij} w_{ksf,ij} + b_{ksf,ij} \quad (6.1)$$

where f denotes the f^{th} filter, x and y represent the indices of the output tensor, k and s define the currently observed patch of the input tensor n given as kernel and stride size, respectively. i and j denote indices inside this patch. Furthermore, w and b define the learned weights and biases, respectively. Considering sequence length of 200, the input is given as a 200×5 tensor for the first convolutional layer. Its output is defined by a 200×128 tensor which gets reduced to a 100×128 after applying max pooling. Max pooling calculates the output as a tensor where each index xyf is calculated as follows:

$$m_{xyf} = \max c_{ksf} \quad (6.2)$$

where f denotes the f^{th} filter, x and y the indices of the output tensor and k and s define the currently observed patch of the input tensor c given as kernel and stride size, respectively. The second convolutional layer does not change the shapes. However, the second max pooling layer again halves the shape to 50×128 , which is our latent feature representation. The decoder has the same shapes in reverse order. All layers are utilizing Relu as the activation function defined by:

$$a(y) = \max(0, y) \quad (6.3)$$

where y is the output of a layer. However, in the last reversal operation of the decoder sigmoid is applied, which is defined by:

$$a(y) = \frac{1}{1 + e^{-y}} \quad (6.4)$$

6.2.1.3 Convolutional Neural Network based Classifier

Figure 6.3 illustrates architecture of proposed classifier that uses latent space and discriminates between circular RNA and other lncRNA.

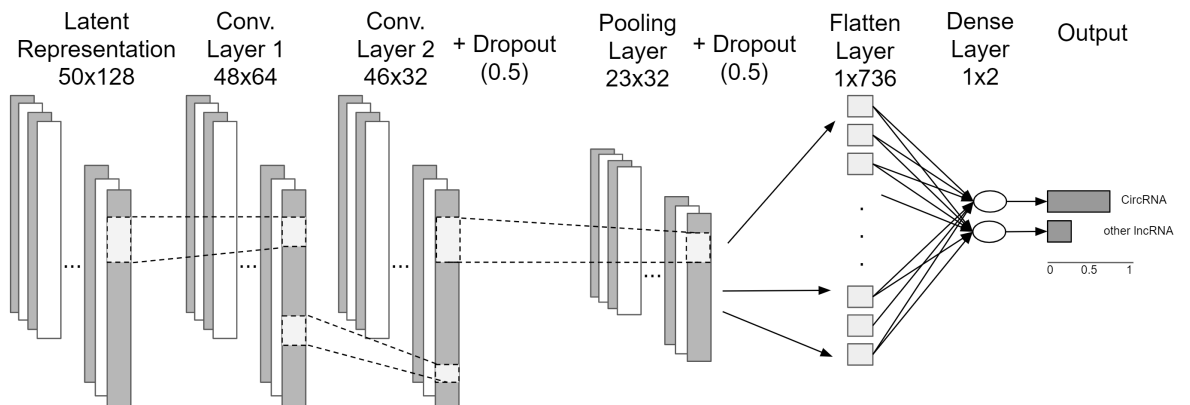


Figure 6.3: Graphical representation of the employed classifier.

Latent space is fed to two one dimensional convolutional layers with kernel size 3, stride size 1 and 64, 32 filters for the first and second layers, respectively. Following this we have a dropout

layer with a probability of 0.5 and max pooling with kernel size and stride size equal to 2. Finally, a flattened layer followed by a dense layer is used. Note, that we do not freeze the weights of the trained encoder in our classification model, but instead fine-tuned the weights during the second training stage.

Convolutional and max pooling layers working paradigm is explained in section 6.2.1.2. Dropout randomly ignores a fixed percentage of neurons during the optimization step. All layers utilize Relu as an activation function, besides the last dense layer which uses softmax defined as:

$$a(y)_i = \frac{e^{y_i}}{\sum_{j=1}^K e^{y_j}} \text{ for } i = 1, \dots, K \text{ and } y = (y_1, \dots, y_K) \in \mathbb{R}^K \quad (6.5)$$

Let's assume we have an input sequence of length 200, our input shape is defined by 200×5 which gets transformed to 50×128 by the encoder, as defined in section 6.2.1.2. Our convolutional based classifiers change the shape to 48×64 and 46×32 for the first and second convolutional layers. Dropout does not change shapes of sequences. On the other hand, max pooling reduces the shape to 23×32 , flattening this shape results in a vector of size 736. Lastly, the dense layer calculates our final prediction with an output size of 2×1 .

6.2.2 Benchmark Dataset

In order to evaluate the integrity of proposed CircNet approach we performed experimentation on the publicly available benchmark dataset provided by Chaabane et al. [68]. It consists of two classes, circular RNAs and other lncRNAs. Chaabane et al. [68] utilized CircRNADb database to extract 31939 circular RNA sequences [82]. On the other hand, the GENCODE database was used to extract 19683 lncRNAs [139]. More details about the dataset, such as the minimal, maximal, average sequence length and the standard deviation of all sequences is summarized in Table 6.2.

Table 6.2: Statistics of benchmark dataset, where minimal and maximal sequence length represents the length of shortest and longest sequences, respectively. On the other hand average and standard deviation of sequence length illustrate the mean and standard deviation of sequences in the corresponding classes.

Measure	Positive class	Negative class	Both classes
Minimal sequence length	201	204	201
Maximal sequence length	3050672	1536213	3050672
Average sequence length	19924	18653	19439
Standard Deviation of sequence lengths	34439	47025	39716

6.3 Evaluation Criteria

In order to ensure a fair performance comparison of proposed CircNet predictor with state-of-the-art circular RNA classification approaches [68, 75, 324], we performed experimentation

with standard data splits provided by Chaabane et al. [68] where benchmark dataset has 75% train, 15% test and 10% validation samples. Following evaluation criteria of existing studies [68, 75, 324], we utilize 6 different evaluation measures namely accuracy, f1-measure, Matthews correlation coefficient, specificity, recall and AUROC. The learnable parameters of CircNet predictor are optimized through RMSProp optimizer, with an initial learning rate of 0.001 and the Mean Squared Error (MSE) as the loss function.

6.4 Results and Discussions

We performed experimentation by scaling the length of highly variable non-coding RNA sequences to 3 different predefined lengths: 200, 500 and 1000. To accomplish this we utilize 3 different fixed-length generation strategies: middle, pre and post. Furthermore, with an aim to verify claim that important information can be extracted from genome sequence regions adjacent to circular RNA sequences, we also performed experimentation by extracting 2 different lengths (50, 100) adjacent nucleotides from genome sequences and combining them with non-coding RNA sequences. A detailed description about predefined lengths values and 3 fixed-length generation strategies is provided in section 6.2.1.1. Considering, all possible settings, 26 experiments have been performed and the results are summarized in Table 6.3.

From Table 6.3, it can be concluded that when CircNet is fed with only circular RNA sequences using three different padding schemes (post, pre and middle), it performs better with middle padding method. This proves that in a sequence more important information lies at the beginning and end of a sequence. When CircNet is fed with an input of 200 nucleotides and padding at the middle it produces performance figures of 0.9827, 0.9860, 0.9633 and 0.9813 in terms of accuracy, f1, MCC and specificity, respectively. However, when input length increases to 500 nucleotides the performance of all three measures improves. The same scenario holds true when the length is increased to 1000 nucleotides.

On the other hand, experimental results also validate that by using adjacent nucleotides performance gets improved. Along with the addition of adjacent nucleotides, here once again middle padding approach performed better as compared to other pre and post padding approaches. Comparing the best performing model which does not use adjacent nucleotides with the worst performing model which uses adjacent nucleotides, there is an increase of 5.64% for accuracy, 4.41% for F1, 12.12% for MCC and 10.92% for specificity for the latter model. Among different experimental settings, 100 adjacent nucleotides and 1000 sequence length with middle fixed-length generation strategy produce the best performance in terms of accuracy, F1 and MCC, while 50 adjacent nucleotides with a sequence length of 500 achieved the best specificity.

Furthermore, we evaluate the integrity of proposed CircNet approach using AUROC curves and the respective AUROC values at different experimental settings. Figure 6.4(a) illustrates AUROC curves when CircNet was fed with original sequences and the curves of Figure 6.4(b)

Table 6.3: Performance statistics of CircNet [372] based on different adjacent nucleotides, scaling methods and sequence lengths.

Number of adjacent nucleotides	Scaling	Seq. Len.	Acc.	F1	MCC	Spec.
0	Middle	200	0.8985	0.9192	0.7832	0.8462
		500	0.9063	0.9266	0.7997	0.8278
		1000	0.9134	0.9315	0.8148	0.8544
	Post	200	0.8301	0.8709	0.6335	0.6774
		500	0.8293	0.8689	0.6315	0.6941
		1000	0.8372	0.8705	0.6518	0.7642
	Pre	200	0.8802	0.9052	0.7434	0.8115
		500	0.8831	0.9067	0.7506	0.8299
		1000	0.8751	0.9040	0.7328	0.7567
50	Middle	200	0.9827	0.9860	0.9633	0.9813
		500	0.9818	0.9853	0.9615	0.9826
		1000	0.9813	0.9849	0.9602	0.9755
	Post	200	0.9771	0.9816	0.9514	0.9653
		500	0.9771	0.9816	0.9514	0.9650
		1000	0.9768	0.9813	0.9506	0.9646
	Pre	200	0.9700	0.9758	0.9366	0.9670
		500	0.9702	0.9759	0.9369	0.9677
		1000	0.9702	0.9759	0.9368	0.9656
100	Middle	200	0.9810	0.9847	0.9598	0.9855
		500	0.9823	0.9858	0.9624	0.9724
		1000	0.9828	0.9862	0.9635	0.9775
	Post	200	0.9770	0.9815	0.9511	0.9639
		500	0.9773	0.9818	0.9517	0.9636
		1000	0.9775	0.9819	0.9523	0.9677
	Pre	200	0.9703	0.9760	0.9371	0.9660
		500	0.9702	0.9759	0.9368	0.9639
		1000	0.9698	0.9756	0.9360	0.9650

represent AUROC values when CircNet was fed with original sequences along with adjacent nucleotides. A detailed description of how we define adjacent nucleotides and what our motivation is in using them, is given in section 6.2.1.1. Briefly, these are nucleotides appearing before and after the circular RNA sequence in the original genome. From Figure 6.4 it can be concluded that the approach denoted with middle, in which we extract nucleotides from the beginning and end of the sequence, improves circNet performance, as compared to its performance when it was fed with pre and post sequence length selection method. In the case of not using adjacent nucleotides, middle length scaling approach achieves an AUROC of 0.96 while pre and post length selection methods achieve AUROC of 0.94 and 0.90, respectively. Moreover, as written in our motivation, important information is contained in adjacent nucleotides, as it can be seen from the AUROC values, where the AUROC value is always higher when including adjacent nucleotides compared

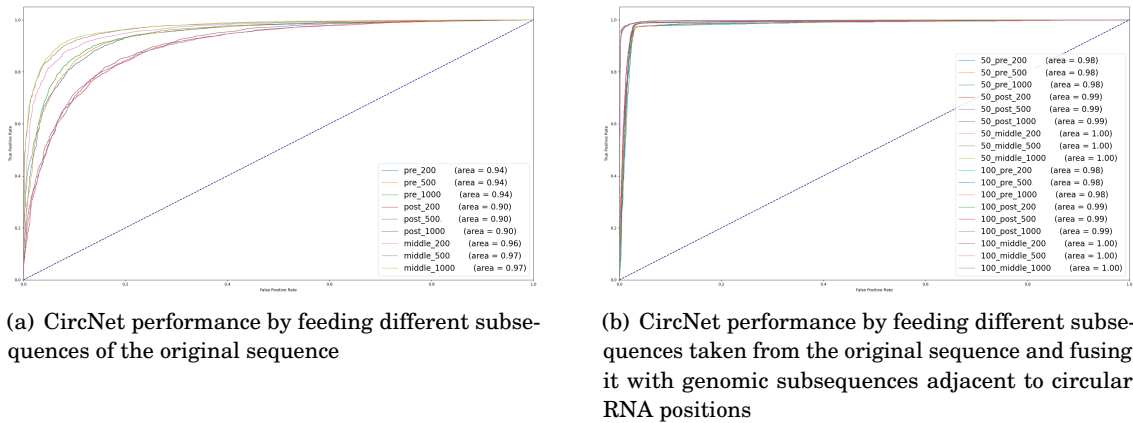


Figure 6.4: CircNet [372] performance in terms of AUROC for different experimental settings by taking subsequences from different positions along with fusion of genome adjacent nucleotide information

to using only the original ones. The worst AUROC measure in the adjacent case is 0.98, while the best AUROC value for the nonadjacent case is 0.97. Visually, this improved performance can also be observed in the AUROC curves, since the curves in the adjacent case converge faster to a high true positive rate, compared to the nonadjacent ones. Lastly, the curves representing the adjacent case all behave very similar and are quite close to each other, unlike the nonadjacent case, where many curves vary largely.

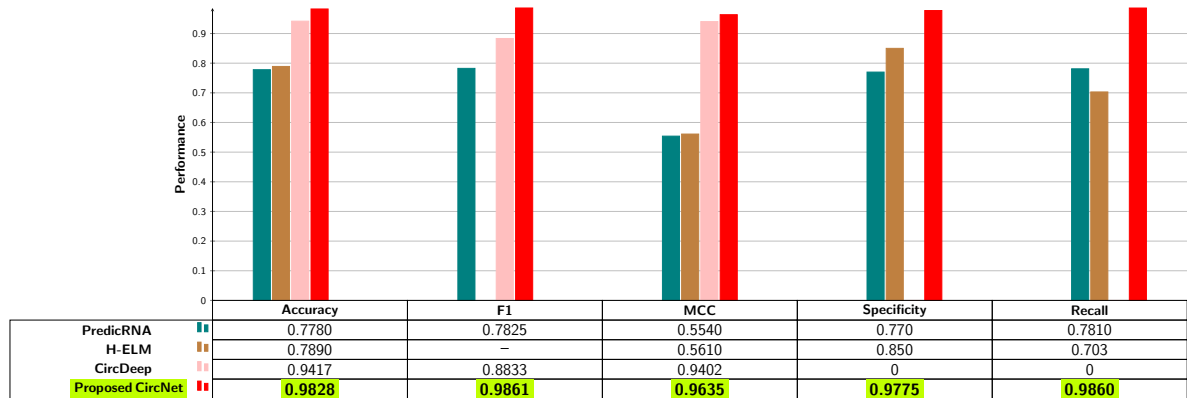


Figure 6.5: Performance comparison of proposed CircNet predictor with existing circular RNA classification approaches.

Figure 6.5 illustrates performance comparison of our best performing CircNet approach setting with three previous machine and deep learning based approaches. PredcircRNA [324] approach makes use of hand crafted features that degrade its performance, so it manages to produce 77% accuracy, 78.1% F1 measure and 55.4% MCC. On the other hand, although H-ELM [75] approach uses hand crafted features, but it removes irrelevant and redundant features through feature selection method that slightly improves its performance and makes it better predictor

than PredcircRNA [324]. CircDeep [68] approach makes use of a combination of only two hand crafted features with a feature representation learned through a deep learning model, based on convolutional neural networks and long short term memory layers. As compared to PredcircRNA [324], together all three encoders improve CircDeep predictor performance values of accuracy, F1, MCC with a significant margin of 16.37%, 10.08% and 38.62%, respectively. In comparison to CircDeep predictor, proposed CircNet approach [372] improves the performance values of accuracy, F1 and MCC with a significant margin of 4.11%, 10.29% and 2.33%, respectively. We do not compare specificity and recall values of proposed CircNet approach with the circDeep [68] approach as authors did not report the values of these measures. In comparison with other two approaches, namely PredcircRNA [324] and H-ELM [75], proposed CircNet approach achieves 12.75% and 20.75% improvement in terms of specificity, respectively. Similarly, CircNet approach also outperforms both existing approaches in terms of recall with a significant margin of 20.5% from PredcircRNA [324] and 28.3% from H-ELM [75] approach.

6.5 Conclusion

This chapter presents CircNet approach that makes use of autoencoder and CNN based classifier competent in categorizing circular RNAs from other lncRNAs. Proposed CircNet approach outperformed state-of-the-art CircDeep [68] predictor with a significant margin of 4.11%, 10.29% and 2.33% in terms of accuracy, F1 and MCC. With an aim to find the most discriminative regions of RNA sequences, we performed extensive experimentation by taking different sequence lengths, scaling methods and extension of the sequences. Extension of sequences is performed by incorporating adjacent nucleotides. We observed that addition of adjacent nucleotides improves the performance of predictor. Based on predictor performance improvement, it can be concluded, that in the genome sequence adjacent regions of circular RNAs preserve information about their classification. Lastly, among 3 different sequence fixed-length generation strategies, proposed predictor produces better performance by taking nucleotides from starting and ending regions of sequences. This performance gain reveals in non-coding RNA sequences more comprehensive information about circular RNA classification is present in the starting and ending regions.

Finally, we hope to accurately distinguish between circular and other long non-coding RNAs, proposed predictor will facilitate to understand the roles of circular RNAs in biological processes, which in turn will expedite diagnosis and treatment of many severe diseases, such as cancer, diabetes and respiratory illness.

RNA SUBCELLULAR LOCATION PREDICTION

Biological functions of a variety of Ribonucleic Acids (RNAs) such as messenger RNA (mRNAs) [27, 252], microRNA (miRNAs) [24, 27], small nucleolar RNA (snoRNAs), long non-coding RNAs [127, 263] and circular RNA rely on their localization in various subcellular compartments such as nucleus, cytoplasm and cytosol [215, 403]. mRNAs localization in nucleus regulate gene expression by eliminating defective RNAs from the cell and tweaking the expression levels of various non-coding RNAs [146, 383]. It provides quantitative as well as spatial control over the production of proteins by localizing in cytoplasm [252]. mRNA localization in cytosol helps to maintain cell membrane and control the use of nutrients for metabolism [293, 468]. miRNAs localization in nucleus plays a key role in cell division where each cell divides into identical daughter cells with an objective to promote organism growth and well-being by replacing worn out cells [294]. Furthermore, miRNAs localization in cytoplasm causes gene silencing by binding to mRNA molecules [294]. Small nucleolar RNAs (snRNAs) play a key role in post-transcriptional regulation by guiding RNA modifications of ribosomal RNAs (rRNA), transfer RNAs (tRNAs) and small nuclear ribonucleic acid RNAs (snRNAs) molecules by localizing in the nucleus [371]. Long non-coding RNAs (lncRNAs) control gene expression through chromatin remodeling by localizing in nucleus [55]. In cytoplasm, lncRNAs avoid mRNAs degradation as well as repress miRNAs to reduce their regulatory effects on mRNAs [456]. Within the nucleus, circular RNAs enhance the

⁰This chapter is an adapted version of the work presented in Asim et al. "EL-RMLocNet: An Explainable LSTM Network for RNA-Associated Multi-Compartment Localization Prediction", In *Computational and Structural Biotechnology Journal* (2022) [23], Asim et al., "MirLocPredictor: A ConvNet-Based Multi-Label MicroRNA Subcellular Localization Predictor by Incorporating k-Mer Positional Information", In *Genes* (2020) [27], Asim et al., "Advances in Computational Methodologies for Classification and Subcellular Locality Prediction of Non-Coding RNAs", In *International Journal of Molecular Sciences* (2021) [19], Asim et al., "Circ-LocNet: A Computational Framework for Circular RNA Subcellular Localization Prediction", In *International Journal of Molecular Sciences* (2022) [306] and Asim et al., "L2S-MirLoc: A Lightweight Two Stage MiRNA Subcellular Localization Prediction Framework", In *International Joint Conference on Neural Networks, (IJCNN)* (2021). *IEEE* [24]

expression of mRNAs and within cytoplasm compartment they perform different regulatory roles by interacting with proteins and miRNAs [181] [258].

The subcellular localization of RNAs is an efficient and a widespread strategy to target the gene products to a particular region of various cells. Localization of various RNA molecules controls the translation of mRNAs into proteins in a temporal and spatial manner. It influences which type and number of proteins will be produced within certain cell by regulating the production of mRNA molecules and the amount of time they reside in the cytoplasm. Likewise, the spatial distribution of the RNA molecules mainly influences cellular concentration as well as location of its corresponding proteins which impact the cell function and its aptitude to interact with neighboring cells or respond to environmental changes. Furthermore, it has the potential to avoid toxicity of various protein products, generates fast cellular responses and determines molecular interactions [354, 360, 454]. It provides the basis for spatial differences in shape, structure and function of a variety of cells in order to ensure that each cell exhibits a unique form of polarization [323, 421]. Characterizing RNA subcellular localization is essential for thorough categorization of different cell types and cell states [353]. In addition to facilitate a deep understanding of molecular and cellular biology, knowledge of RNA subcellular localization is also beneficial for the development of heterogeneous biomedical applications [353]. Like subcellular localization of messenger RNAs (mRNAs) assists to identify and treat Huntington's disease by eliminating active mRNAs of disease specific gene in nucleus and cytoplasm [107]. Also, mRNAs guide protein synthesis by localizing in cytoplasm [252], paving way for the production of the most effective recombinant proteins [214]. Furthermore, considering the association between RNA expression levels in different subcellular compartments with a variety of diseases such as Cancer [107], accurately determining RNA subcellular localization can largely assist to demystify their roles in various disease as well as to design optimized therapeutics responsible to increase or decrease various RNAs expression levels in the target subcellular compartment.

7.1 Related Work

Considering the efficiency and robustness of computational approaches shown in various fields such as Natural Language Processing [320] and Bioinformatics [442], to date, a number of Artificial Intelligence based RNA subcellular localization predictors have been developed which are summarized in Table 7.1. The paradigms of existing approaches can be broadly classified into 2 categories, single compartment localization prediction (SCLP) [89, 132, 426, 437, 468?] and multi-compartment localization prediction (MCLP) [403]. To better illustrate SCLP and MCLP paradigms, consider a hypothetical corpus C which contains 5 RNA sequences $X = X_1, X_2, X_3, X_4, X_5$ that belong to 5 subcellular compartments $L = Nucleus, Cytoplasm, Mitochondria, Cytosol, Exosome$. In SCLP, each RNA sequence X_i belongs to exactly one subcellular compartment L_i , such as X_1 belongs to Nucleus, X_2 belongs to Cytoplasm and so on.

Whereas, in MCLP, each RNA sequence X_i belongs to more than one subcellular compartment L_i at the same time, such as X_1 belongs to {Nucleus, Exosome}, X_2 belongs to {Mitochondria, Cytosol, Cytoplasm } and so on.

Table 7.1: A summary of existing computational subcellular localization predictors for miRNA, lncRNA, mRNA and circular RNA molecules

Approach	Sequence Encoding Methods Cardinality	Nucleotide Encoding	Classifier
RNA Type: miRNA			
Our L2S-MirLoc [24]	Multi-Label	Electron Interaction PseudoPotentials (EIIP)	Random Forest (RF)
miRNALoc [294]		pseudo dinucleotide compositions and di-nucleotide properties	Support Vector Machine (SVM)
Our MirLocPredictor [27]		positional and semantic information of k-mers (kmerPR2Vec)	Convolutional Neural Network (CNN)
MirGOFs [437]		functional similarity based encoding matrix	microRNA-based similarity inference model
MiRLocator [426]		K-mer embeddings using Word2vec (RNA2Vec)	BiLSTM encoder-decoder model
RNA Type: LncRNA			
iLoc-LncRNA 2.0 ¹	Multi-Class	fusing mutual information algorithm and incremental feature selection strategy	SVM
lncLocation [133]		k-mer frequency, physicochemical properties and secondary structure features Autoencoder and binomial distribution based feature selection	SVM, RF, Logistic regression, XGBoost, lightGBM, DNN and CNN
Locate-R [4]		K-mer composition and Pearson based filtering	Deep SVM
lncLocator 2.0 [263]		Glove embeddings	CNN, BiLSTM, MLP
lncLocator [65]		k-mer frequency and stacked autoencoder	stacked ensemble classifier (SVM, RF)
iLoc-lncRNA [373]		binomial distribution-based feature selection, Pseudo K-tuple Nucleotide Composition	SVM
DeepLncLoc [456]		subsequence embeddings	CNN
lncLocPred		k-mer, triplet and PseDNC VarianceThreshold, binomial distribution and F-score based feature selection	Logistic Regression
Yang et al LncRNAPred [436]		kmer nucleotide composition, Analysis Of Variance (ANOVA) based feature selection	SVM
DeepLncRNA [158]		k-mer, RNA binding motifs Genomic loci	feed-forward multi-layer deep neural network
KD-KLNMf [463]		k-mer and dinucleotide based spatial autocorrelation, KLD non-negative matrix factorization based feature selection	SVM
RNA Type: mRNA			
mLoc-mRNA [293]	Multi-Label	k-mer frequency and elastic-net based feature selection	RF
DM3Loc [402]		One-hot encoding	Attention based CNN
Zhang mRNAloc [468]	Multi-Class	9-mer, binomial distribution and one-way analysis of variance based features	SVM
RNATracker [429]		One-hot encoding	Hybrid (CNN+ LSTM+Attention)
mRNALoc [146]		pseudo k-tuple nucleotide composition	SVM
mRNALocater [383]		pseudo k-tuple nucleotide composition electron-ion interaction pseudopotential, correlation coefficient filtering	Ensemble(CatBoost+ LightGBM+XGBoost)
SubLocEP [252]		Nucleotide physicochemical properties	Weighted LightGBM
NN-RNALoc [30]		k-mer frequency, distance-based subsequence profiling and PCA for dimensionality reduction	Multi-Layer DNN
RNA Type: Circular RNA			
Our Circ-LocNet [306]	Multi-Class	K-Mer, Reverse Complement Kmer, Pseudoknc, Xxkgap, Z-Curve, Electron-Ion Interaction Pseudopotentials of Trinucleotide (Eiip)	RF, Xgboost, Naive Bayes, SVM, AdaBoost
RNA Type: miRNA, mRNA, lncRNA, snoRNAs			
Multi-compartment localization predictor [403]	Multi-Label	K-Mer4, K-mer1234, Reverse Complement Kmer, NAC, DNC, TNC, composition of k-spaced nucleic acid pair (CKSNAP)	SVM

Table 7.1 categorizes existing RNA subcellular localization predictors in terms of SCLP and MCLP, where 5 MCLP have been developed for miRNA molecules and 2 predictors have

been developed for the mRNA molecules. A total of 6 SCLP have been developed for mRNA and 11 predictors have been developed for lncRNA biomolecules. One SCLP is developed for Circular RNA. It is evident from Table 7.1, most of the existing RNA subcellular localization predictors [65, 89, 132, 426, 437, 468] handle the problem of SCLP. However, these predictors are not effective to decode RNA association with various biochemical and pathological processes mainly happen through RNA concurrent presence in multiple compartments [403]. Furthermore, MCLPs for miRNA, lncRNA, mRNA and snoRNAs can be developed by utilizing publicly available databases that contain annotated localization information against these 4 different types of RNA molecules. However, public databases do not contain much MCLP information about circular RNAs that hinders the development of MCLPs. To best of our knowledge, there is only one generic multi-compartment localization predictor [403] for multiple RNA types (mRNAs, snoRNAs, miRNAs, lncRNAs) and species (*Homo sapiens*, *Mus musculus*). However, this approach is computationally expensive and relies on manually curated features which is why it lacks to produce promising performance for the subcellular localization prediction of different types of RNAs across multiple species.

Regardless of whether an existing predictor addresses the problem of SCLP or MCLP are not well generalized as they are designed to predict subcellular localization of one particular RNA type. Due to utilization of suboptimal feature extraction methods existing approaches are not powerful enough to handle different kinds of RNAs which vary in terms of sequence length, nucleotides composition, chemical structures and molecular interactions. Furthermore, majority of existing approaches are based on deep neural networks which are known as black box predictors as they do not explain which features are important for the accurate identification of subcellular compartment of particular RNA and species. The poor degree of model explainability hinders the researchers to accurately estimate the effects of diverse trade-offs in a model. Building on the need of a robust and explainable RNA subcellular localization predictor, this chapter presents another contribution of the dissertation, we develop an end-to-end deep learning approach "EL-RMLocNet" [23] for multi-compartment localization prediction of 4 different RNAs (mRNAs, snoRNAs, miRNAs, lncRNAs) across 2 distinct species (*Homo sapiens*, *Mus musculus*). It presents novel approaches to optimize multi-compartment subcellular localization predictive pipeline at different levels:

- This chapter presents a novel approach GeneticSeq2Vec to generate a statistical representation of RNA sequences. By treating nucleotide k-mers as vertices and their interactions as edges, GeneticSeq2Vec captures heterogeneous relations of vertices to generate k-hops proximity matrices. The k-hops proximity matrices are decomposed to generate the most informative components based on precise representation. It concatenates k-hops precise representation to encode nucleotide k-mers translational invariance, their local and global interaction patterns and correlations with target RNA in statistical sequence vectors.
- Considering accurate subcellular localization prediction of target RNA class and species

relies on the most relevant features, EL-RMLocNet [23] makes use of Long Short Term Memory (LSTM) and attention mechanism to find the most discriminative features and their heterogeneous dependencies.

- To better illustrate the decision making of EL-RMLocNet approach and quantify the practical significance. EL-RMLocNet [23] performs reverse engineering to map the weights of statistical feature space to nucleotide k-mers patterns for 4 different RNA classes (mRNA, snoRNA, miRNA, lncRNA) and 2 species (Homo sapiens, Mus Musculus).
- To objectively evaluate the efficiency and generalizability of EL-RMLocNet approach, we perform a comprehensive performance comparison of proposed EL-RMLocNet with state-of-the-art RNA associated subcellular localization predictor across 4 different RNA classes (Homo sapiens, Mus Musculus) and 2 species (Homo sapiens, Mus Musculus).
- To enable the scientific community to infer RNA subcellular localization on the go, we develop an interactive and user-friendly web server which is publicly available at https://rna_subcellular_predictor.opendfki.de/.

7.2 Materials and Methods

This section describes different modules of proposed EL-RMLocNet approach and benchmark datasets used to evaluate the performance of proposed approach.

7.2.1 Proposed EL-RMLocNet Approach

Working paradigm of proposed EL-RMLocNet approach can be categorized in two distinct phases. Firstly, k-mer embeddings are generated in an unsupervised manner using graph based approach which is explained in section 7.2.2 In second stage, an explainable deep learning classifier makes use of generated pretrained k-mer embeddings and raw sequences to predict subcellular compartments. A comprehensive details of proposed classifier is illustrated in section 7.2.3.

7.2.2 A K-hop Neighbourhood Relation based Statistical Representation Scheme for RNA Sequences (GeneticSeq2Vec)

Considering the effectiveness of graph based representation learning approaches for a variety of Natural Language Processing [396] and Bioinformatics tasks [442] mainly due to their ability to capture comprehensive semantic information and translational invariance of words. We present a novel graph based approach GeneticSeq2Vec to generate a rich statistical representation of RNA sequences, complete working paradigm of which is summarized by the pseudo-code in Figure 7.2.

Generation of statistical representation of raw RNA sequences using the proposed GeneticSeq2Vec approach is mainly comprised of four steps: 1) an un-directed k-mer graph generation,

2) k-hop proximity matrices construction, 3) k-hop proximity matrices factorization, 4) k-hop representation concatenation. In the 1st step, sequences of particular RNA class (mRNA, snoRNA, miRNA, lncRNA) and species (Homo sapiens, mus musculus) are divided into nucleotide k-mers. Then, nucleotide k-mers of all the RNA sequences are concatenated to generate a nucleotide k-mers list. Using nucleotide k-mers list, unique nucleotide k-mer pairs are generated by rotating a window of 2 with the stride size of 1. To effectively model the correlations of nucleotide k-mers at different granularity, an un-directed graph $G = (V, E)$ is generated where the set of nucleotide k-mers are represented as vertices $V = \{v_i, v_j, \dots, v_z\}$ and their interaction as edges $E = \{e_{i,j}, \dots, e_{o,p}\}$ primarily treating nucleotide k-mer pairs collection as connection reference. To perform computational analysis of $V * V$ sized un-directed graph G , a numerical representation of the graph G is generated through an adjacency matrix $S \in \mathbb{R}^{|V| * |V|}$ where $S_{i,j} = 1$ as well as $S_{j,i} = 1$ if there is an edge $e_{i,j}$ between vertex v_i and vertex v_j . On the other hand, if there is no edge between vertex v_i and vertex v_j then $S_{i,j} = 0$ and $S_{j,i} = 0$, revealing each entry in adjacency matrix indicates whether the pair of vertices have any association.

With an aim to capture proximity which measures diverse relational information and semantic closeness of one vertex to another vertex, in 2nd step, it transforms adjacency matrix into proximity matrix by performing multiple operations. Firstly, by computing the summation of every row of adjacency matrix S , a normalized adjacency matrix $X \in \mathbb{R}^{|V| * |V|}$ is generated. To match the size of adjacency matrix S , normalized adjacency matrix X is extended to the size $|V| * |V|$ by repeating its only row. Afterward, using Equation 7.1, transition probability of each vertex v_i to its immediate neighbouring vertex is computed to produce proximity matrix A , where $A_{i,j}$ is the transition probability from vertex v_i to its immediate neighboring vertex v_j .

$$A = \log \frac{\text{adjacencymatrix}(S)}{\text{normalizedadjacencymatrix}(X)} - \log \frac{1}{\text{vertexvocabularysize}(\beta)} \quad (7.1)$$

The proximity matrix A is multiplied by an identity matrix to generate a first-order (1-hop) proximity matrix A^1 . The first-order (1-hop) proximity matrix A^1 models whether there exists a direct connection between vertices by modeling the pairwise closeness between vertices. In Figure 7.1, analysis of the edges connecting different vertices within the boundary of red dotted circle reveals that first-order proximity (1-hop) captures two kinds of information: 1) vertex A1 is directly connected to vertex A2 as well as vertex A3, 2) vertex A1 and vertex A2 has strong relation represented with thick line and vertex A1 and vertex A3 has weak connection represented with thin line. By extending this paradigm to all vertices pairs present in the vocabulary, first-order (1-hop) proximity matrix captures the most fundamental relation between vertices. Considering, the extraction of information regarding whether two vertices are directly connected (1-hop) is not sufficient to capture heterogeneous relations of k-mer vertices. Hence, it is important to capture higher-order (k-hop) proximity which can effectively model the complex relationships of vertices. More specifically, the second-order (2-hops) proximity information A^2 captures the common neighbors among two vertices, the more neighbors are shared among vertices, the stronger the

connection is. In Figure 7.1, analysis of the vertices connection within the boundary of green dotted circle indicates that, vertex A1 and A2 has 4 common neighbors (B1, B2, B3, B4), hence these vertices have a far more stronger connection as compared to A2 and A3 vertices which have only common neighbour (B5). This paradigm is extended to all vertices pairs to generate second-order (2-hops) proximity matrix. Clearly, second-order (2-hops) proximity information is important to determine the strength of vertices connection on the basis of number of common neighbors, extracting key nucleotide k-mers information such as most frequently co-occurring nucleotide k-mers as well as common contexts.

Further, the third-order (3-hops) proximity information A^3 is essential to measure the impact of common neighbors on the strength of long range connection between vertices. In Figure 7.1, analysis of the trajectory A1-B-C-A2 within the boundary of aqua color dotted circle reveals that despite the strong connection among vertex A1 and vertex B, the connection between vertex A1 and A2 can be significantly weakened because of two weaker connections between vertex B and vertex C as well as vertex C and vertex A2. On the contrary, the trajectory A1-B- C_i indicates that the relationship between vertex A1 and A2 remains very strong primarily due to the decent number of common neighbors between vertex A2 and vertex B which greatly strengthens their relationship. Likewise, the fourth-order (4-hops) proximity information is also crucial to capture global relations of vertices. In Figure 7.1, analysis of the vertices connection inside the boundary of orange dotted circle reveals that the relation between vertex A1 and vertex A2

remains very strong because their connection partners B1 and B2 have four common neighbors D1-to-D4 which strengthens the relation of vertex A1 and A2. On the other hand, vertex A1 and vertex A2 become totally unrelated if we only consider their relation with vertex D5 and vertex D6, respectively mainly, because no path is left which connects vertex A1 to vertex A2. By extending the paradigms of third-order (3-hops) and fourth-order (4-hops) proximity to all possible vertices trajectories, global relations of the vertices can be captured in 3-hops and 4-hops proximity matrices which corresponds to long range contextual information of nucleotide k-mers.

It is evident from a thorough analysis of high order (k-hops) proximity modeling that each higher order (k-hop) proximity matrix captures different kind of relations among k-mer vertices.

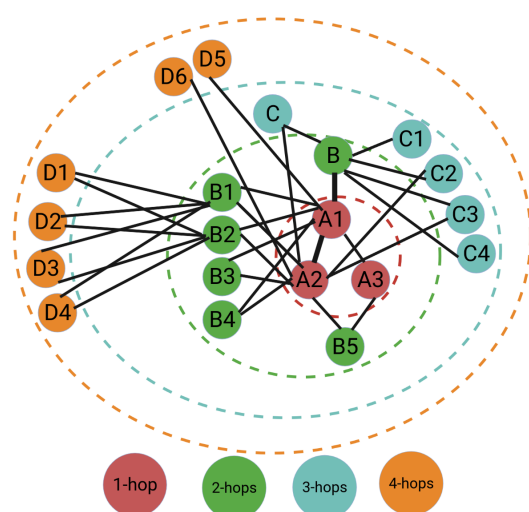


Figure 7.1: Illustration of k-order (K-hop) proximity information, red dotted circle represents first-order proximity (A^1), green dotted circle indicates second-order proximity (A^2), aqua dotted circle represents third-order (A^3) proximity and orange dotted circle indicates fourth-order proximity (A^4).

Therefore, instead of mapping heterogeneous nucleotide k-mer relations in a common subspace, GeneticSeq2Vec generates k-hop proximity matrices to retain heterogeneous relational information in different subspaces. Considering sequences vary across different RNA subtypes in terms of sequence length, nucleotide k-mer distribution, the idea of generating different subspaces helps to find optimal value of k-hop proximity for each RNA subtype as it avoids the influence of higher order proximity modeling to lower order proximity modeling. Building on, first order (1-hop) proximity matrix A^1 is computed through the multiplication of proximity matrix A to an identity matrix. Higher order (k-hop) proximity matrices A^k can be computed by multiplying the proximity matrix A k-times to itself.

Algorithm 1: A K-hop Neighbourhood Relation based Statistical Representation Scheme for RNA Sequences

Input:
k-mer pair collection
maximum value of hop k
Vertex Vocabulary size β
Dimension of representation vector d

1. Generate an undirected k-mer Graph G
Generate adjacency matrix of the graph S
2. Generate normalized adjacency matrix X
Compute basic proximity matrix (A)
 $A = \log(S/X) - \log(1/\beta)$
Calculate A^1, A^2, \dots, A^K respectively
Get each k-hop representations

for $K = 1$ to K **do**

if $K=1$ **then**
 $A^k = A * Identitymatrix(I)$
Construct the representation vector W^k

else
Calculate higher order proximity matrix
 $A^k = [A.A.A\dots]^k$
Construct the representation vector W^k

3. Factorizing higher order proximity matrix
 $U^k \Sigma^k, (V^k)^T$
 $= SVD(A^k)$
 $W^k = U_d^k (\Sigma_d^k)^{1/2}$

4. **Concatenate all the k-hop representations**
 $W = [W^1, W^2, \dots, W^k]$

Output: Matrix of the graph representation W

Figure 7.2: A k-hop neighbourhood relation based statistical representation scheme for RNA sequences

Where the proximity from the vertex v_i to v_j is mainly an entry in i^{th} row and j^{th} column of k-order (k-hops) proximity matrix A^k . The k-hop multiplications of proximity matrix A help to capture diverse interactions and global relations of the vertices, indicating higher order

proximity matrices encode translational invariance information of nucleotide k-mers to generate heterogeneous context aware representations. More specifically, the 2nd step produces k-hop representation matrices $W_b, W_c, \dots, W_k \in \mathbb{R}^{|V| \times |d|}$ for the input graph G where the i^{th} row of each W_i represents a continuous value vector of d dimension for the nucleotide k-mer vertex v_i learned by modeling its proximal k-hop relations with respect to all nucleotide k-mer vertices present in the vocabulary.

In 3rd step, proposed GeneticSeq2Vec factorizes proximity matrices produced by different k-hops using Singular Value Decomposition (SVD) approach in order to learn precise k-hops representation matrices $W_b, W_c, \dots, W_k \in \mathbb{R}^{|V| \times |d|}$. Using equation 7.2, SVD decomposes each k-hops proximity matrix into the product of three matrices, two of them U and V are orthogonal matrices and Σ serves as a diagonal matrix which is comprised of an ordered set of singular values.

$$W^k = U^k \sum (V^k)^T \quad (7.2)$$

Finally, in 4th step, it combines the precise representation produced by different k-values to generate k-order (k-hops) relations aware representations of all vertices, which can be expressed as follows:

$$W = [W^1, W^2, W^3, \dots, W^k] \quad (7.3)$$

7.2.3 Explainable Deep Learning based RNA Associated Multi-Compartment Localization Predictor

To accurately predict subcellular localization patterns of different RNA classes in multiple species, we have developed an explainable deep learning classifier "EL-RMLocNet". EL-RMLocNet leverages the stochastic embedding layer to optimize the embedding matrix generated through GeneticSeq2Vec approach. It uses LSTM to find and retain most informative features as well their long range dependencies from statistical vectors of RNA sequences. Unlike a trivial recurrent neural network (RNN), LSTM does not face the problem of vanishing gradients because it utilizes a gating mechanism to regulate the flow of information. The distribution of nucleotide k-mers vary across sequences of different RNA types and classes, indicating accurate subcellular localization of target RNA classes rely on certain set of nucleotide k-mers patterns. EL-RMLocNet captures potential nucleotide k-mers patterns using attention mechanism which weights the features on the basis of their potential to accurately predict subcellular localization of target RNA classes. By revealing potential nucleotide k-mers patterns for different RNA classes and species, attention mechanism also makes the decision making of deep learning model quite transparent. To significantly reduce the classification error, predictive potential and generalizability of proposed classifier are optimized using multiple neural strategies such as normalization, dropout and learning rate decay. Considering, the performance of the deep learning model is largely influenced by different hyperparameters such as number of layers, learning rate, batch size, etc., we optimize hyperparameters using grid search and facilitate optimal values of different hyperparameters in

Table 7.2. Architecture of proposed deep learning model EL-RMLocNet is given in Figure 7.3 and details of various inherent layers are provided in following subsections.

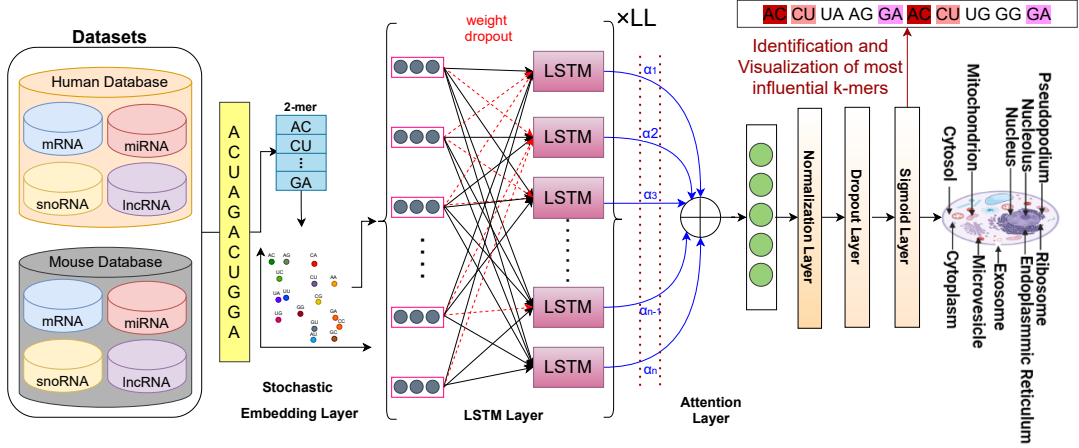


Figure 7.3: Workflow of an explainable deep learning model for RNA associated multi-compartment subcellular localization prediction

7.2.3.1 Stochastic Embedding Layer

The process of predicting RNA associated subcellular localization starts by dividing the RNA sequences into nucleotide k-mers by sliding a window of size w with the stride size of s . For every RNA sequence, statistical vector of each nucleotide k-mer is retrieved at the embedding layer mainly using embedding matrix of size vocabulary \times vector-dimensions produced by novel graph based representation learning module, discussed in section 7.2.2. To optimize embedding matrix, 2 distinct embedding dropout tricks are utilized in order to avoid model over-fitting which happens due to over-specialization of only few features. In k-mer embedding dropout, entire k-mer has the dropout probability of dp whereas in k-mer vector dimension dropout, each k-mer vector dimension has the likelihood of dp to be replaced by zero. Optimized d -dimensional statistical vectors of RNA sequences are obtained by averaging the respective k-mer statistical vectors. The d -dimensional RNA sequence vectors are passed to LSTM network having ll layers, ld hidden units which find and retain the most informative features along with their dependencies.

7.2.3.2 Optimized Long Short Term Memory (LSTM) Layer

Contrary to the traditional recurrent neural network, LSTM controls the information flow by making use of 3 distinct gates. Update gate or Input gate or update gate, indicated as \bar{I}_u (Equation 7.14) mainly regulates the flow of naive information in current time step. Forget gate, indicated as \bar{I}_f (Equation 7.15) decides whether memory information of last time step shall be dropped to taken forward. Third gate known as output gate is indicated by \bar{I}_o (Equation 7.16). It determines up to what extent information from previous time step will be transferred to next time step by

taking currently available information into account. In these mathematical expressions, $[W^i, W^f, W^o, U^i, U^f, U^o]$ refer to weight matrices, b_u, b_f, b_o indicate bias vectors, x_t represents d-dimensional nucleotide k-mer vector fed at particular time-step t , $t+1$ and $t-1$ refer to next and previous time steps, respectively, h_t refers to current hidden state, c_t indicates memory cell state and \odot represents element-wise product.

$$\bar{I}_u = \sigma(W^i .x_t + U^i .h_{t-1} + b_u) \quad (7.4)$$

$$\bar{I}_f = \sigma(W^f .x_t + U^f .h_{t-1} + b_f) \quad (7.5)$$

$$\bar{I}_o = \sigma(W^o .x_t + U^o .h_{t-1} + b_o) \quad (7.6)$$

$$cin_t = \tanh(W^c .x_t + U^c .h_{t-1}) \quad (7.7)$$

$$c_t = (\bar{I}_u \odot cin_t + \bar{I}_f \odot c_{t-1}) \quad (7.8)$$

$$h_t = (\bar{I}_o \odot \tanh(c_t)) \quad (7.9)$$

These 3 different gates mainly get activated or de-activated on the basis of corresponding weight matrices and behave on the basis of the corresponding activation function (e.g., sigmoid (σ), tanh). In equation 7.15, weight matrix W^f controls the working of forget gate. For example, if forget gate vector \bar{I}_f is completely zero, then c_{t-1} content will not be considered at all, indicating all information provided by the c_{t-1} will be discarded. Contrarily, if forget gate vector \bar{I}_f contains one, then the model preserves the information. These 3 different gates perform a variety of operations to regulate nucleotide k-mers information represented as a floating point vector falling in range of 0-to-1. Each cell of LSTM is comprised of these three gates. To preserve long term information of nucleotide k-mers, hidden state h of every cell is saved at each time step.

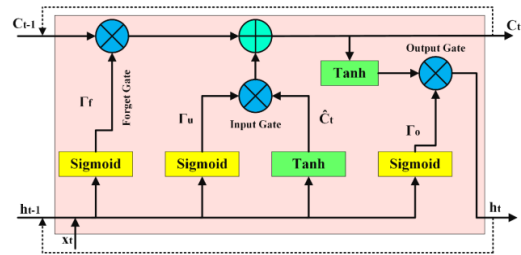


Figure 7.4: Information flow in standard LSTM cell

To regularize LSTM ll layers, considering, dropping hidden state of LSTM layers can significantly hinder the aptitude of LSTM to retain long term dependencies. We optimize LSTM layers by applying weight dropout on recurrent weight matrices $[U^i, U^f, U^o]$ as well non-recurrent weight matrices $[W^i, W^f, W^o]$ of LSTM layers where we randomly drop subset of weights in the network instead of dropping subset of activations. While weight dropout on recurrent weights avoid overfitting on the recurrent connections of LSTM layers, weight dropout on non-recurrent weight matrices enhance the LSTM ability to extract important residue dependencies. In this manner, LSTM layers produce d-dimensional feature vectors for RNA sequences which are passed

to an attention layer.

7.2.3.3 Attention Layer

One of the most important feature of the human perception is its ability to focus on only the most important parts of the input to make sense of the information present in outside world. Similarly, the significance of various nucleotide k-mers patterns for accurate RNA associated subcellular localization prediction varies across RNA classes and species, some nucleotide k-mers patterns are more discriminative while others are completely redundant. Considering, accurate multi-compartment subcellular localization prediction of various RNA classes and species mainly depends on the set of most relevant features. We utilize attention paradigm to optimize input d-dimensional RNA sequence vectors by weighting the features on the basis of their importance for hand on task.

The workflow of attention paradigm involves the generation of attention weights and optimize input features using attention weights is summarized in the Figure 7.5. First of all, we map the input d-dimensional LSTM feature vectors represented as x^t to h_t using Equation 7.10, where f_1 refers to nonlinear activation function and $h_t \in R^s$ represents hidden state at the time step t with size s .

$$h_t = f_1(h_{t-1}, x_t) \quad (7.10)$$

In order to avoid the issue of long-term dependencies which can significantly derail multi-compartment subcellular localization prediction performance, we utilize LSTM as nonlinear activation function f_1 . Then attention mechanism is developed using a deterministic attention based deep learning model. For a particular sequence $x^k = (x_1^k, x_2^k, \dots, x_m^k)^T \in R^m$, using previous hidden state represented as h_{t-1} as well as cell state c_{t-1} within LSTM cell, α_t^k and β_t^k can be defined using Equation 7.11 and Equation 7.12, respectively:

$$\alpha_t^k = v^T \tanh(W_1 * [h_{t-1}, C_{t-1}] + W_2 x^k) \quad (7.11)$$

$$\beta_t^k = \text{softmax}(\alpha_t^k) = \frac{\exp(\alpha_t^k)}{\sum_{i=1}^n \exp(\alpha_t^k)} \quad (7.12)$$

In these equations, matrices W_1, W_2, W_3, \dots and v are hyperparameters of the attention model that can be learned through backpropagation. The α_t^k vector is of length m where i^{th} value estimates the significance of k^{th} given feature sequence for a particular time step t . These values

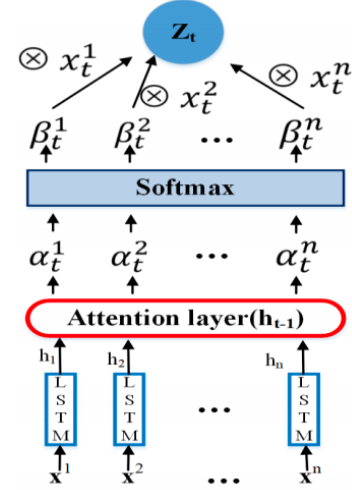


Figure 7.5: Architecture of the Attention model

are normalized through softmax. Whereas β^k represents attention weight that contains a value indicating the amount of attention should be placed on k^{th} input feature sequences. Output produced by attention model can be obtained at a particular time step t where the weighted and optimized input feature sequence represented as z_t will be equivalent to (Equation 7.13):

$$z_t = (\beta_t^1 x_t^1, \beta_t^2 x_t^2, \dots, \beta_t^n x_t^n)^T \quad (7.13)$$

By replacing the normal d -dimensional LSTM feature vector x_t with z_t and updating attention model, we manage to obtain optimized attention based feature vectors for RNA sequences. Unlike x_t where all input features are treated equally, z_t assigns higher weights to the most potential features effectively by eliminating the impact of redundant features for target RNA associated subcellular localization prediction. Optimized ad dimensional attention based feature vectors are passed forward in the network.

7.2.3.4 Bag of Tricks for Optimizing the Training and Prediction of EL-RMLocNet Approach

To optimize the training of deep learning model EL-RMLocNet, 3 distinct optimization tricks are utilized. The ad dimensional vectors produced by attention layer are passed to the normalization layer [202]. Normalization addresses the issue of co-variance shift which de-stabilizes the neural network by standardizing the input before feeding it to a hidden layer for every batch. It ensures that input-to-output mapping of a neural network does not overspecialize one particular region of protein sequences, resulting in faster training, convergence and improved generalizability [202].

Equation 7.14 describes the overall paradigm of normalization which normalizes each sequence x_i by tuning 2 parameters γ and β .

$$Y_i = BN_{\gamma, \beta}(x_i) \quad (7.14)$$

Equation 7.15 illustrates the way mean of a given batch is computed where x_i represents the current sequence from m sequences present in a given batch b .

$$u_b = 1/m \sum_{i=1}^m (x_i) \quad (7.15)$$

Equation 7.16 describes the way variance of every batch b is computed where each sequence x_i is subtracted from the mean of entire batch (u_b) before aggregating and computing average using m number of sequences present in given batch b .

$$0_b^2 = 1/m \sum_{i=1}^m (x_i - u)^2 \quad (7.16)$$

Equation 7.17 subtracts each sequence x_i from mean of the batch u_b and takes fraction by standard deviation to normalize the values between 0 and 1, which is represented with \hat{x}_i .

$$\hat{x}_i = \frac{x_i - u_b}{\sqrt{0_b^2 + \epsilon}} \quad (7.17)$$

In order to enable the network to adapt mean and variance of distribution, 2 parameters γ and β are learned and updated along with biases and weights during training. Final, normalized, scaled and shifted version of hidden distribution can be represented using equation 7.18.

$$y_i = \gamma * \hat{x}_i + \beta \quad (7.18)$$

Further, we also apply traditional dropout to avoid model overfitting occurred due to neuron co-adaptation where neurons stop operating independently and rely on other neurons to make decisions. Through random sampling based on the Bernoulli distribution (Equation 8.6), we apply traditional dropout on hidden neurons where each hidden neuron has the likelihood of dp to be dropped.

$$y = f(Wx) \bullet m, m_i \sim \text{Bernoulli}(p) \quad (7.19)$$

Considering choosing an optimal learning rate lr for deep learning model is not a straightforward task, another optimization trick used in proposed deep learning model EL-RMLocNet is learning rate decay. Learning rate decay trick smartly updates the learning rate in such a manner that global minima is computed and model converges to the best possible weights. By making use of adaptive moment estimation based on weight decay (ADAMW) optimizer, learning rate lr value is optimized using decay rate of ld during weight update, which can be mathematically expressed as:

$$w_{i+1} = w_i - 2\lambda w_i - \left\langle \frac{\delta L}{\delta w} \middle| w_i \right\rangle \quad (7.20)$$

Using one-hot encoded actual subcellular localization compartments, probability score s_i for each subcellular localization compartment present in benchmark dataset is computed through the application of sigmoid $f(s_i)$ before computing cross-entropy loss CE , which can be mathematically expressed as:

$$f(s_i) = \frac{e^x}{1 + e^x} CE = -t_1 \log(f(s_1)) - (1 - t_1) \log(1 - f(s_1)) \quad (7.21)$$

Using the batch size b through the process of backpropagation, proposed EL-RMLocNet predictor learns hyperparameters that facilitate accurately inferring the multi-compartment subcellular localization of various RNAs across multiple species.

7.2.4 Benchmark RNA-Associated Subcellular Localization Prediction Datasets

We collect 8 different RNA subcellular localization datasets belonging to Homo sapiens and mus musculus species from literature [403]. To prepare these datasets, Wang et al. [403] utilized a public metathesaurus RNALocate [464] to get raw sequences and subcellular localization information related to 4 RNA classes namely mRNA, miRNA, snoRNA and lncRNA.

Further, in each RNA class sequences which have more than 80% similarity were removed using CD-HIT tool. For 8 benchmark datasets, statistical distribution of 4 different RNAs in diverse subcellular compartments is provided in Figure 7.7. More specifically, 4 pie graphs in first row of the Figure 7.7 indicate the statistical distribution of mRNA, miRNA, snoRNA and lncRNA

sequences in multiple subcellular compartments for Homo sapiens species, whereas second row pie graphs reveal the statistical distribution of 4 different RNAs in diverse cellular compartments. Comparing the variations in sequence length across all 8 benchmark datasets indicates that all 4 RNA subtypes datasets have slightly longer sequences in Homo sapiens species as compared to mus musculus species.

Further, in order to analyze the variation in sequence length across all 8 benchmark datasets, donut chart in Figure 7.6 reports the minimum, maximum and average sequence length of 4 different RNA subtypes datasets for Homo sapiens species represented as *H_mRNA*, *H_miRNA*, *H_snoRNA*, *H_lncRNA* and for mus musculus species represented as *mRNA*, *miRNA*, *snoRNA*, *lncRNA*. For Homo sapiens species, *H_lncRNA* dataset contains the most lengthy sequences whose average length falls around 16,335 nucleotides. The *H_mRNA* dataset contains the second most lengthy sequences followed by *H_snoRNA* and *H_miRNA* dataset with average length of 3,675, 111 and 43 nucleotides, respectively. For mus musculus species, *lncRNA* dataset contains longer sequences followed by *mRNA*, *snoRNA* and *miRNA* dataset with average sequence length of 11,052, 3,547, 116 and 50 nucleotides, respectively.

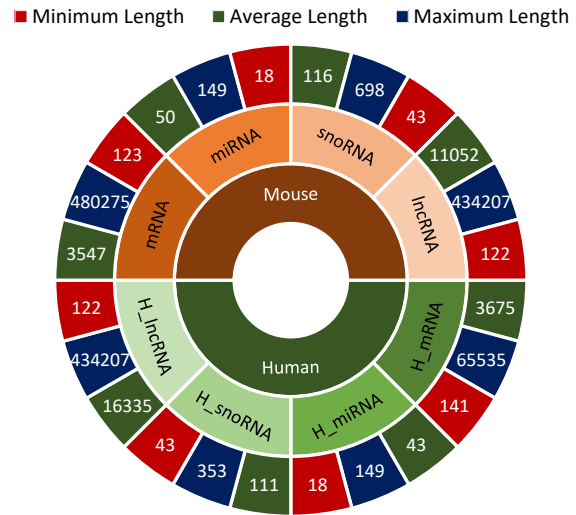


Figure 7.6: A comparison of variations in sequence length across 8 benchmark RNA associated multi-compartment subcellular localization datasets

7.3 Evaluation Criteria

Table 7.2: Optimal parameter values of proposed EL-RMLocNet approach for 8 benchmark datasets belonging to 4 different RNA classes and 2 species

Benchmark Dataset	K-mer	Stride Size (s)	Embedding Dimension (d)	Embedding Dropout (ed)	LSTM Layers (ll)	LSTM Hidden Units (ld)	Attention Dimension(ad)	Dropout (dp)	Learning Rate (lr)	Learning Rate Decay (ld)	Batch Size (b)
Homo sapiens species											
mRNA	3	2	200	0.005	1	200	50	0.01	0.05	0.001	32
miRNA	1	1	32	0.0025	1	32	60	0.005	0.06	0.1	32
snoRNA	2	2	64	0.0025	1	64	50	0.005	0.06	0.01	32
lncRNA	2	2	200	0.005	1	200	50	0.1	0.05	0.1	64
Mus Musculus species											
mRNA	2	1	200	0.0025	4	64	90	0.05	0.06	0.1	32
miRNA	1	1	32	0.0025	1	32	60	0.005	0.06	0.1	32
snoRNA	2	2	16	0.0025	1	16	50	0.005	0.06	0.0001	32
lncRNA	3	2	200	0.0025	4	60	50	0.05	0.05	0.01	128

In order to perform a fair performance comparison of proposed approach with existing state-of-the-art RNA multi-compartment localization predictor, 10-fold cross validation is performed. We use GridSearch [259] to optimize a variety of hyperparameters. To capture hidden pattern

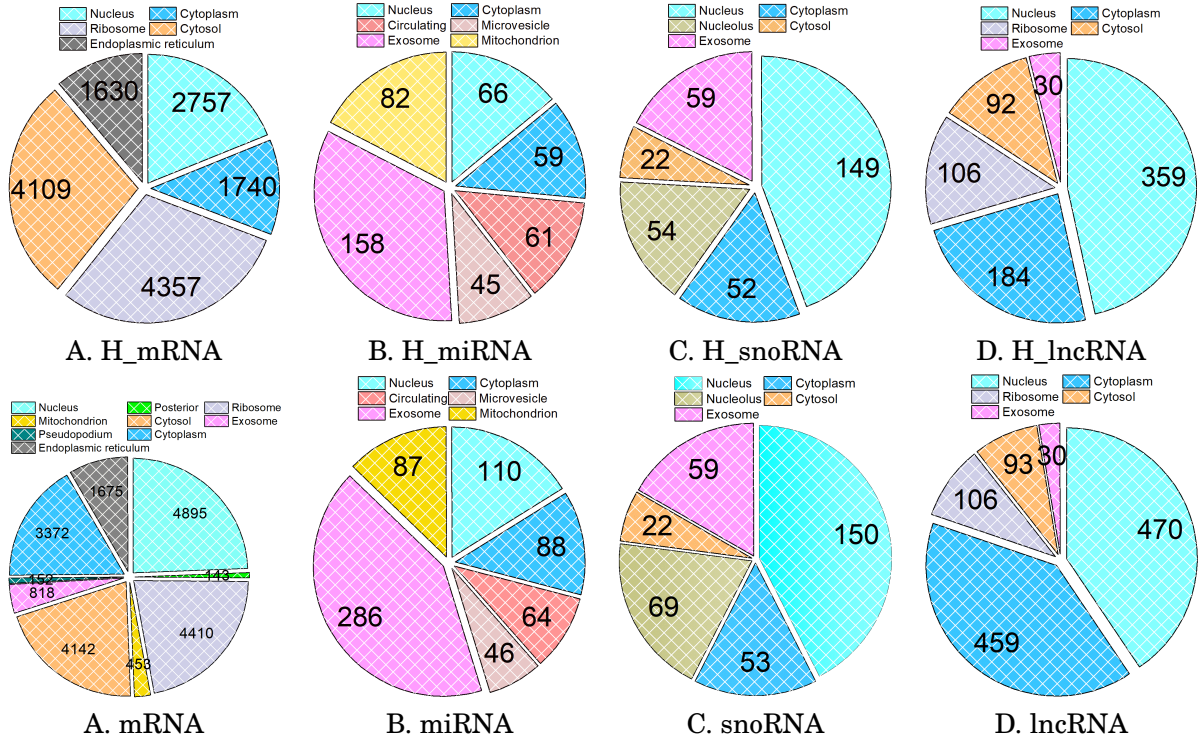


Figure 7.7: Statistical distribution of benchmark RNA associated multi-compartment localization prediction datasets for Homo sapiens (A-D) and Mus Musculus species (E-H)

of nucleotides, considering RNA sequences are comprised of only 4 unique bases, we perform experimentation with 5 different k-mers ranging from 1-to-5 generated using stride size of 1-to-3. To capture comprehensive relations and positional in-variances of nucleotide k-mers, novel k-hop neighborhood based statistical representation learning scheme performs experimentation with 2 to 7 hop based proximity matrices to generate rich d-dimensional vectors for RNA sequences.

Proposed EL-RMLocNet classifier is trained by tweaking an embedding dropout from 0.004 to 0.005, LSTM neurons from 100-to-400, batch size from 32-to-128, adaptive moment estimation based on weight decay (ADAMW) as an optimizer, learning rate from 0.04-to-0.05, decay rate from 1e-05-to-1e-07, standard dropout from 0.1-to-0.05 and categorical cross entropy as a loss function. Model checkpoint which achieves lowest training error is saved to make prediction on test sequences for the task of RNA subcellular localization prediction. To ensure the reproducibility of reported results, optimal values of different hyperparameters are summarized in Table 7.2.

7.4 Results and Discussions

This section quantifies the impact of 6 different sequence fixed-length generation approaches over the performance of the proposed EL-RMLocNet [23] approach for RNA multi-compartment subcellular localization prediction. Further, it performs a comprehensive assessment of the predictive

performance and generalizability of proposed EL-RMLocNet [23] approach for RNA associated multi-compartment subcellular localization prediction using a variety of evaluation metrics. It compares the performance of proposed EL-RMLocNet [23] approach with state-of-the-art RNA associated multi-compartment subcellular localization predictor using 8 benchmark datasets. It also performs intrinsic analysis of the key nucleotide k-mers patterns found by proposed approach EL-RMLocNet [23] to accurately predict the subcellular localization of different RNA classes in distinct species.

7.4.1 Performance Assessment of EL-RMLocNet for Multi-Compartment RNA Localization Prediction

It is evident from the donut chart 7.6 that in both Homo sapiens and mus musculus species, sequence length of all 4 RNA subtypes including mRNA, miRNA, snoRNA and lncRNA significantly differ from each other. Considering machine and deep learning classifiers operate on fixed-length genomic sequences, we perform experimentation with 6 different settings based on copy padding, sequence truncation and hybrid paradigms to fix the length of RNA sequences across all 8 benchmark datasets of 2 distinct species.

In copy padding paradigm, first of all, maximum possible sequence length is computed by comparing all the sequences of particular dataset. Afterward, all the sequences whose lengths are less than maximum threshold, are extended to justify maximum length by inserting a specific constant at starting or ending region of sequences. Another paradigm to fix the length of sequences is sequence truncation where first of all minimum possible sequence length is computed. Then, nucleotides from starting or ending region of all those sequences whose lengths are greater than minimum threshold are truncated in order to reduce the length up to minimum threshold. Considering copy padding paradigm may create an unnecessary bias to fade out discriminative sequence patterns and sequence truncation paradigm is vulnerable to lose important nucleotide distribution information. Hybrid paradigm first finds average sequence length and then utilize copy padding trick to fix the length of those sequences whose lengths are shorter than average length threshold and leverage sequence truncation trick for sequences whose lengths are greater than average length threshold.

Considering accurate RNA subcellular localization prediction relies on certain distributional patterns of nucleotides which can be present in any region of the sequences. We perform experimentation with all 3 sequence fixed-length generation paradigms using 6 different settings. Table 7.3 quantifies the impact of 6 different sequence fixed-length generation settings over the performance of proposed EL-RMLocNet [23] approach in terms of average precision. In Table 7.3, 2 settings related to copy padding are represented as *start_max*, *end_max*, sequence truncation settings are shown as *start_min*, *end_min* and hybrid paradigm settings are shown as *start_average*, *end_average*, where the setting names reveal the region of the sequences targeted for extension or truncation along with length threshold criteria. As evident from the

Table 7.3: Comparative analysis of 6 different fixed-length sequence generation approaches based on proposed EL-RMLocNet [23] approach over 8 benchmark datasets of 2 different species in terms of average precision

RNA Subtype	Sequence Length Variation					
	Start_Max	End_Max	Start_Average	End_Average	Start_Min	End_Min
Homo sapiens						
mRNA	0.72	0.70	0.77	0.72	0.73	0.71
miRNA	0.85	0.86	0.85	0.84	0.77	0.77
lncRNA	0.83	0.84	0.83	0.84	0.82	0.85
snoRNA	0.77	0.83	0.80	0.78	0.80	0.80
Mus Musculus						
mRNA	0.66	0.65	0.71	0.68	0.60	0.63
miRNA	0.86	0.87	0.86	0.86	0.84	0.83
lncRNA	0.73	0.70	0.77	0.73	0.72	0.69
snoRNA	0.82	0.81	0.82	0.81	0.80	0.81

Table 7.3, for *Homo sapiens* species, from both copy padding settings, EL-RMLocNet [23] approach achieves better average precision with *end_max* setting across all RNA subtypes except *H_mRNA* where *start_max* setting performs better. A similar performance trend can be seen with sequence truncation settings where EL-RMLocNet [23] attains better average precision with *end_min* as compared to *start_min* across most RNA subtypes. Unlike copy padding and sequence truncation settings, from 2 hybrid paradigm settings, EL-RMLocNet [23] approach produces better average precision with *start_average* across all RNA subtypes except lncRNA where its counterpart setting performs better. Overall, EL-RMLocNet [23] achieves peak performance with *end_max* setting for miRNA and snoRNA biomolecules, with *start_average* for mRNA biomolecule and with *end_min* for lncRNA biomolecule, obtaining the average precision of 86%, 83%, 77% and 85%, respectively. This indicates that all 3 sequence fixed-length generation paradigms (copy padding, sequence truncation and hybrid) manage to achieve good performance for one or the other RNA multi-compartment subcellular localization prediction.

Analyzing the performance trends for *mus musculus* species (Table 7.3) indicates that from 2 copy padding settings, EL-RMLocNet [23] approach achieves superior average precision using *start_max* for 3 RNA subtypes including mRNA, lncRNA and snoRNA attains better performance using *end_max* for miRNA biomolecule. Whereas from 2 sequence truncation settings, EL-RMLocNet approach produces good performance with *start_min* setting for miRNA and lncRNA biomolecules and with *end_min* for mRNA and snoRNA biomolecules. Contrarily, from 2 hybrid paradigm settings, EL-RMLocNet approach produces better average precision with *start_average* setting as compared to *end_average* setting across all 4 different RNA subtypes. Overall, EL-RMLocNet approach achieves peak performance with *start_average* setting for mRNA, lncRNA biomolecules, with *end_max* for miRNA biomolecule and with *start_max* for snoRNA biomolecule, obtaining the highest average precision of 71%, 77%, 87% and 82%, respectively.

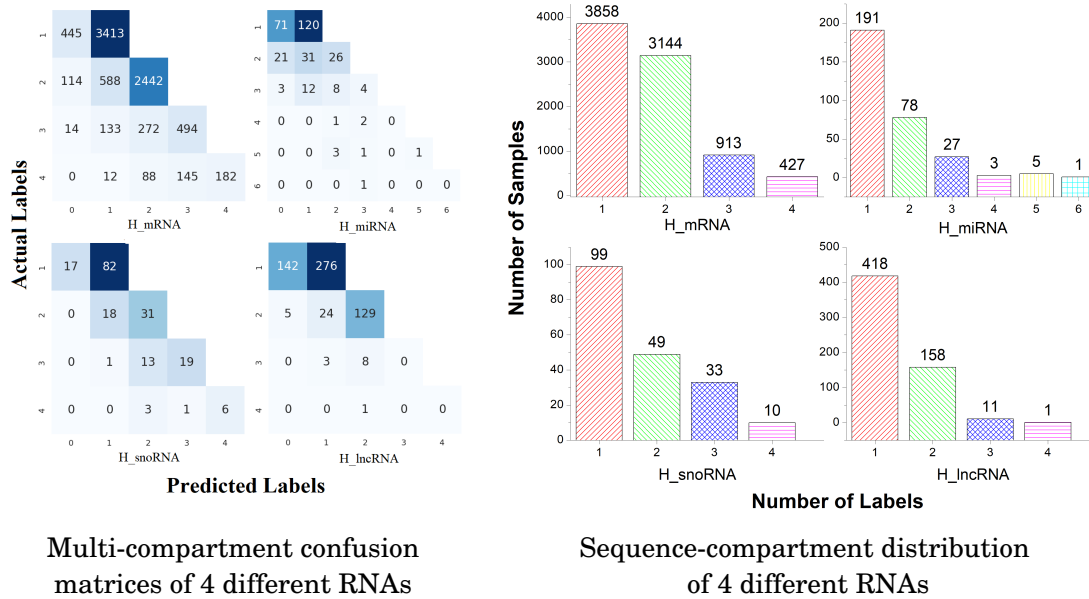


Figure 7.8: Multi-compartment localization prediction performance produced by EL-RMLocNet on 4 benchmark Homo sapien datasets of mRNA, miRNA, snoRNA and lncRNA corresponding to unique sequence-compartment distribution

Further, to analyze up to what extent EL-RMLocNet approach manages to correctly predict various combinations of subcellular compartments on account of heterogeneous subcellular compartment cardinality across 8 different benchmark datasets, multi-compartment confusion matrices along with sequence-to-compartment distributions bar graphs for Homo sapiens species and mus musculus species are given in Figure 7.8 and Figure 7.9, respectively. We leverage one-versus-rest strategy in order to generate confusion matrices across all 8 benchmark datasets where false negatives (fn), false positives (fp), true negatives (tn) and true positives (tp) are computed by considering one subcellular compartment as positive and other subcellular compartments as negative. By averaging fn, fp, tn and tp using total number of available subcellular compartments, confusion matrix for target RNA associated subcellular localization dataset is computed. This is primarily to assess the robustness of EL-RMLocNet when positive subcellular compartment has few number of RNA sequences and negative subcellular compartment has large number of RNA sequences.

From accuracy confusion matrices (Figure 7.8) produced by proposed EL-RMLocNet approach for Homo sapiens species, performance analysis for mRNA multi-compartment localization prediction indicates that, from 3,858 uni-compartment RNA sequences, subcellular localization of 3,413 sequences are correctly predicted by proposed EL-RMLocNet approach, indicating over 88% of uni-compartment RNA sequences are correctly predicted. From 3,144 bi-compartment RNA sequences, 2,442 RNA sequences are correctly classified into 2 cellular compartments, making it 78% of total bi-compartment sequences. For tri-compartment and tetra-compartment cardinalities, almost 54% and 43% RNA sequences of respective cardinalities are correctly classified

in appropriate subcellular compartments. For Homo sapiens miRNA subcellular localization, 63% of total uni-compartment, 33% of total bi-compartment and 15% of total tri-compartment RNA sequences are accurately categorized in respective subcellular localization compartments by EL-RMLocNet approach. For Homo sapiens snoRNA subcellular localization prediction, EL-RMLocNet approach accurately categorizes 83% of uni-compartment 63% of bi-compartment, 58% of tri-compartment and 60% of tetra-compartment RNA sequences. Further, for lncRNA multi-compartment subcellular localization prediction, 66% of uni-compartment and 82% of bi-compartment RNA sequences are correctly predicted. Whereas, no tri-compartment or tetra-compartment RNA sequence is correctly classified in respective subcellular compartment by EL-RMLocNet approach.

It is evident that a significant number of genomic sequences having different subcellular compartment cardinalities are accurately predicted by EL-RMLocNet approach across different RNA classes. Overall, for Homo sapiens species, EL-RMLocNet achieves better performance on mRNA followed by snoRNA, lncRNA and miRNA biomolecules. It manages to correctly predict 88% of mRNA uni-compartment, 82% of lncRNA bi-compartment, 58% of snoRNA tri-compartment and 60% of snoRNA tetra-compartment RNA sequences. Unlike existing RNA associated multi-compartment localization predictors whose performance significantly drops on account of different sized datasets as well as with the increase of subcellular compartment cardinality, proposed EL-RMLocNet approach shows promising performance across multiple datasets and shows robustness for different subcellular compartment cardinalities.

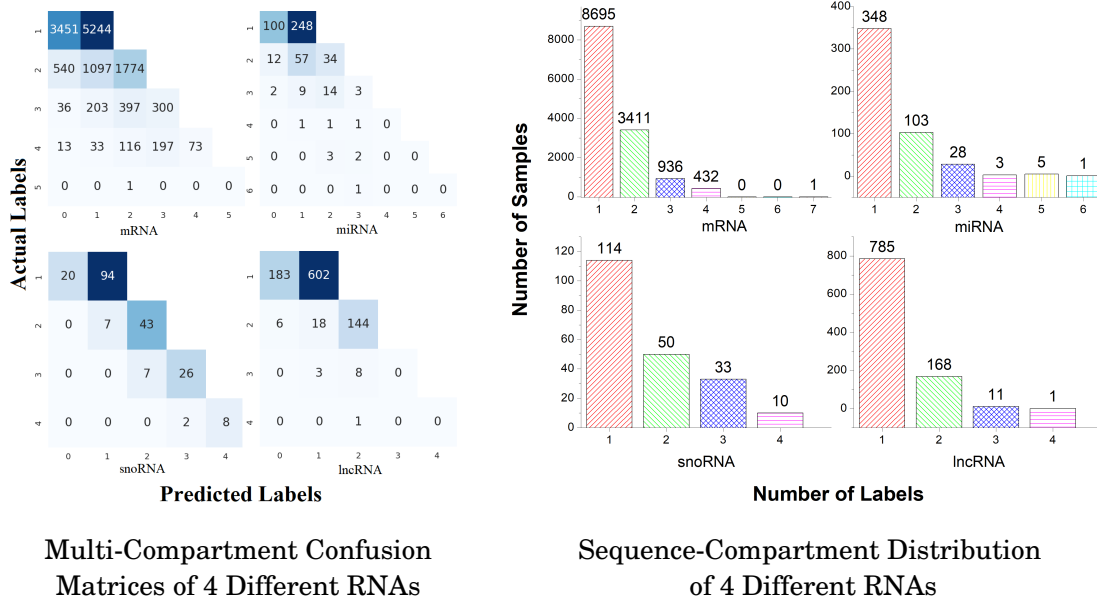


Figure 7.9: Multi-compartment localization prediction performance produced by EL-RMLocNet [23] on 4 benchmark Mus Musculus datasets of mRNA, miRNA, snoRNA and lncRNA corresponding to unique sequence-compartment distribution

Turning towards the accuracy confusion matrices produced by proposed EL-RMLocNet approach for 4 different RNAs belonging to mus musculus species, performance analysis of mRNA multi-compartment localization prediction indicates that from 8,695 uni-compartment RNA sequences, 5,244 are correctly predicted which makes up to 60% of uni-compartment sequences. Further, 28% of bi-compartment, 32% of tri-compartment and 17% of tetra-compartment RNA sequences are accurately inferred in respective cellular compartments. For miRNA subcellular localization, decent percentages of uni-compartment, bi-compartment and tri-compartment RNA sequences are accurately predicted which fall around 71%, 33% and 11%, respectively. For snoRNA subcellular localization prediction, 82% of uni-compartment, 86% of bi-compartment, 79% of tri-compartment and 80% of tetra-compartment RNA sequences are corrected predicted by EL-RMLocNet approach. Similarly, for lncRNA subcellular localization prediction, 77% of uni-compartment and 86% of bi-compartment RNA sequences are accurately predicted into respective localization compartments. To summarize accuracy confusion matrices performance across both species, it is easy to understand that unlike existing computational approaches whose performance decline on account of different species, proposed EL-RMLocNet achieves promising performance across all 4 different RNA classes. Contrary to Homo sapiens species, for mus musculus species, EL-RMLocNet achieves better performance on snoRNA followed by lncRNA, mRNA and miRNA biomolecules. It manages to accurately predict 82% of snoRNA uni-compartment, 86% of snoRNA and lncRNA bi-compartment, 79% of snoRNA tri-compartment and 88% of snoRNA tetra-compartment RNA sequences, revealing once again a promising robustness towards different subcellular compartment cardinalities.

In a nutshell, a comprehensive and multi-dimensional assessment indicates that proposed EL-RMLocNet approach marks promising performance for multi-compartment subcellular localization of 4 different RNAs across 2 different species. It achieves higher performance figures for mus musculus species for most RNA classes. While the novel approach based on the idea of using RNA-As-Graphs assists to capture comprehensive semantic and structural information of nucleotide k-mers. The gating mechanism of LSTM helps to find and retain long range dependencies of the features and attention mechanism assists to find most relevant features for target RNA class and species. By optimizing feature extraction and target specific subcellular localization prediction, proposed EL-RMLocNet manages to achieve promising performance over multiple different sized benchmark datasets for RNA associated multi-compartment subcellular localization prediction.

7.4.2 Comparison of EL-RMLocNet with Existing Multi-Compartment RNA Localization Predictors

Considering the significance of determining co-localization of biomolecules in multiple subcellular compartments for deep understanding of cellular biology and to develop diverse biochemical applications [19], Wang et al. [403] developed the state-of-the-art multi-compartment localization

predictor for 4 different RNA classes of 2 distinct species. They utilized 6 different nucleotide composition and statistics based sequence encoding schemes including nucleotide property composition, nucleotide k-mers composition, reverse compliment k-mer, nucleic acid composition, di-nucleotide composition, tri-nucleotide composition and composition of k-spaced nucleic acid pairs to adequately represent the nucleotide information present in RNA sequences. By fusing multivariate information using Hilbert-Schmidt independence criterion based multiple kernels learning, they found an optimal combined kernel for SVM classifier for multi-compartment localization prediction of mRNAs, miRNAs, snoRNAs and lncRNAs for home sapiens and mus musculus species.

Table 7.4 compares the performance produced by proposed EL-RMLocNet approach with stat-of-the-art approach [403] for the subcellular localization of 4 different RNAs (mRNAs, miRNAs, snoRNAs and lncRNAs) for home sapien species. As indicated by the Table 7.4, proposed approach EL-RMLocNet outperforms state-of-the-art approach [403] across all 4 benchmark datasets belonging to different RNAs in terms of 5 different evaluation measures. EL-RMLocNet achieves the average precision increment of 7%, 1%, 1% and 10% as compared to state-of-the-art [403] performance for miRNA, mRNA, snoRNA and lncRNA multi-compartment localization prediction. EL-RMLocNet improves state-of-the-art accuracy by 11%, 5%, 1% and 13% for miRNA, mRNA, snoRNA and lncRNA multi-compartment localization prediction. Performance analysis in terms of coverage, ranking loss and one-error where lower value indicates better predictive performance, EL-RMLocNet surpasses the previous best performance by a decent margin for all 4 RNAs across all evaluation metrics.

Table 7.4: Performance comparison of proposed EL-RMLocNet approach with state-of-the-art approach for multi-compartment localization prediction of miRNA, mRNA, snoRNA and lncRNA using 8 benchmark datasets of Homo sapiens (Human) and Mus Musculus (Mouse) species

Species	Datasets	Average Precision		Accuracy		Coverage		Ranking Loss		One error	
		State-of-the-art [403]	Proposed EL-RMLocNet [23]	State-of-the-art [403]	Proposed EL-RMLocNet [23]	State-of-the-art [403]	Proposed EL-RMLocNet [23]	State-of-the-art [403]	Proposed EL-RMLocNet [23]	State-of-the-art [403]	Proposed EL-RMLocNet [23]
Human	miRNA	0.79	0.86	0.52	0.63	1.46	0.70	0.17	0.11	0.29	0.26
	mRNA	0.76	0.77	0.41	0.46	1.69	0.68	0.24	0.23	0.37	0.35
	snoRNA	0.82	0.83	0.54	0.55	1.54	0.45	0.18	0.17	0.24	0.20
	lncRNA	0.75	0.85	0.42	0.55	1.18	0.45	0.22	0.17	0.37	0.20
Mouse	miRNA	0.79	0.87	0.58	0.69	1.31	0.50	0.18	0.10	0.31	0.28
	mRNA	0.70	0.71	0.34	0.37	1.71	0.87	0.14	0.13	0.44	0.40
	snoRNA	0.80	0.82	0.52	0.56	1.59	0.29	0.21	0.20	0.25	0.20
	lncRNA	0.76	0.77	0.43	0.47	0.95	0.60	0.19	0.18	0.40	0.36

Furthermore, performance comparison of proposed EL-RMLocNet approach with stat-of-the-art approach [403] for the subcellular localization of 4 different RNAs (mRNAs, miRNAs, snoRNAs and lncRNAs) for *Mus Musculus* species (Table 7.4) indicates that proposed EL-RMLocNet approach once again outperforms previous best performance across all 4 benchmark datasets in terms of five different evaluation metrics. EL-RMLocNet outperforms state-of-the-art average precision by 8%, 1%, 2% and 1% for miRNA, mRNA, snoRNA and lncRNA multi-compartment subcellular localization. In terms of accuracy, EL-RMLocNet outperforms previous best performance by 11%, 3%, 4% and 4% for all 4 miRNA, mRNA, snoRNA, lncRNA multi-compartment localization prediction. Similarly, performance analysis in terms of coverage, ranking loss and one-error reveals that EL-RMLocNet achieves lower error values across most evaluation metrics for all 4 different RNA classes.

To sum up, proposed EL-RMLocNet approach [23] achieves better performance across most datasets from 8 benchmark datasets belonging to 4 different RNAs and 2 species. Overall EL-RMLocNet achieves higher performance increment for *Homo sapiens* species as compared to *Mus musculus*. It outperforms stat-of-the-art approach [403] by an average accuracy figure of 8% for *Homo sapiens* species and 6% for *Mus musculus* species. Unlike traditional nucleotide frequency and physicochemical properties based sequence encoding schemes used by stat-of-the-art approach [403] which lacks to capture comprehensive relations of nucleotides. EL-RMLocNet uses a novel weighted graph based statistical representation learning scheme which treats nucleotide k-mers as nodes and their interactions as edges to better characterize nucleotide k-mers relations. Further, unlike machine learning based stat-of-the-art approach [403], proposed EL-RMLocNet makes use of a precisely deep neural network which utilizes gating mechanism to retain informative features and their dependencies and attention mechanism to find RNA class and species specific discriminative distribution of features to accurately predict target species RNA subcellular localization.

7.4.3 Visualization of Most Informative Nucleotide k-mers Patterns

Proposed EL-RMLocNet approach effectively predicts the subcellular localization of various RNAs mainly by finding the most discriminative features with the help of attention mechanism. The mapping of statistical feature space having certain attention weights to their corresponding nucleotides k-mers is essential to elaborate on which nucleotide k-mer distribution is most informative to accurately predict various subcellular compartments of target RNA subtype of particular species. The acquisition and interactive visualization of such information effectively interpret and explain the decision making of deep learning model, actualize the generalizability and practical significance of the model to facilitate biomedical researchers and practitioners. Considering, sequence length largely fluctuates across different RNA subtypes and species ranging from few hundreds of nucleotides to thousands of nucleotides. We visualize the importance given by attention mechanism of proposed EL-RMLocNet approach to nucleotide k-mer distribution

within the range of 100 nucleotides across all 8 benchmark datasets of 4 different RNAs (mRNA, miRNA, snoRNA and lncRNA) and 2 species (Homo sapiens, Mus Musculus) to avoid repetition of information and improve readability.

Considering, attention mechanism can even assign different weights to same nucleotide k-mer and same weight to different nucleotide k-mers depending on the short and long range contextual information. Figure 7.10 highlights nucleotide k-mer distribution of 4 different RNAs across 2 species on a gradient scale from light to darker shade of a specific color and size scale from shorter to larger fonts, indicating more darker and standout nucleotide k-mers are the most informative for target RNA subtype. For instance, for Homo sapiens lncRNA multi-compartment subcellular localization prediction, nucleotide bi-mer “CG” is the most informative across different distributions of nucleotides. For Mus Musculus miRNA multi-compartment subcellular localization prediction, nucleotide bi-mers CC,GC, AG, GG are most informative followed by AA and TT within certain nucleotide distributions. Similarly, for other RNA subtypes across both species, most informative and least informative nucleotide k-mers and their different nucleotide distributions (unique color shaded) are evident in the Figure 7.10. We believe that an interactive intrinsic analysis of various RNAs helps to identify the most appropriate degree of nucleotide k-mer (e.g., bi-mer, tri-mer), identify region containing most useful nucleotide k-mer distribution, providing a direction to optimize the performance and generalizability of various other RNA sequence analysis tasks.

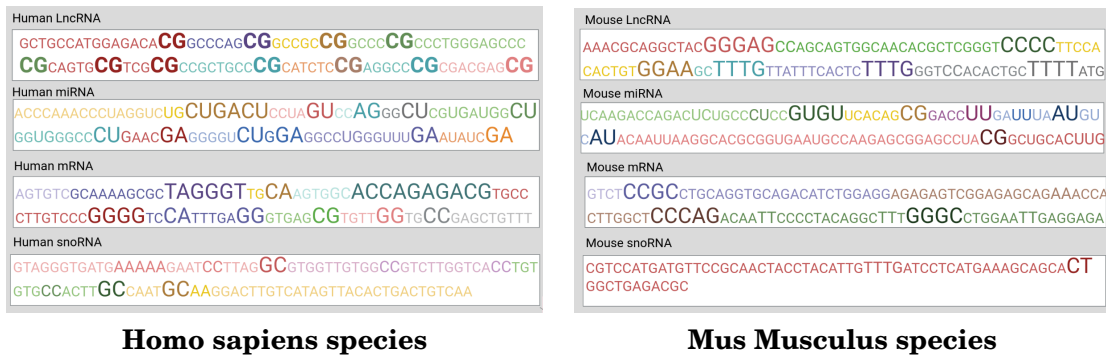


Figure 7.10: Most and least informative nucleotide k-mers patterns for 4 different RNAs belonging to Homo sapiens and Musculus species identified by attention layer of proposed EL-RMLocNet [23] approach

7.5 Conclusion

In this study, we establish an effective multi-compartment localization prediction landscape for 4 different RNA classes and 2 distinct species to better understand the functional dynamics of RNAs. Unlike existing computational approaches which lack to capture context of residues at different granularity while generating statistical representation of RNA sequences as well as

potential residue patterns important for accurate multi-compartment localization prediction. Our proposed approach EL-RMLocNet generates a comprehensive local and global residue contextual information aware statistical vectors of RNA sequences by treating RNA-As-Graph captures. It makes use of LSTM network to extract features, short and long range dependencies and attention mechanism to assign weights to the features on the basis of their importance for accurate multi-compartment localization of target RNA class. Visualization of important higher order residue patterns can assist researchers to draw important insights while comparing sequences of homogeneous or heterogeneous RNA classes. A comprehensive comparison of proposed EL-RMLocNet approach with state-of-the-art approach using 8 benchmark datasets of 4 different RNA classes and 2 distinct species proves that EL-RMLocNet is the first most effective generic yet explainable model for RNA multi-compartment localization prediction. We expect public availability of EL-RMLocNet will prove a valuable asset for subcellular localization prediction of various RNAs across multiple species, as well as an additional tool for the classification and localization prediction of other biomolecules.

Part II

Proteomics Sequence Analysis

PROTEIN-PROTEIN INTERACTION PREDICTION

Proteins are large and complex biomolecules that perform a multitude of crucial functions within living organisms mostly by interacting with other proteins [40]. Protein-protein interaction (PPI) analysis is important to understand diverse biological processes including cell proliferation [314], signal transduction [326], DNA transcription, replication [398, 462], hormone regulation [469], cycle control [231] and neuro-transmission [374]. It also helps to identify disease-related signaling pathways and symbolizes unfamiliar targets for therapeutic intervention [447]. In-depth exploration of PPIs is critical for a thorough understanding of protein functionalities, genetic mechanisms [7, 408], discovery of novel drug targets [16] and development of effective preventive or therapeutic strategies to combat diseases [330].

8.1 Related Work

A number of experimental approaches such as tandem affinity purification (TAP) [147], mass spectrometric protein complex identification [178], protein chips [477] and yeast two-hybrid (Y2H) [204, 230] have been utilized to infer PPIs. However, these experimental methods are expensive and time-consuming [357]. Furthermore, because of high specificity between proteins, these experimental approaches produce significant false positive results which mark the need for additional methodologies to cross-check the obtained results. Due to slow sequence analysis process, these approaches have been typically applied to identify intra-species PPIs, whereas inter-species interactome remained comparatively understudied [357]. Advancements in high-

⁰This chapter is an adapted version of the work presented in Asim et al., "CONR-NET: A Collection of Neural Refinements for protein-protein Interaction Prediction", In *iScience* (2022) and Asim et al., "BoT-Net: A Lightweight Bag of Tricks based Neural Network for Efficient lncRNA-miRNA Interaction Prediction", In *Interdisciplinary Sciences: Computational Life Sciences* (2022) [25]

throughput approaches and the influx of PPI data related to different species have given rise to many databases including the Database of Interacting proteins (DIP) [348], the Molecular Interaction Database (MINT) [260] and the Human protein References Database (HPRD) [327]. The public availability of such humongous annotated data has opened new horizons for the development of computational approaches for economical, fast and more accurate analysis of PPIs.

In order to predict PPIs to date, a plethora of computational approaches have been developed [213, 334] which can be broadly segregated into three classes: 1) Structure based, 2) Network based and 2) Sequence based. Structure based approaches estimate the likelihood of PPIs by leveraging primary and higher-level spatial structures like secondary, tertiary or quaternary structures [315]. Those proteins are more likely to interact in which compatibility levels of interacting regions are high or in which spatial structures more often appear on protein-protein binding-motif regions. [123, 315, 366]. Following this principle, Hue et al. [199] performed the pioneer work to predict PPIs in which they fed structural information of protein pairs to support vector machine (SVM) classifier. Zhang et al. [462] performed similar work by using protein structural information and Bayesian classifier for PPI interaction prediction. Hosur et al. [186] utilized protein structural information to compute the interaction confidence score for each protein pair using a boosting classifier. Structural information based PPI predictors neglect the mutual influence of local structures [28, 142]. Such approaches are more vulnerable to overlook important information for accurate PPI prediction which might be present in primary sequences and likely to get lost while extracting structural information [28, 142].

Network based PPI prediction approaches utilize the link information present in existing PPI networks. PPI networks are hierarchical illustrations of interacting proteins and exist in form of ontologies where each node represents a particular protein and interaction of two different proteins is represented by an association link. Proteins residing in upper hierarchy act as parents and their attached interacting partners of lower hierarchy act as child. Network based PPI prediction approaches extract the names of proteins from existing ontologies to find their biological characteristics in other resources and heterogeneous relations between proteins in order to predict interactions between unseen proteins on the basis of prior learning. Initial network based approaches considered that these proteins are more likely to interact which share more common interacting partners in PPI network [226]. However, these approaches have become obsolete after the discovery of Kovacs et al [226] that two proteins are more likely to interact if at least one of them is very similar to other's interacting partners. But Kovacs et al [226] approach has limited practical significance as it lacks to determine the interactions between the long distant proteins. To address this problem, Wang et al. [410] predicted PPIs without defining the length of different network paths in advance, however, their approach heavily relies on the quality of PPI network. Most recent paradigm of network based approaches considers that proteins of same functional module are more likely to interact as compared to the proteins of different functional module

[190]. Using the already known information of the functional modules, Hu et al [189] integrated biological information of proteins particularly gene ontology into PPI network to predict PPIs. Likewise, Ioan et al. [201] proposed attention based deep learning model which used graph based embeddings to learn deep semantic relations of gene ontology to distinguish interactive and non-interactive protein sequence pairs. A closer look at different network based PPI prediction approaches reveals that these approaches completely rely on pre-computed PPI networks and biological information, both of which need periodic updates to cater huge proteins related data produced by high throughput technologies. Furthermore, such resources are characterized by high false-positive as well as false negative rates which eventually hamper the performance of PPI predictors. Therefore, raw sequence based PPI prediction approaches are widely considered more appropriate to perform large scale PPI analysis.

To date, several raw sequence based machine and deep learning based approaches have also been proposed [450] for PPI prediction. For example, most recently, Yu et al. [450] proposed a machine learning based PPI predictor GcForest-PPI. It utilized amino acids composition information and physicochemical characteristics to generate statistical representation of protein sequences. It used Elastic Net [480] to extract a discriminative set of features that were passed to an ensemble classifier based on three different models namely XGBoost, Random Forest and Extra-Tree. GcForest-PPI achieved the accuracy of 95.44% and 89.26% on benchmark *S.cerevisiae* (*S.cerevisiae*) and *Helicobacter Pylori* (*H.pylori*) datasets. Kong et al. [223] presented another machine learning approach namely FCTP-WSRC. They utilized amino acid physicochemical properties, composition and transition information to generate statistical representations of protein sequences. They utilized principal component analysis to reduce redundant features and generate better feature space. Using reduced statistical representations and WSRC [223] classifier, they managed to achieve the accuracy of 86.73% and 78.70% on 2 benchmark *S.cerevisiae* and *H.pylori* datasets. Jia et al. [210] also proposed a machine learning based PPI predictor namely "iPPI-Esml". They combined amino acid composition information, physicochemical characteristics and protein chain specific wavelet transform information to generate statistical representations of protein sequences which were passed to a deep forest classifier. The iPPI-Esml approach achieved the accuracy of 95% on benchmark *S.cerevisiae* and 90% on *H.pylori* datasets.

Apart from machine learning based PPI predictors, Yao et al. [439] proposed a deep learning based predictor namely DeepFE-PPI. They utilized Word2vec based embedding generation approach [300] to generate statistical representations of protein sequences which were passed to a Multi-Layer Perceptron model for PPI prediction. DeepFE-PPI achieved the accuracy of 95% on benchmark *S.cerevisiae* dataset. Du et al. [116] presented DeepPPI which utilized amino acid's physicochemical properties to generate statistical representations of protein sequences. They utilized a Multi-Layer Perceptron model which extracted the high-level discriminative features from statistical vectors to make accurate PPI prediction. DeepPPI achieved the accuracy of 94%

and 86% on 2 benchmark *S.cerevisiae* and *H.pylori* datasets.

Critical analysis of machine and deep learning based PPI predictors (i.e., GcForest-PPI [450] WSRC [223], DeepFE-PPI [439]) reveals that amino acid composition or physicochemical properties based protein sequence encoding methods overlook the relationships that exist between different amino acid segments as a function of context of long protein sequences [210, 450]. Furthermore, selecting an optimal set of physicochemical properties from a huge available collection requires extensive empirical evaluation [210, 450]. Besides, concatenation of statistical representations generated through different types of encoding methods also gives birth to redundant features. To remove redundant features, existing PPI predictors [223, 450] utilize dimensionality reduction or feature selection approaches to generate an effective feature space. However, dimensionality reduction approaches generally prove inefficient for large and weakly non-linear data [69, 369]. Also, determining the number of principal components for the generation of compressed representation varies across different datasets, indicating that optimal principal components are found through comprehensive empirical evaluation. Similarly, major disadvantage of using elastic-net as a feature selection approach [450] is the high computational cost as one needs to cross-validate the relative weights of L1 and L2 penalty. Elastic-net leverages a combination of L1 and L2 penalties in order to shrink coefficient of un-important features to near zero, which is a computationally expensive and a time consuming process [349].

Furthermore, Word2vec [300] based PPI prediction approaches [439] also lack to generate an effective statistical representation of protein sequences. Because Word2vec [300] treats k-mers as atomic entities to generate their distinct vectors in which it neglects the distribution of amino acids within each k-mer. FastText [47] is an extension of Word2vec [300] where vector of each k-mer is computed by considering the distribution of k-mers and distribution of amino acids inside the k-mers. Also, our previous work [26] found that among three different neural embedding generation approaches namely: Word2Vec, FastText and Glove; FastText approach most effectively captures semantic information of k-mers.

We use FastText approach to generate comprehensive contextual information aware statistical vectors for k-mers present in protein sequences. Furthermore, we generate fixed-length protein sequences using six traditional and four novel fixed-length generation approaches. We propose a novel attention based deep hybrid model namely ADH-PPI, which makes best use of different neural network layers and optimization strategies for accurate PPI prediction. ADH-PPI makes use of Long Short-Term Memory, convolutional and attention layers to find the most discriminative features along with their short and long range dependencies important to effectively distinguish interactive protein sequence pairs from non-interactive protein sequence pairs. To avoid under-fitting and over-fitting, training of the ADH-PPI is optimized using different kinds of dropout, normalization and learning rate decay strategies.

A comprehensive empirical evaluation indicates that proposed ADH-PPI approach outperforms several machine and deep learning based PPI predictors across 6 different species bench-

mark datasets with a decent margin. To better describe the decisions of proposed ADH-PPI approach, we map the weights of statistical feature space to potential k-mer distributions which contribute the most to accurate PPI prediction through the reverse engineering strategy.

8.2 Materials and Methods

This section explains different modules of proposed predictor and describes protein-protein interaction prediction benchmark datasets.

8.2.1 Methodology of Proposed ADH-PPI Predictor

The working of the proposed ADH-PPI predictor can be categorized into three different modules. First module generates effective statistical representations of k-mers present in protein sequences by applying transfer learning in an unsupervised manner. Second module generates fixed-length protein sequences using traditional and novel sequence fixed-length generation methods. Using fixed-length protein sequences and k-mer embeddings, third module trains a novel attention based deep hybrid neural network for PPI prediction. A brief description of each module is provided in the following sub-sections.

8.2.1.1 K-mer Embedding Generation

To generate k-mer embeddings, first step is to divide the protein sequences into k-mers. Overlapping k-mers are generated by rotating a fixed-size window over a protein sequence where the stride size is always less than the size of window. On the other hand, non-overlapping k-mers are generated by rotating a window with a stride size equal to window size.

Protein sequences are made up of 20 distinct amino acids. Hence, in both overlapping or non-overlapping k-mer generation, the unique vocabulary size is equal to 20^k . The value of k determines the size of vocabulary which impacts model complexity, memory cost, run time cost, as well as up to what extent amino acid contextual information is taken into account, hence the choice of k is very crucial. Following the work of Le et al. [243] and Asim et al. [20], we generate different overlapping and non-overlapping k-mers by varying the window size from 2-to-7 and stride size from 1-to-7.

For different sizes overlapping and non-overlapping k-mers, we generate k-mers embeddings of different dimensions using FastText embedding generation model, working of which is graphically illustrated in Figure 8.1.

With an aim to capture comprehensive information of amino acids distributions, we take two benchmark *S.cerevisiae*, *H.pylori* datasets and four independent test sets in order to most effectively train the FastText model over large dataset of 26,886 protein sequences. For all 26,886 protein sequences, we generate overlapping and non-overlapping k-mers by varying the window size from 2-to-7 and stride size from 1-to-7. This produces number of different k-mers=6

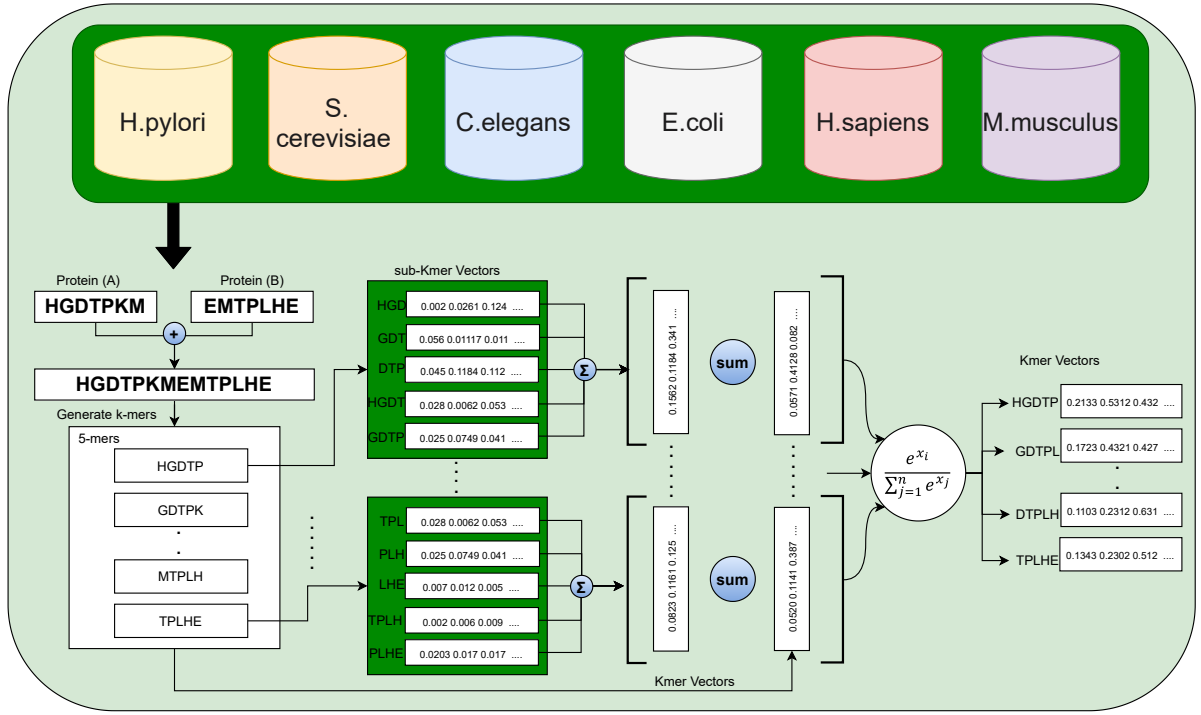


Figure 8.1: Workflow of unsupervised transfer learning applied using 6 datasets of distinct species to learn distributed representation of higher order sequence amino acids

× maximum possible different stride size = 6 equal to 36 different versions of protein sequences corpus based on different overlapping k-mers and 6 versions of protein sequences corpus based on different non-overlapping k-mers. For each version of protein sequences corpus, we train the FastText model to generate k-mer embeddings of different dimensions d ranging from 100, 120, 240, to 300. For example, by considering non-overlapping 3-mers, 26,886 protein sequences are divided in 3-mers which generates a vocabulary of 20^3 unique 3-mers and FastText generates d -dimensional statistical vectors for each 3-mer. FastText embedding generation model is an extension of Skipgram model [300]. Given a training k-mer sequence $k_1, k_2, k_3, \dots, k_T$, objective function of Skipgram model can be defined as follows:

$$J = \max \frac{1}{T} \sum_{t=1}^T \sum_{c \in C_t} \log p(k_c | k_t) \quad (8.1)$$

Where C_t represents the collection of surrounding k-mers of current k-mer k_t , given current k-mer k_t , $p(k_c | k_t)$ denotes the probability of observing its surrounding k-mer k_c .

$$p(k_c | k_t) = \frac{e^{s(k_t, k_c)}}{\sum_{j=1}^W e^{s(k_t, k_j)}} \quad (8.2)$$

Here $s(k_t, k_c)$ represents the scoring function. Skipgram model considers the scoring function as scalar product $s(k_t, k_c) = u_{k_t}^T v_{k_c}$, where u_{k_t} and v_{k_c} represent the vectors of two k-mers k_t and

k_c , respectively. However, Skipgram can only generate a distinct vector for each k-mer without exploiting their sub-kmer information. To overcome this problem, FastText represents a k-mer as a bag of sub-kmers. For instance, k-mer "HGDTP" will be represented by sub-kmers such as <#HGD, HGDT, GDTP, DTP#> and k-mer itself <HGDTP>. Unlike Skipgram model, FastText defines the scoring function $s(k_t, k_c)$ as the $\sum_{g \in (1, \dots, G)} z_g^T v_c$ where $(1, \dots, G)$ denotes the sub-kmers collection of k_t , z_g represents the vector of sub-kmer and v_c represents the vector of k-mer k_c . In this manner, FastText learns the embeddings of sub-kmers. Using sub-kmers embeddings, a k-mer embedding is learned as the sum of distributed representations of its sub-kmers. Major advantage of FastText embedding generation model is that it takes k-mer distributions as well as distributions of amino acids within k-mers into account to generate effective distributed representation of k-mers. Another advantage is that it shares the distributed representation of sub-kmers across all the k-mers which is extremely useful to generate optimal embeddings for less frequent k-mers. FastText embedding generation model is trained with an objective to maximize the probability of target k-mer over all k-mers present in the vocabulary using a softmax layer. Embedding matrix along with output layer parameters are learned by back propagating the error using stochastic gradient descent and negative sampling approach. Using FastText, we generate effective d -dimensional vectors for k-mers where the value of d is varied from 100, 120, 240, to 300.

8.2.1.2 Fixed-Length Generation of Protein Sequences

Exploratory analysis of 2 core PPI datasets (Figure 8.4) indicates that minimum sequence length for both *S.cerevisiae* and *H.pylori* datasets is 10 amino acids, average protein sequence length for *S.cerevisiae* dataset is around 1100 amino acids and for *H.pylori* dataset, average sequence length is around 734 amino acids. It is evident that protein sequences have high length variability. Considering machine learning approaches require fixed-length protein sequences, existing PPI prediction approaches transform variable length protein sequences into fixed-length sequences using traditional copy padding or sequence truncation approaches [450]. We perform experimentation with 6 different variations of copy padding and sequence truncation approaches in order to quantify their efficacy for PPI prediction. Furthermore, we present a unique way to generate fixed-length protein sequences by finding and retaining only the most informative amino acids distributions. This section briefly summarizes five different settings to generate fixed-length sequences.

In 1st setting, performance of 6 traditional copy padding and sequence truncation approaches is evaluated. In copy padding approach, first maximum length of sequence is computed by comparing corpus sequences. Then, all the sequences having length less than maximum length are extended to make them equal to maximum length by adding certain constant. Sequence truncation is another way to make fixed-length sequences where minimum sequence length is computed by comparing corpus sequences. Amino acids from all those sequences whose length is

larger than minimum length are truncated to make them equal to minimum sequence length. Another trend is to utilize both copy padding and truncation approaches where average length of corpus sequences is computed. Certain constant is added in sequences which are shorter than the average length, whereas, amino acids from the sequences that are larger than the average length are truncated.

In copy padding trick, it is an important question whether the start of the sequences is an ideal location for the addition of constant or the end of the sequences. Likewise, in the sequence truncation approach, it is questionable whether extra amino acids need to be truncated from start of the sequences or end of the sequences. For copy padding trick, we first add constant at the start of sequences and in another variation, we add constant at the end of the sequences to find out which strategy is more appropriate. Similarly, for the sequence truncation approach, we truncate sequences from the start of sequences and from the end of sequences in other variation. In hybrid sequence fixed-length generation paradigm based on average length, we also extend or truncate corpus sequences from the start of sequences or end of the sequences. A graphical representation of all 6 strategies is presented in Figure 8.2 under the hood of setting-1.

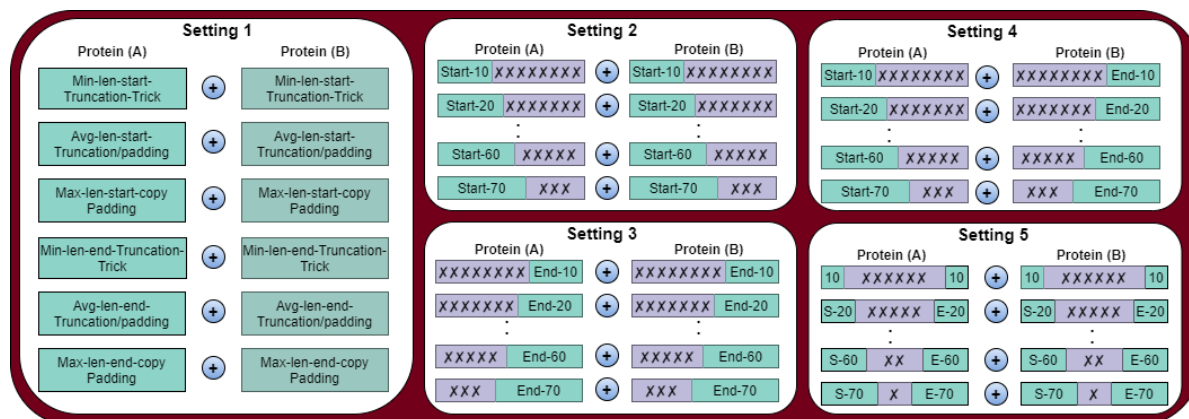


Figure 8.2: A variety of experimental settings to generate fixed-length sequences based on traditional copy padding or sequence truncation and proposed bag of most informative amino acids distribution tricks.

Considering, the vulnerability of traditional copy padding approach to create unnecessary bias through the addition of too many constants and sequence truncation to lose important amino acid distribution while handling flexible protein sequences. Here we propose a unique idea to optimize fixed-length sequence generation process where fixed-length sequences are generated using only the few amino acids from different regions of protein sequences which contains the most informative distribution of amino acids for the task of PPI prediction. More specifically, in Figure 8.2, under the hood of 2nd setting, ADH-PPI selects X amino acids solely from the starting region of one protein A and Y amino acids solely from starting region of protein B. In 3rd setting, performance of X amino acids taken from the ending region of protein A and Y amino acids taken merely from the ending region of protein B is evaluated. Whereas, in 4th setting, X amino acids of

protein A taken from the starting region of protein sequence are combined with Y amino acid of protein B taken from the ending region of protein sequence to assess the discriminative aptitude of start-end region. In last setting, performance is assessed by combining X amino acids taken from start-end regions of protein A with Y amino acid taken from start-end region of protein B. To identify up to what number of amino acids can capture the discriminative essence of protein sequences, in all 4 proposed subsequence based fixed-length generation settings, we select as minimum number of amino acids as possible (e.g., 10, depending on the minimum sequence length of the benchmark dataset) and iteratively increment this number with a step size of 10 amino acids up to 50% of average sequence length of benchmark core PPI prediction datasets. In all 4 settings, number of amino acids range varies from 10-to-70 with an increment of 10 amino acids. By fusing protein A sequences with protein B sequences, the fixed-length protein sequence generated through traditional and novel preprocessing strategies are passed to an attention based deep hybrid neural network for PPI prediction.

8.2.1.3 An Attention based Deep Hybrid Neural Network (ADH-PPI)

In the marathon of developing robust and precise deep learning based end-to-end frameworks for diverse Genomics and Proteomics sequence analysis tasks, we are witnessing the explosion of deep learning approaches, core architectures of which are mainly formed by deep feed forward neural networks [256], deep belief networks [481], convolutional neural networks [256], autoencoders [299] and long short-term memory networks [299]. Predominantly, efforts are being made under the hood of two different paradigms to develop more efficient deep learning models for diverse sequence analysis tasks [256, 299, 481]. The main focus of one paradigm is to develop deep neural networks based on series of neural layers (i.e., convolutional layer, recurrent layer) to effectively capture the non-linearity of genomic and proteomic sequences [256, 284, 299, 317, 481]. Whereas, other paradigm pays more attention to develop shallow or ensemble neural networks which utilize neural layers (i.e., convolutional layer, recurrent layer) in different parallel channels and combine the features extracted by different channels to perform target prediction. The chapter in hand develops an attention based deep hybrid model (ADH-PPI) for PPI prediction following the structure of first paradigm. Workflow of proposed ADH-PPI approach is illustrated in Figure 8.3, a brief description of different components of ADH-PPI approach is given in the following subsections.

Stochastic Embedding Layer:

Stochastic embedding layer takes k-mers of protein sequences and k-mer embeddings learned in unsupervised manner (generation of which is explained in section 8.2.1.1) to generate an embedding weight matrix $E \in R^{|\text{unique_kmers}| \times \text{embedding_size}}$. To fine-tune embedding matrix in a more generic way, we apply two different kinds of dropouts on the embedding matrix, where there is a probability $p_{\text{embeddings}}$ to fully replace k-mer embedding vectors with zeros and probability

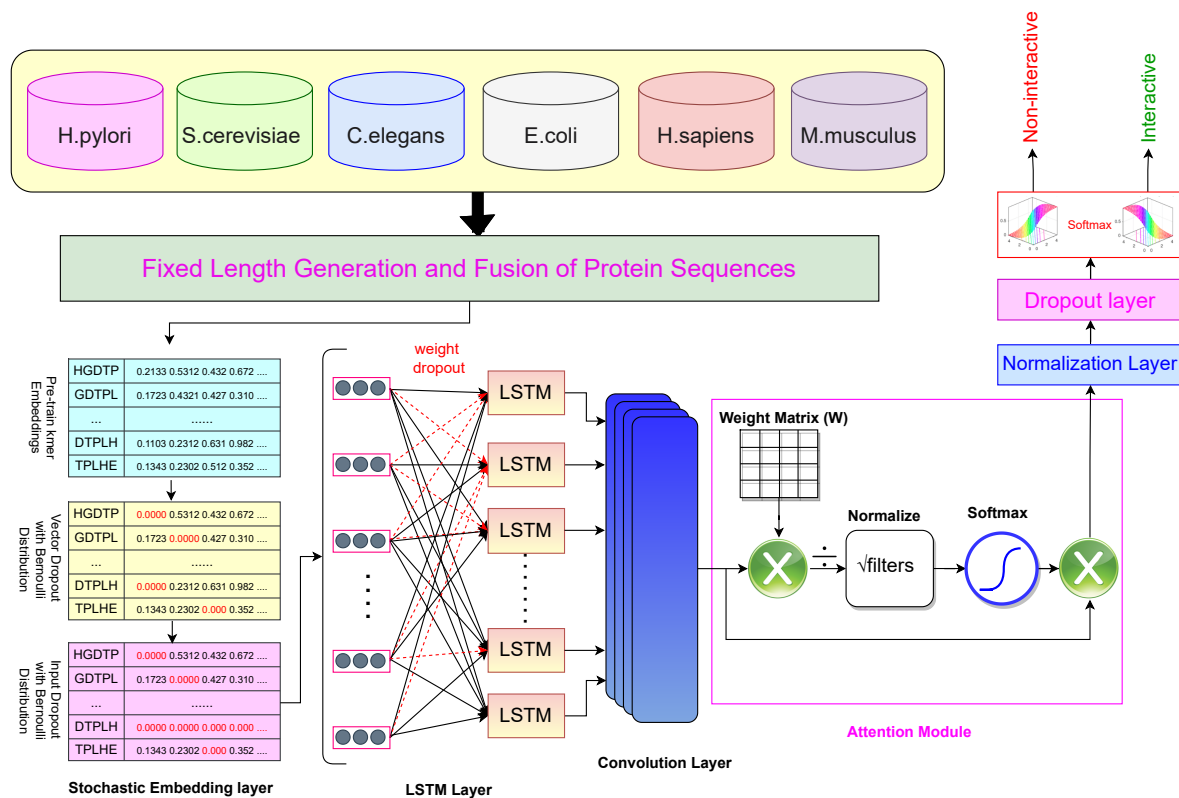


Figure 8.3: Workflow of proposed attention based deep hybrid methodology ADH-PPI for protein-protein interaction prediction

$p_{embeddings_dim}$ to replace individual continuous values with zero in remaining k-mer embedding vectors. First kind of dropout drops few k-mer embedding vectors whereas second kind of dropout drops few continuous values of remaining k-mer embedding vectors [143, 297]. This regularization avoids model over-fitting by ensuring that model does not over-specialize certain k-mers to extract the most informative features for various classes. While performing experimentation, the k-mer embedding vector $p_{embeddings}$ and dimension $p_{embeddings_dim}$ dropout probabilities vary from 0.002-to-0.008 where we find that $p_{embeddings}$ of 0.004 and $p_{embeddings_dim}$ of 0.005 performs better. The optimized embedding matrix containing 120-dimensional embedding vectors for unique k-mers is passed to a Long Short Term Memory layer.

Optimized Long Short Term Memory Layer:

Long short-term memory (LSTM) layer is a special kind of recurrent layer that avoids gradient explosion and gradient disappearance issues faced by the neural network during the modeling of long sequences [197]. Furthermore, LSTM is effective for the extraction of long dependencies of features which is very critical for accurate PPI prediction [197]. Unlike a traditional recurrent neural network, LSTM makes use of multiple gates for the extraction of informative features. A brief description of information flow and extraction in LSTM cell is briefly describe in chapter 7. The 120-dimensional feature vectors produced by the LSTM layer are passed to the convolutional

layer.

Convolutional Layer:

Like simulating the cells with local receptive fields within human brain, the convolutional layer performs an operation known as convolution which uses local connection and shared weights to extract hidden informative features [151]. Convolution operation applied at a particular l^{th} layer produces a feature map $A^{[l]}$ that can be mathematically expressed as:

$$A^{[l]} = f(A^{[l]} \otimes W^{[l]} + b^{[l]}) \quad (8.3)$$

Where $W^{[l]}$ represents the weight matrix of convolutional kernel of the l^{th} layer, symbol \otimes denotes the convolutional operation, $b^{[l]}$ represents the offset vector and $f(x)$ denotes the activation function. We use ReLu as an activation function to sparse the final output of convolutional layer which leads to speed up the training process and maintain the steady convergence rate to prevent vanishing gradient issue. CNN layer uses 50 kernels of size 3 to produce 50-dimensional feature vectors which are passed to an attention layer.

Attention Layer:

Attention layer is widely used to adjust the weights of feature vectors in such a manner that most crucial features are emphasized and less important features are penalized [151]. Attention function can be considered a mapping from a Query vector (Q) and Key-Value vectors (K-V) to an output vector. Here Q, K and V are linear projection of given protein sequence statistical representation and output is the new protein sequence statistical representation of same dimensions incorporating comprehensive mutual association of higher order amino acids present in protein sequences. The entire process involves three steps: acquiring Query, Key and Value linear projections, estimating the weight through placing Query and Key into a certain compatibility function and obtaining the output by estimating the weighted sum of value using the pre-computed weight. There are many types of compatibility functions which produces many flavors of attention mechanism. We use the least space and time efficient version of the compatibility function namely Scaled Dot-Product Attention (SDPA). The SDPA computes the dot product of Query and Key which is divided by $\sqrt{d_k}$ where d_k denotes the Key dimension and finally applies the softmax over it to obtain the weight.

$$Weight = softmax \frac{QK^T}{\sqrt{d_k}} \quad (8.4)$$

In equation 8.4, weight represents a square matrix having number of rows/columns equivalent to length of protein sequences calculated in terms of number of higher order amino acids. Each i^{th} row j^{th} column value denotes the interaction intensiveness among i^{th} higher order amino acid and j^{th} higher order amino acid. After computing weight, every row of output that represents the statistical vector of a higher order amino acid, can be estimated as the weighted sum of all higher order amino acids. This is primarily implemented through a single-matrix multiplication which can be mathematically expressed as follows:

$$Output = Weight * V = softmax \frac{QK^T}{\sqrt{d_k}} V \quad (8.5)$$

Given, 50-dimensional statistical vectors of protein sequences, attention layer updates the values of statistical features on the basis of their usefulness for PPI prediction.

Normalization Layer:

Neural network faces the issue of internal co-variance shift which de-stabilizes the neural network due to change in input distribution to hidden layers of neural network when model weights are updated after the execution of every batch [202]. Internal co-variance shift makes the optimal weights learned by the network during previous iterations obsolete [202], disturbs the convergence and generalizability of the model [202].

Normalization addresses this issue by standardizing the input before feeding it to a hidden layer for every batch. It ensures that input to output mapping of a neural network does not over-specialize one particular region of protein sequences, resulting in faster training, convergence and improved generalizability [202]. A comprehensive detail of normalization layer in terms of mathematical expressions is given in chapter 7.

Standard Dropout Layer:

Dropout is a de-facto standard to regularize neural networks, which generally improves the quality of the hidden features by alleviating the likelihood of hidden units co-adaptation problems. More specifically, for every hidden unit, dropout avoids co-adaptation by iteratively tweaking the presence and absence of other hidden units to ensure that a hidden unit cannot rely on other hidden units to fix its mistakes.

In proposed ADH-PPI methodology, each hidden unit has the probability p to be dropped where the value of p falls in range of 0.01-to-0.4. Mathematically (Equation 8.6), likelihood of omitting a hidden unit is done according to the Bernoulli distribution with probability p . Through an element wise product of hidden unit vector with a mask where each element is randomly sampled from Bernoulli distribution, hidden units are dropped during training. Whereas, for testing (Equation 8.7), instead of dropping the hidden unit, probability for a hidden unit not to be dropped $1 - p\%$ is estimated.

$$y = f(Wx) \cdot m, m_i \sim Bernoulli(p) \quad (8.6)$$

$$y = (1 - p)f(Wx) \quad (8.7)$$

Softmax Layer:

Using a dense 50-dimensional representation of protein sequences, softmax layer discrimi-

nates interactive protein pairs from non-interactive protein pairs. Categorical cross-entropy also known as softmax loss is used as a loss function which is a simple softmax activation plus a cross entropy loss. Working of softmax activation and categorical cross entropy is described in equation 8.8 and equation 8.9, respectively.

$$f(s_i) = \frac{e_i^s}{\sum_j^C e_j^s} \quad (8.8)$$

$$CE = - \sum_i^C t_i \log(f(s_i)) \quad (8.9)$$

In these equations, t represents one-hot encoded ground truth, s_i represents probability score for each class in C and $f(s_i)$ refers to softmax activation applied before the computation of cross-entropy loss.

8.2.2 Benchmark Datasets

In order to prove the integrity of proposed ADH-PPI approach and to perform a fair comparison with existing PPIs prediction approaches, we evaluate ADH-PPI performance over PPIs datasets of 6 different species including humans, Drosophila, Yeast, Bacterium, Caenorhabditis elegans and Escherichia coli.

From Yeast specie, performance of ADH-PPI is evaluated on a well-known public benchmark dataset namely Saccharomyces cerevisiae (S.cerevisiae), which is extensively utilized by several researchers for PPI prediction [450]. PPIs of Saccharomyces cerevisiae (S.cerevisiae) were first extracted by Guo et al. [161] from Database of Interacting proteins (DIP) [424]. Authors eliminated those protein pairs where any one of the protein was comprised of less than 50 amino acids and obtained a dataset of 5,943 protein pairs with positive interactions. To eliminate redundancy, researchers utilized a renowned program CD-HIT [141]. From 11,188 PPIs, a total of 5,594 PPIs were retained considering that they had less than 40% pairwise sequence similarity with each other. An equal number of negative PPIs were generated using 3 different approaches. In first approach, non-interacting protein pairs were generated by random pairing of proteins which were not present in the positive dataset. In second approach, negative dataset was generated by combining proteins having similar subcellular localization patterns extracted from Swiss-Prot database [32]. In third approach, negative dataset was generated using data augmentation approach. Another widely used PPI prediction dataset [450] Helicobacter pylori belongs to Bacterium specie, which was compiled by Martin et al. [290]. It contained 2,916 protein pairs out of which 1,458 protein pairs were positive and 1,458 protein pairs were negative. From a collection of protein pairs which were not explicitly declared as interactive, a bunch of protein pairs were selected as non-interacting proteins. Statistics of both core datasets S.cerevisiae and H.pylori are described in Figure 8.4.

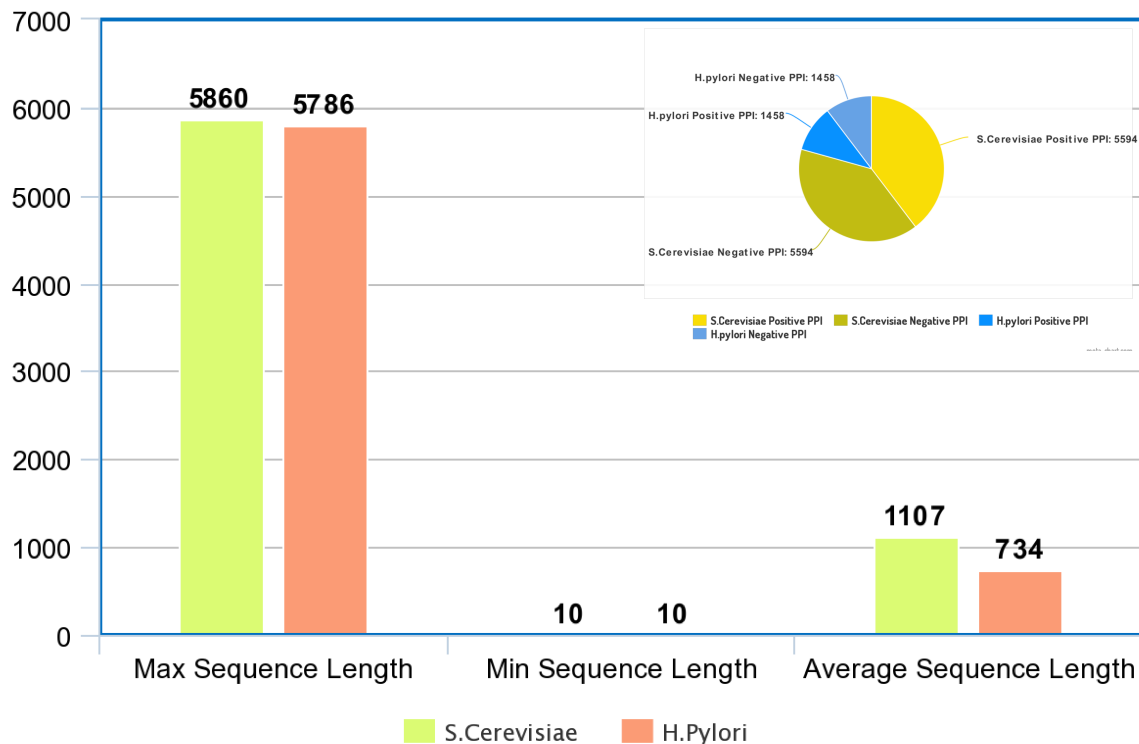


Figure 8.4: Statistics of 2 core protein-protein interaction prediction datasets

In order to perform a fair performance comparison with existing PPI predictors and to further prove the versatility of proposed methodology ADH-PPI, we also evaluate ADH-PPI over 4 independent test sets developed by Zhou et al. [474]. These datasets have been extensively used in literature [170, 196, 450]. As the procedure used to develop 4 different independent test sets has been described in existing studies [170, 196, 450, 474], here we only shed light on the statistics of 4 independent test sets. E.coli consists of 6,954 protein pairs with positive interactions, C.elegans contains 4,103, Homo sapiens (H.sapiens) consists of 1,412 and Mus musculus (M.musculus) is composed of 313 protein pairs with positive interactions.

8.3 Evaluation Criteria

Following the evaluation criteria of previous PPI prediction studies [450], performance of ADH-PPI for 2 core benchmark datasets is evaluated using 10-fold cross-validation. We utilize 6 most commonly used evaluation measures namely: recall, precision accuracy (ACC), Matthews correlation coefficient (MCC), F1-score and area under receiver operating characteristics (AUROC) to compute the model performance from different perspectives.

Exploratory analysis of 2 core PPI datasets (Figure 8.4) indicates that minimum sequence length for both datasets is 10 amino acids, average protein sequence length for S.cerevisiae

dataset is around 1100 amino acids as compared to H.pylori dataset whose average sequence length is around 734 amino acids. To identify up to what number of amino acids can capture discriminative essence of protein sequences, ADH-PPI performs experimentation by selecting as minimum number of amino acids as possible (e.g., 10, depending on minimum sequence length of benchmark dataset) and increments this number with the defined step size up to 50% of average sequence length of benchmark dataset. For 2 core PPI datasets, using the step size of 10, ADH-PPI finds that within 70 amino acids based subsequences (which is almost just 15% of average sequence length) manage to capture the discriminative essence of long protein sequences for the task of PPI prediction. Therefore, in our experimentation, we report the performance by varying amino acids from 10-to-70 using the step size of 10.

Considering the efficacy of grid search for automated parameter search [259, 361], we employ grid search to find optimal values of hyperparameters. We find that FastText [47] captures rich inherent relationships with 5-mers and represents protein sequences using 120-dimensional vectors generated through the concatenation of character k-mers and amino acid level vectors. For embedding generation, experimentation is performed by varying the window size from 3-to-10 and stride size from 1-to-5. To regularize embedding matrix, higher order amino acid vector dropout varies from 0.002-to-0.08 and vector dimension dropout varies from 0.002-to-0.008. Learning rate initial range is defined as 0.05-to-0.01 which is tweaked using the weight decay range of 0.00001-0.01 if validation loss stops improving. Standard dropout probability is varied from 0.001-to-0.6 to avoid over-fitting the proposed ADH-PPI approach for the task of PPI prediction.

Table 8.1: Optimal values of different hyperparameters of proposed ADH-PPI methodology for 2 core datasets and 4 independent test sets for the task of PPI prediction.

PPI Dataset	Degree of Higher Order Amino acid (K-mer)	Stride Size	Sequence Embedding Dimension	Learning Rate	Weight Decay	Dropout Rate	Subsequence Regions
S.Cerevisiae	5	5	FastText-120	0.03	0.1	0.3	P-A_S-40, P-B_E-40
H.Pylori	5	1	FastText-120	0.05	0.01	0.01	P-A_S-40, P-B_E-40
C.elegans	5	5	FastText-120	0.05	1.00E-05	0.1	P-A_S-40, P-B_E-40
H.sapiens	5	5	FastText-120	0.05	1.00E-05	0.1	P-A_S-40, P-B_E-40
M.musculus	5	5	FastText-120	0.05	1.00E-05	0.1	P-A_S-40, P-B_E-40
E.coli	5	5	FastText-120	0.05	1.00E-05	0.1	P-A_S-40, P-B_E-40

ADH-PPI is trained using the batch size of 64, adaptive moment estimation based on weight decay (ADAMW) as an optimizer and categorical cross entropy as a loss function. ADH-PPI is trained on complete core S.cerevisiae dataset to evaluate its performance on 4 independent test sets. Extensive empirical evaluation using defined ranges of diverse hyperparameters is performed to optimize these hyperparameters, the best values of most crucial hyperparameters with respect to 2 core benchmark datasets and 4 independent test sets are summarized in Table 8.1. For both core benchmark datasets, ADH-PPI is trained for 10 epochs where the best model having least validation loss is saved to perform evaluation.

8.4 Results and Discussions

This section comprehensively describes the performance produced by 6 traditional preprocessing strategies used to generate fixed-length sequences. It compares the performance of 4 distinct settings based on subsequences to showcase which region of protein sequences contain the most crucial information about PPI prediction. It also makes a comprehensive comparison between traditional preprocessing strategies and proposed subsequence generation settings. Further, it assesses the performance of optimal most informative subsequence generation setting using protein sequence pairs generated through 2 different protein subsequence orders to validate the robustness of ADH-PPI approach. Finally, it performs a fair comparison of proposed ADH-PPI approach with existing PPI predictors using 2 core datasets and 4 independent test sets belonging to 6 different species.

8.4.1 A Comprehensive Performance Analysis of Traditional Sequence Preprocessing Strategies

Figure 8.5 illustrates the performance values produced by proposed ADH-PPI predictor under the hood of 6 traditional copy padding and sequence truncation approaches used to generate fixed-length sequences across two benchmark core datasets.

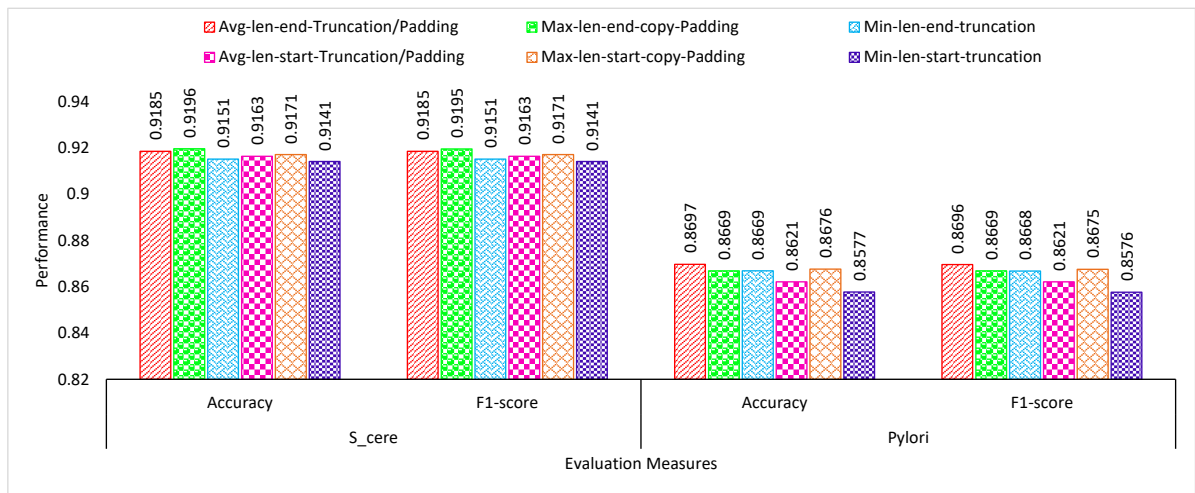


Figure 8.5: Performance comparison of proposed ADH-PPI approach across 2 different datasets including S.Cerevisiae and H.Pylori using 6 traditional sequence fixed-length generation approaches

Performance analysis of 6 commonly used preprocessing strategies over *Saccharomyces cerevisiae* (S.cerevisiae) dataset indicates that, mapping protein sequence to maximum possible length and applying copy padding at the end of protein sequences marks the best performance of 92% in terms of accuracy and F1-score. Mapping protein sequences to average length and

applying padding or sequence truncation trick at the end of protein sequences achieve second best performance. Among all 3 settings which apply copy padding or sequence truncation at the end of protein sequences, mapping protein sequences to minimum possible length attains lowest performance across both evaluation measures. On the other hand, from 3 settings where copy padding or sequence truncation trick is applied at the start of protein sequences, once again mapping protein sequences to maximum possible length achieves overall 3rd best and slightly better performance than other 2 settings based on average and minimum length.

Contrarily, over *Helicobacter pylori* (H.pylori) dataset, mapping protein sequences to average sequence length and applying copy padding or sequence truncation at the ending region of protein sequence marks the best performance followed by the performance produced by maximum sequence length setting where copy padding is applied at the starting region of protein sequences across both evaluation metrics. Setting based on maximum length where copy padding is applied at the end of protein sequence and setting based on minimum length where sequence truncation is applied at the end of sequence length achieve almost similar performance of around 87% in terms of accuracy and F1-score. Whereas, average sequence length based setting where copy padding or sequence truncation is applied at starting region of protein sequences mark slightly better performance than minimum sequence length based setting.

Among all 6 traditional copy padding or sequence truncation approaches, sequence fixed-length generation approaches which apply copy padding or sequence truncation at the ending regions of protein sequences using average or maximum sequence length mark better performance across both core datasets.

8.4.2 Performance Analysis of Proposed Subsequence based Preprocessing Approaches

To showcase the impact of 4 different subsequence based fixed-length generation strategies on the performance of proposed classifier, Figure 8.6 illustrates which protein regions contain most informative distribution of amino acids for PPI predictions across core datasets of 2 distinct species.

A critical analysis indicates that over *S.cerevisiae* dataset, performance of 2 settings where amino acids taken from the start of protein A are combined with the amino acids taken from the end of protein B and amino acids of starting region of protein A are combined with amino acids of starting region of protein B mark similar performance trends across different thresholds of amino acids. While former setting achieves the performance of 95.5%, latter setting attains the performance of 94% until 20 amino acids. With the increase of amino acids, performance of both settings slightly fluctuate before finishing at 95% and 93.5%, respectively at 70 amino acids across both evaluation metrics. Former setting achieves the peak performance using 40 amino acids whereas latter setting mark the best performance with 10 amino acids. Performance of setting-5 which explores the start-end regions of protein A and protein B almost gradually

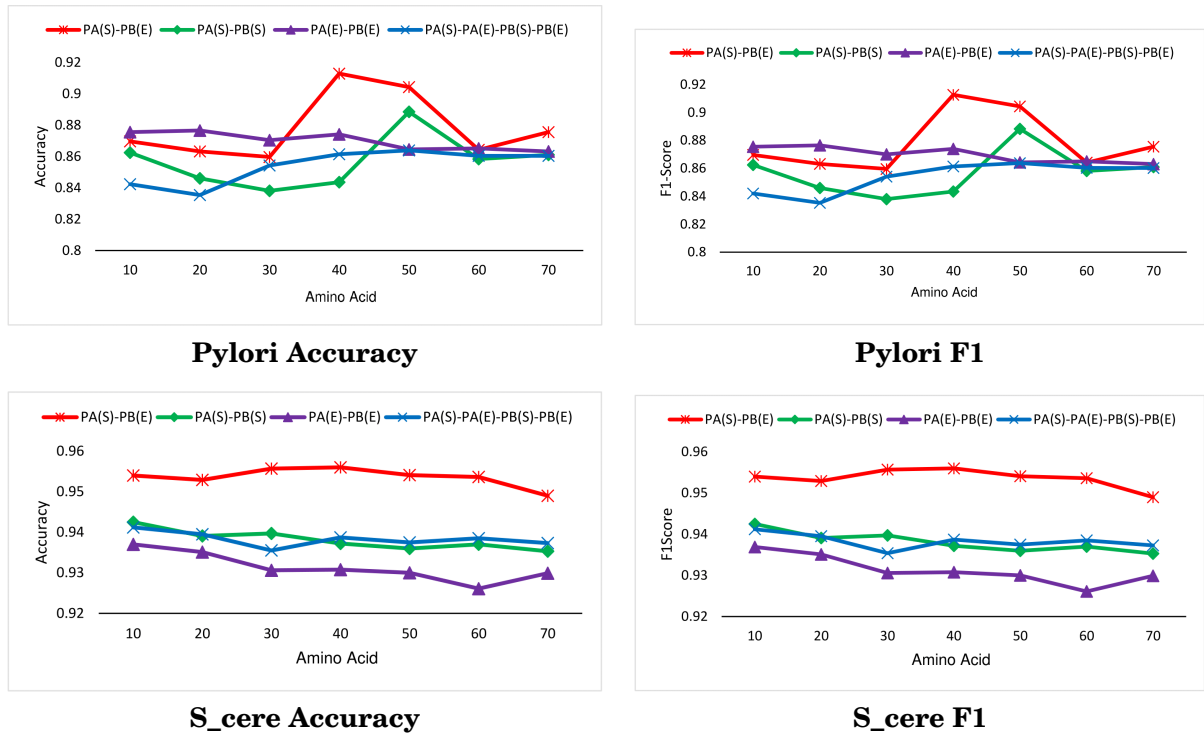


Figure 8.6: Proposed classifier performance analysis under the hood of 4 different subsequences based strategies used to generate fixed-length sequences. Here ‘PA’ represents protein A and ‘PB’ refers to protein B, whereas S indicates the starting amino acids of protein sequence and E represents the ending amino acids of protein sequence.

declines until 30 amino acids, increases up to 94% with 40 amino acids before flattening off across rest of amino acids thresholds. Likewise, performance of setting-3 which selects amino acids from the ending regions of protein A and protein B also progressively decreases from the peak of 93.5% until 30 amino acids before leveling off until 50 amino acids and finishing at 93% at 70 amino acids in terms of accuracy and F1-score. Among all 4 settings, setting-4 which selects amino acids from starting region of protein A and ending region of protein B marks best performance followed by setting-5 which explores the start-end region of both proteins. Whereas, setting-3 marks the lowest performance among all settings based on protein discriminative subsequences.

Over *H.pylori* dataset (Figure 8.6), performance of setting-4 remains around 86% until 30 amino acids before jumping to the peak of 91% with 40 amino acids which declines afterward and finished at 87% with 70 amino acids. Here, performance of setting-2 slightly fluctuates until 40 amino acids before declining and leveling off at 87%. Performance of setting-2 almost gradually decreases from 86% to 84% until 30 amino acids, jumps to the peak of 88.5% until 50 amino acids before decreasing and ending around 86%. Setting-5 performance shows upward trend at most amino acids thresholds and finishes around 86% across both evaluation metrics. Like *S.cerevisiae* dataset, once again, setting-4 which explores the starting region of protein A and ending region

of protein B marks the best performance in terms of accuracy and F1-score. However, for *H.pylori* dataset, peak performances of all 4 settings are comparatively lower than the figures achieved over *S.cerevisiae* dataset across both evaluation metrics.

Overall, among all protein subsequence based settings, setting-4 which selects amino acids from starting region of protein A and ending region of protein B marks best performance across both core PPI datasets in terms of accuracy and F1-score.

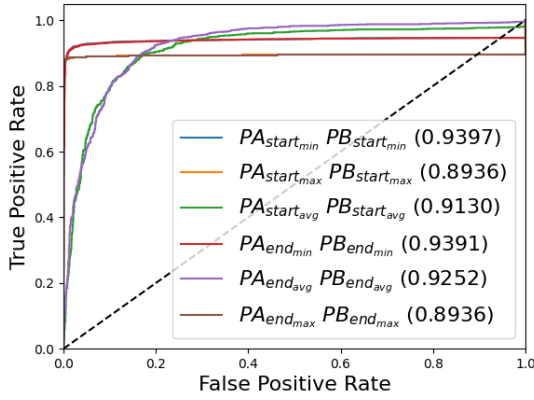
8.4.3 Performance Comparison of Proposed Subsequence Approaches with Traditional Sequence Fixed-Length Generation Approaches

In order to compare the performance of traditional copy padding or sequence truncation based settings with 4 other settings which explore the performance potential of distinct regions of protein sequences by selecting different number of amino acids, Figure 8.7 indicates area under receiver operating characteristics (AUROC) produced by 5 different settings over *S.cerevisiae* and *H.pylori* datasets. As is indicated by the Figure 8.7, over *S.cerevisiae* dataset, in setting-1, applying traditional copy padding or sequence truncation approaches at the ending region of protein sequence slight achieve better degree of separability as compared to those approaches which pad or truncate starting region of protein sequence. Former approaches attain the peak of 95% and latter approaches acquire the peak of 94%. Among all 6 approaches, mapping protein sequences to average length and applying copy padding or sequence truncation at the end of protein sequences mark the best performance followed by another ending region based setting which maps protein sequences to minimum length.

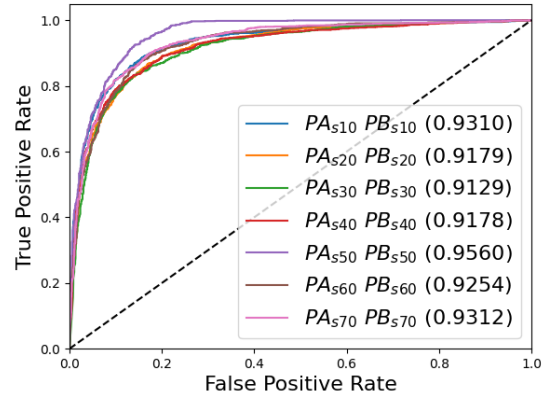
Furthermore, in setting-2 based on partial protein sequences, with the influx of amino acids, ADH-PPI degree of separability gets improved up to the peak of 98% until 30. Afterward, ADH-PPI performance fluctuates across different amino acid thresholds before finishing at 97%. However, all setting-2 amino acid variants achieve better performance than traditional copy padding or sequence truncation approaches (setting-1), indicating the prime performance potential of protein subsequences.

In setting-3 which explores the performance potential of merely ending region of protein pairs, varying the amino acids from 10-to-70, performance of ADH-PPI remains almost constant at 96% which is still better than the performance attained by the most commonly used sequence fixed-length generation approaches (setting-1). Likewise, in setting-4 which selects different amino acids from starting region of protein A and ending region of protein B, ADH-PPI achieves the degree of separability of 98% across 7 different amino acid thresholds, showing best AUROC among all 5 settings. Whereas setting-5 based on start-end region of protein pairs attains the performance of 97% across all 7 amino acid thresholds, indicating degree of separability comparable to setting-2.

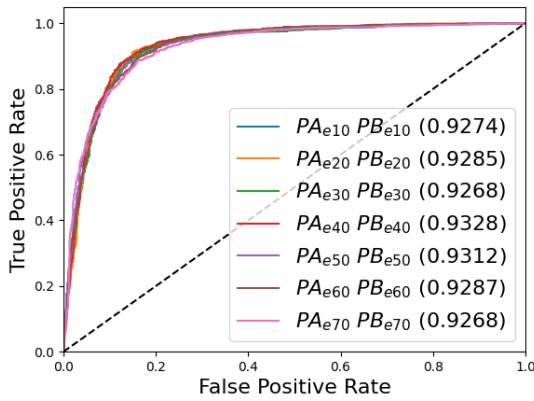
On the other hand, over *H.pylori* dataset, applying copy padding or sequence truncation approaches at the starting region of protein pairs attain slightly superior degree of separabil-



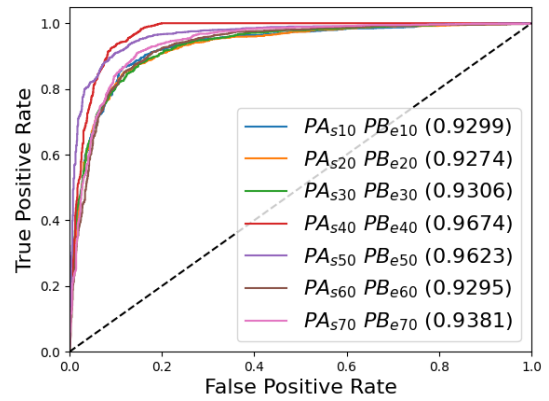
Pylori_Setting-1



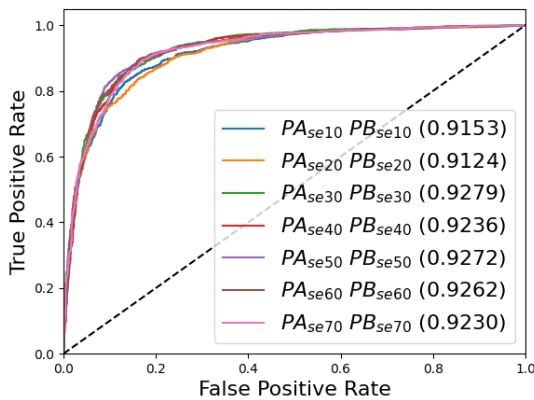
Pylori_Setting-2



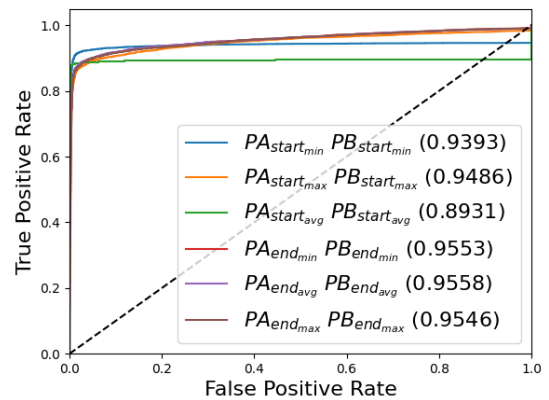
Pylori_Setting-3



Pylori_Setting-4



Pylori_Setting-5



S_cere_setting-1

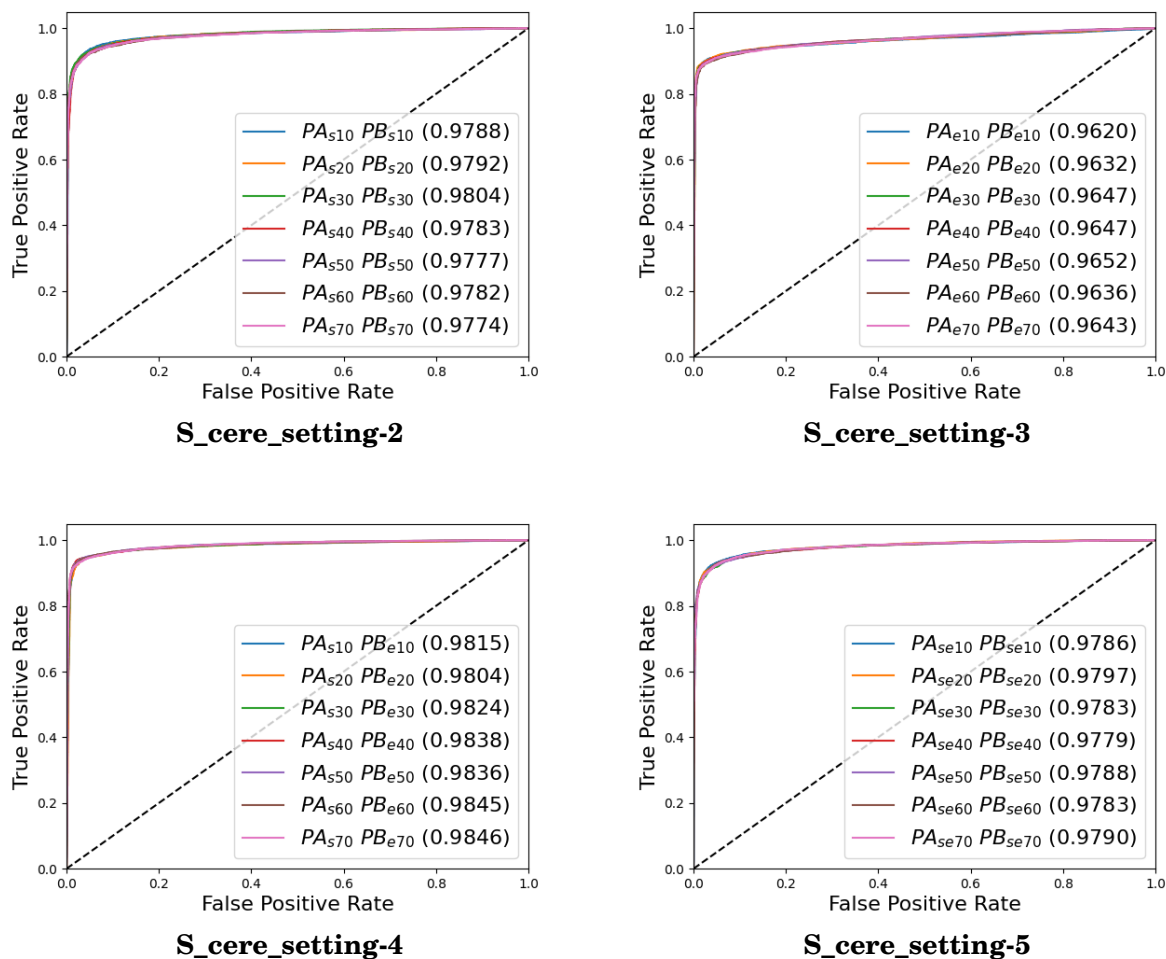


Figure 8.7: Impact of 5 different settings on the performance of proposed ADH-PPI approach across 2 different datasets including S.Cerevisiae and H.Pyloir for the task of PPI prediction in terms of area under receiver operating characteristics. Setting 1 is based on traditional copy padding, sequence truncation and hybrid approaches. Settings 2, 3, 4 and 5 are based on subsequences criteria where X number of amino acids from starting and ending regions of protein A and protein B are taken. The value of X varies from 10 to 70 amino acids with the difference of 10 amino acids. Setting 2 takes X number of amino acids from starting region of protein A and ending region of protein B. Setting-3 takes X number of amino acids from ending region of protein A and protein B. Setting 4 takes X number of amino acids from starting region of protein A and protein B. Setting 5 takes X number of amino acids from starting and ending region of protein A and starting and ending region of protein B.

ity as compared to approaches based on ending region of protein sequences. However, unlike *S.cerevisiae* dataset, here, both kinds of approaches mark better performance by mapping the protein sequences to minimum length. In setting-2, performance of ADH-PPI declines from 93% to 91% until 30 amino acids, however, jumps to 95% until 50 amino acids before finishing at 93%. Overall, it outperforms traditional copy padding or sequence truncation approaches by 2% in terms of AUROC. In setting-3 which merely selects amino acids from ending region of protein pairs, performance of ADH-PPI fluctuates by the figure of 1%. ADH-PPI attains the degree of separability of 93% with 40 amino acids, indicating overall better performance than setting-1 but slightly lower performance than setting-2. Like *S.cerevisiae* dataset, here once again, setting-4 based on starting region of protein A and ending region of protein B achieves the best degree of separability among all 5 settings. With the influx of amino acids, ADH-PPI performance jumps to 96% until 40 amino acids before slightly fluctuating and ending at 93%. Whereas, performance of setting-5 based on start-end region of protein pairs increases up to 92% until 30 amino acids and gets flattened afterward across rest of the amino acid thresholds.

In a nutshell, prime objective of developing Artificial Intelligence based predictors is to make the best use of raw protein sequences, extract distinct distribution of amino acids in the sequences in order to discriminate interactive protein sequences from non-interactive protein sequences. However, protein sequences are highly variable in length and deep learning models require fixed-length input sequences. Commonly used sequence fixed-length generation approaches are copy padding and sequence truncation. In copy padding approach, all sequences are mapped to maximum sequence length by padding certain letter to shorter sequences, whereas in sequence truncation approach, all sequences are mapped to minimum sequence length by eliminating extra amino acids from longer sequences. Distribution of amino acids varies in different subregions of sequences and the performance of deep learning algorithms primarily rely on the extraction of discriminative distribution of amino acids. Copy padding approach creates unnecessary bias through the repetition of same padding letters which make sequences quite similar to each other, similarly, sequence truncation approach is vulnerable to lose the most informative distribution of amino acids. Subsequences based fixed-length generation is more effective as it does not insert any hypothetical letter. Furthermore, it skips constant regions that usually lie in center of the sequences and does not lose informative distribution because it takes both starting and ending regions of the sequences into account. Experimental results reveal that the most discriminative distribution of amino acids lies in first 40 amino acids of protein A and last 40 amino acids of protein B, indicating the success of subsequence based setting for capturing the informative and discriminative essence of protein sequences.

8.4.4 Performance Impact of CNN Layer

To better illustrate the necessity of CNN layer in proposed ADH-PPI predictor, we have performed experimentation on *H.pylori* dataset under the hood of two different settings. In first setting,

we take LSTM, CNN and Attention layers, whereas in second setting, we only take LSTM and Attention layers. Table 8.2 illustrates the predictive performance of both settings in terms of five different evaluation measures namely accuracy, precision, recall, F1-score and MCC.

Table 8.2: Performance analysis of proposed model using pipeline of LSTM, CNN and Attention layers and only LSTM and Attention layers over H.pylori species dataset to quantify the impact of CNN layer

Evaluation Measures	Proposed Model with LSTM, CNN and Attention Layers	Proposed Model with only LSTM and Attention Layers	Performance Difference
Accuracy	0.926	0.919	Around 1%
Precision	0.928	0.921	Around 1%
Recall	0.961	0.945	Around 2%
F1-score	0.944	0.912	Around 3%
MCC	0.855	0.848	Around 1%

Among different subsequence based settings, using 40 amino acids from starting region of protein A and 40 amino acids from ending region of protein B, proposed model with LSTM and Attention layers achieve the accuracy of 0.919, recall of 0.945, precision of 0.921, F1-score of 0.912 and MCC of 0.848. However, this performance is less than the performance achieved using LSTM, CNN and Attention layers in proposed predictor by the F1-score of 3%, accuracy of 2%, precision, recall and MCC of 1%. Overall, exclusion of CNN layer slightly drops the predictive performance and better performance is achieved when LSTM, CNN and Attention layers are used in proposed predictor. This proves the necessity of CNN layer in proposed predictor that essentially captures local dependencies and translational invariance of amino acids present in protein subsequences which complement predictive performance.

8.4.5 Performance Assessment of ADH-PPI Robustness for Different Order Protein Sequence Pairs

Empirical evaluation reveals that proposed ADH-PPI achieves the highest performance on 2 core benchmark datasets and 4 independent test sets on account of protein sequence pairs generated by combining the subsequence of protein A with subsequence of protein B. Among different subsequence generation settings, setting-4 (Figure 8.2) which focuses on the starting region of protein A and ending region of protein B develops the most informative amino acid distribution based protein sequence pairs. However, it is important to note that we have randomly chosen one protein as protein A and other protein as protein B. Building on the equal possibility of generating conversely ordered protein sequence pairs, here we validate the idea that regardless of protein sequence order, starting region of one protein and ending region of other protein contains the most informative amino acid distribution for PPI prediction.

Mainly, experimentation is performed with optimal subsequence generation setting across 2 core datasets and 4 independent test sets by treating one protein subsequence as protein A, other

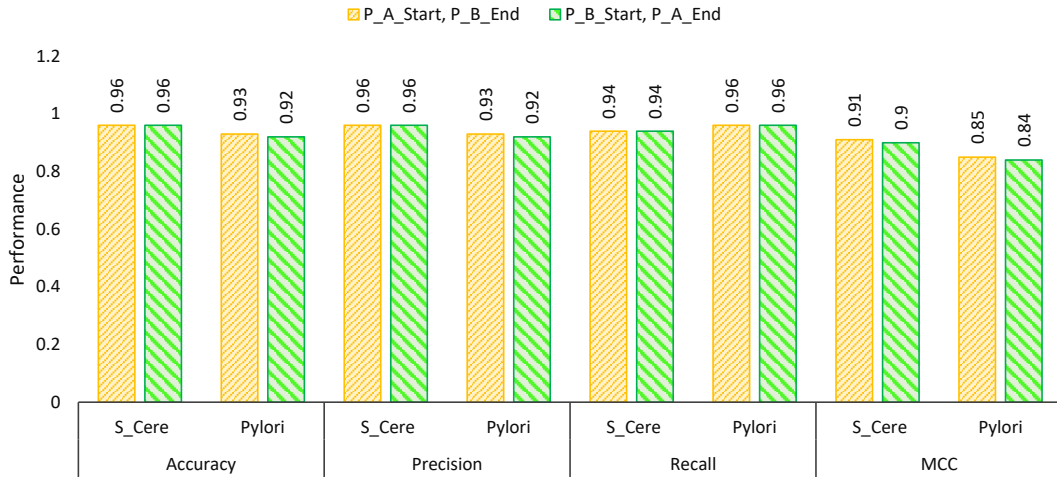


Figure 8.8: Performance assessment of most optimal informative subsequence generation setting using 2 differently ordered protein sequence pairs over S.Cerevisiae and H.Pylori core datasets. Here P_A represents the protein A and P_B refers to protein B, whereas start and end represent the starting and ending region of respective protein

protein subsequence as protein B and exchanging the order of protein subsequences. Furthermore, we use same parameters (e.g., subsequence window size, model parameters (Table 8.2) values described in previous sections 8.4.3 for each dataset across both kinds of paradigms in order to accurately reveal the robustness of ADH-PPI approach.

Figures 8.8 and 8.9 illustrate the performance produced by setting-4 using protein sequence pairs generated by treating one protein as protein A and other protein as protein B as well as reversing the order on 2 core benchmark datasets and 4 independent test sets, respectively. As indicated by the Figures (8.8, 8.9), ADH-PPI achieves almost same performance across all datasets with protein sequence pairs generated using 2 different protein subsequence orders. This indicates that although changing the combination order of protein-subsequences change the characteristic of protein-sequence pairs, however, proposed ADH-PPI is robust enough to capture most informative distribution of protein sequences important for PPI prediction.

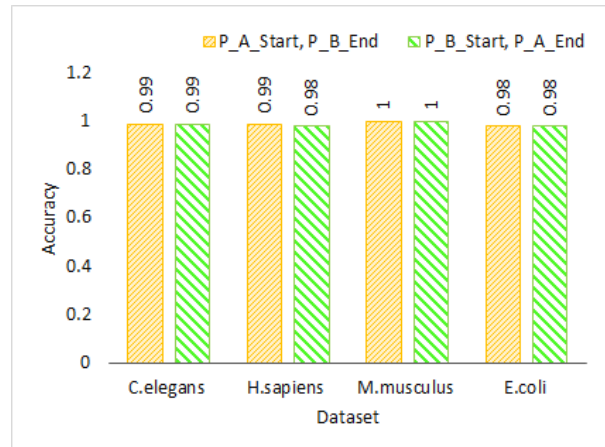


Figure 8.9: Performance assessment of most optimal informative subsequence generation setting using 2 differently ordered protein sequence pairs over C.Elegans, H.sapiens, M.musculus and E.coli independent test sets after training the model on core S.cerevisiae dataset.

8.4.6 Performance Comparison of Proposed ADH-PPI Predictor with Existing PPI Predictors using two Benchmark Core Datasets

In order to prove the integrity of proposed ADH-PPI predictor, rich performance comparison with existing PPI predictors is performance using two core datasets in terms of 4 different evaluation metrics.

Table 8.3: Performance comparison of proposed ADH-PPI predictor with 12 existing PPI predictors on benchmark *S. cerevisiae* dataset, where results of existing PPI predictors are taken from Yu et al. [450] paper.

Method	ACC (%)	Recall (%)	Precision (%)	MCC
ACC+SVM [161]	0.8933 ± 2.67	0.8993 ± 3.68	0.8887 ± 6.16	N/A
Code4+KNN [161]	0.8615±1.17	0.8103±1.74	0.9024±1.34	N/A
MCD+SVM [449]	0.9136 ±0.36	0.9067 ±0.69	0.9194 ±0.62	0.8421±0.0059
MLD+RF [445]	0.9472±0.43	0.9434±0.49	0.9891±0.33	0.8599±0.0089
PR-LPQ+RF [446]	0.9392±0.36	0.9110±0.31	0.9645±0.45	0.8856±0.0063
MIMI+NMBAC+RF [110]	0.9501±0.46	0.9267±0.50	0.9716±0.55	0.9010±0.0092
LRA+RF [448]	0.9414 ± 1.8	0.9122 ± 1.6	0.9710 ± 2.1	0.8896 ± 0.026
DeepPPI [116]	0.9443±0.30	0.9206±0.36	0.9665±0.59	0.8897±0.0062
ippi-esml [210]	0.9515±0.25	0.9221±0.36	0.9797±0.60	0.9045±0.0053
WSRC [223]	0.8673	0.8993	NA	0.7693
DeepFE-PPI [439]	0.9478	0.9299	0.9645	0.8962
GcForest-PPI [450]	0.9544	0.9272	0.9805	0.9102
Proposed ADH-PPI	0.9573	0.9394	0.9575	0.9144

Table 8.3 compares the performance of proposed ADH-PPI predictor with 12 machine and deep learning based predictors over *S. cerevisiae* dataset. As indicated by the Table 8.3, proposed ADH-PPI predictor outperforms auto co-variance and SVM based PPI prediction methodology [161] by 7%, 5% and 7% and KNN based methodology [161] by 10%, 13% and 6% in terms of accuracy, recall and precision, respectively. It outperforms WSRC classifier [223] by 9%, 4% and 14% in terms of accuracy, recall and MCC and ippi-esml [210] approach by 3%, 5%, 3% and 4% in terms of accuracy, recall, precision and MCC, respectively. Multi-scale continuous and discontinuous (MCD) feature representation and SVM classifier based approach [449] takes the previous best accuracy of 89% to 91%, amino acid substitution matrix based feature representation and RF classifier based approach [448] attains the accuracy of 94%. RF classifier achieves the accuracy of 95% using multivariate mutual information (MMI) of protein feature representation [110] and 94% using multi-scale local descriptor (MLD) based feature representation [445]. Proposed ADH-PPI predictor outperforms SVM and random forest based PPI prediction methodologies by the comparable margin. From existing machine learning based PPI predictors, GcForest-PPI [450] achieves top performance in terms of most evaluation metrics. Proposed ADH-PPI predictor surpasses the performance of GcForest-PPI [450] by 1% in terms of accuracy and recall and equalizes the MCC performance value.

Turning towards existing deep learning based PPI prediction methodologies, DeepPPI [116] and DeepFE-PPI [439] achieve almost similar performance on *S.cerevisiae* dataset in terms of four different evaluation metrics. Proposed ADH-PPI predictor outperforms deep learning based PPI predictors by 1% in terms of accuracy, recall and MCC.

Table 8.4: Performance comparison of proposed ADH-PPI predictor with 10 existing predictors on benchmark *H. pylori* dataset, where results of existing PPI predictors are taken from Yu et al. [450] paper.

Method	ACC (%)	Recall (%)	Precision (%)	MCC
SVM [6]	0.8340	0.7990	0.8570	N/A
WSR [308]	0.8370	0.7900	0.8700	N/A
Ensemble of HKNN [309]	0.8660	0.8670	0.8500	N/A
DCT+WSRC [195]	0.8674	0.8643	0.8701	0.7699
MCD+SVM [449]	0.8491	0.8324	0.8612	0.7440
MIMI+ NMBAC+RF [110]	0.8759	0.8681	0.8823	0.7524
DeepPPI [116]	0.8623	0.8944	0.8432	0.7263
ippi-esml [210]	0.9047±0.84	0.9115±1.42	0.8999±2.06	0.8100±0.0163
WSRC [223]	0.7870	0.7321	NA	0.7693
GcForest-PPI [450]	0.8926	0.8971	0.8895	0.7857
Proposed ADH-PPI	0.9263	0.9609	0.9284	0.8547

Moreover, performance produced by proposed ADH-PPI predictor and ten existing PPI predictors on *H.pylori* dataset is shown in Table 8.4. Analysis of Table 8.4 reveals that proposed ADH-PPI predictor achieves even more promising figures than existing PPI predictors across all evaluation metrics. Proposed ADH-PPI predictor outshines best performing machine learning based PPI predictor namely GcForest-PPI [210] by 7%, 7%, 4% and 4% in terms of recall, MCC, precision and accuracy, respectively. It outperforms another top performing MIMI and Random forest based PPI predictor [110] by Matthews correlation coefficient of 13%, recall of 10%, precision of 5% and accuracy of 6%. In comparison to deep learning based PPI predictors, proposed ADH-PPI predictor outperforms DeepPPI [116] predictor by 7%, 7%, 9% and 12% in terms of accuracy, recall, precision and MCC.

To summarize, proposed ADH-PPI predictor outperforms both machine and deep learning based PPI prediction methodologies by decent margin for *S.cerevisiae* and by significant margin for *H.pylori* dataset. It is important to mention that Kong et al. [223] proposed FCTP-WSRC predictor results are not comparable to proposed ADH-PPI predictor. Generally, dimensionality reduction approaches such as principal components analysis (PCA) is applied on training data to learn the reduced matrix and the transformation is applied on testing data where test data is projected to reduce feature space. However, Kong et al. [223] applied PCA on training and testing data separately which introduces biasness. In our experimentation, to find the valid performance figures of FCTP-WSRC predictor [223], we have applied the PCA in correct manner and reported

the valid results on 2 core benchmark datasets (Tables 8.3, 8.4) and independent test sets (Figure 8.10).

8.4.7 Performance Comparison of Proposed ADH-PPI Predictor with Existing PPI Predictors using four Independent Test Sets

To further prove the effectiveness of proposed ADH-PPI predictor, comparison between 6 existing PPI predictors and proposed ADH-PPI predictor is performed. Following experimentation criteria of existing predictors, we train the proposed predictor over core *S.cerevisiae* dataset and perform evaluation over 4 different independent test sets belonging to *C.elegans*, *E.coli*, *H.sapiens* and *M.musculus* species [196, 450]. Figure 8.10 compares the accuracy of proposed ADH-PPI predictor with existing predictors. As shown by the Figure 8.10, proposed ADH-PPI predictor achieves the best performance across all four independent test sets which is higher than the peak performance achieved by deep learning based PPI predictor DeepPPI [116] by 4% for *C.elegans*, 6% for *E.coli*, 5% for *H.sapiens* and 9% for *M.musculus* species test sets.

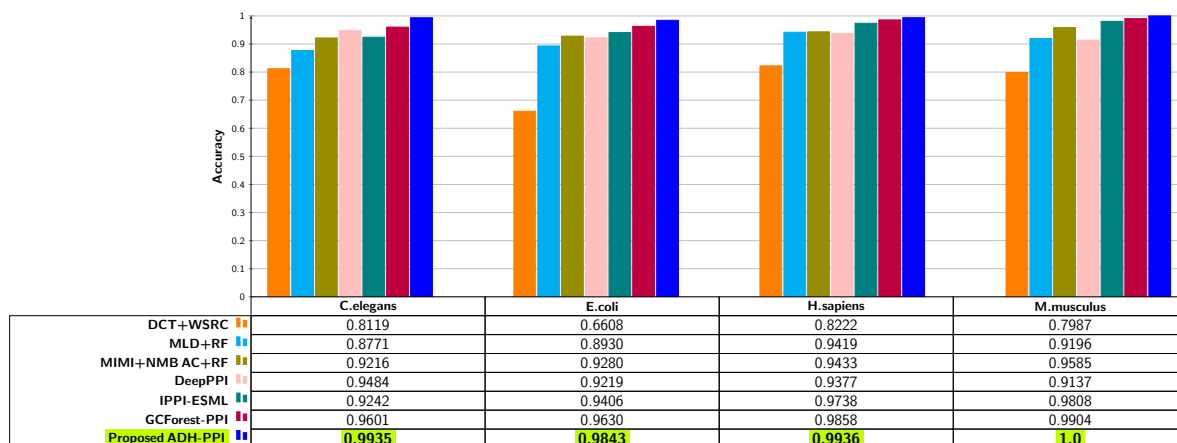


Figure 8.10: Accuracy comparison of ADH-PPI and recent PPI predictors on 4 independent test sets

Furthermore, proposed ADH-PPI predictor outperforms machine learning based state-of-the-art PPI predictor namely GCForest-PPI [450] by 3%, 2%, 1% and 1%, achieving more than 98% performance over all four different species independent test sets.

In a nutshell, over two core datasets and four independent test sets, among all existing PPI predictors, machine learning based PPI predictors perform better than deep learning based predictors. Proposed ADH-PPI predictor outshines state-of-the-art PPI predictor across all 6 datasets of different species including: humans, *Drosophila*, Yeast, Bacterium, *Caenorhabditis elegans* and *Escherichia coli* in terms of most evaluation metrics. The paradigm of considering both k-mer distributions as well as amino acid distributions within k-mer best characterize the protein sequences. Furthermore, the utilization of LSTM, CNN and attention ensures the

extraction of comprehensive discriminative features along with long range dependencies which are essential to accurately predict PPIs across different species. The best utilization of multiple strategies not only enhances the predictive power of proposed ADH-PPI approach but also makes the decisions of proposed ADH-PPI predictor interpretable. Therefore, we believe ADH-PPI will prove a great computational asset for biological researchers and practitioners which can be used to find protein-protein interactions, protein non-coding ribonucleic acid interactions or even interactions between different biomolecules.

8.4.8 A Case Study: Objective Evaluation of Proposed Strategies for Fixed-Length Generation of Sequences

We have seen in previous sections, while generating fixed-length sequences of proteins, novel paradigm of retaining only most informative subsequences helps the proposed ADH-PPI approach to most precisely predict protein–protein interactions. To validate the versatility, generalizability and practical significance of subsequence based fixed-length generation strategies, to perform a case study analysis, similar to protein-protein interaction prediction [25], we consider two tasks namely lncRNA–protein interaction prediction and lncRNA-miRNA interaction prediction. In lncRNA-miRNA interaction prediction, length of lncRNAs varies in thousands of nucleotides, to generate fixed-length sequences, we perform experimentation by utilizing traditional fixed-length generation strategies (copy padding, sequence truncation and hybrid approach) and taking only few nucleotides from starting region, ending region and from both starting and ending regions. Selected subsequences of lncRNAs and miRNAs were passed to LSTM based classifier with random embeddings. Based on experimental results, among traditional and proposed subsequence based sequence fixed-length generation strategies, classifier produces better performance by generating fixed-length of lncRNA sequences by retaining nucleotides from starting region only. Furthermore, LSTM based classifier along with subsequence based strategy manages to outperform existing lncRNA-miRNA interaction predictors.

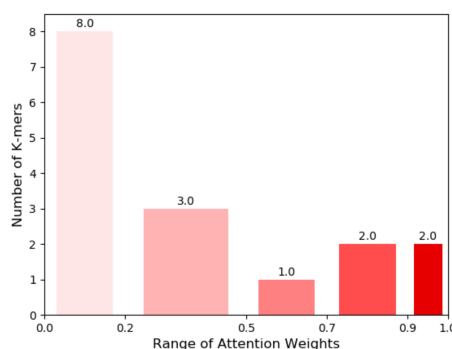
On the other hand for lncRNA-protein interaction prediction, both types of molecules lncRNA and protein have high variability in the length of sequences. To generate fixed-length of lncRNA and protein sequences, we utilize traditional and proposed subsequence based fixed-length generation strategies. Experimental results reveals the superior performance of LSTM based classifier by taking few nucleotides from the starting region of both sequences. Over public benchmark dataset, proposed classifier along with subsequence based strategy outperforms existing lncRNA protein interaction predictors.

8.5 Explainability of Proposed ADH-PPI Predictor

With an aim to overcome a very common black box modeling issue of deep neural networks by decoding the importance of individual amino acids and k-mers, we analyze the attention weights

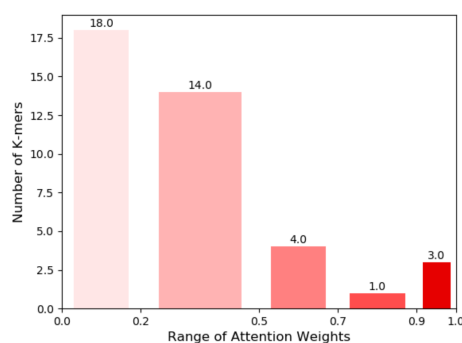
associated with different k-mers to illustrate which k-mers contribute the most in making accurate PPI predictions.

MSGGK GKGAG SAAKA
 SQSRS AKAGL TFPVG RVHRL
 LRRGN KLVVV VLEEP SASIM
 VELKL PLEIN YAFSM DEEFK
 NMDCI



a) Benchmark *S.cerevisiae* Dataset

MRFYF RFYFK FYFKF YFKFL FKFLW KFLWL FLWLL LWLLG
 WLLGI LLGIF LGIFL GIFLI IFLIF FLIFY LIFYR IFYEL FYFLD
 YFLDF FLDFK LDFKG DFKGS FKGSS KGSSS GSSSY SSSYI
 SSYIS SYISD YISDR ISDRI SDRIK DRIKN RIKNA IKNAL
 KNALM NALMN ALMNA LMNAM MNAMV NAMVY AMVYL



b) Benchmark *H.pylori* Dataset

Figure 8.11: Most and least informative amino acid and k-mers patterns identified by Attention layer of proposed ADH-PPI predictor in two test protein sequences belonging to benchmark *S.cerevisiae* and *H.pylori* datasets

To more precisely demonstrate the explainability of the proposed ADH-PPI approach, we arbitrarily take two protein sequence pairs from test sets of benchmark *S.cerevisiae* and *H.Pylori* datasets. Following the working paradigm of proposed ADH-PPI predictor, we generate 5-mers of both test protein sequence pairs and feed both test protein sequence pairs 5-mers along with pre-trained embeddings to two different classifiers trained on *S.cerevisiae* and *H.Pylori* training sets. These classifiers decide whether given protein sequence pairs are interactive or non-interactive. Classifiers make decisions based on the attention weights associated with 5-mers. We extract attention weights and feed these attention weights to decision explainable module which performs reverse engineering to map these attention weights to different 5-mers. To illustrates better,

decision explainable module categorizes different 5-mers into five different groups based on different thresholds applied at attention weights ranging from 0-to-1. Each group is represented with a unique shade of red color, the higher the intensity of red color is the higher the attention weight is for particular k-mer, indicating the darkest red color 5-mers and their inherent amino acids contribute the most in making accurate PPI predictions and the lightest red color 5-mers and their inherent amino acids make least contributions in making accurate PPI predictions. The attention weights range of five different groups of 5-mers is shown on the x-axis of the bar graph (Figure 8.11) whereas y-axis of the bar graph shows the count of 5-mers in each group.

It is evident in the Figure 8.11, for *S.cerevisiae* dataset, only two 5-mers GGKAG and SAAKA fall in first group which has the best range of attention weights 0.90-to-1.0. Two 5-mers fall in second group which has second best range of attention weights 0.70-to-0.89. Similarly, one 5-mer falls in third group and three k-mers fall in fourth group which have attention weight ranges of 0.50-to-0.69 and 0.20-to-0.49, respectively. Among all groups, fifth group has most eight 5-mers, attention weights of which falls in range of 0.1-to-0.20. Furthermore, it can be seen that starting 5-mers distribution has the top attention weights where amino acids G and A are most frequent, which contribute the most in making accurate PPI predictions on *S.cerevisiae* dataset.

Unlike *S.cerevisiae* dataset, in *H.pylori* dataset, eighteen 5-mers fall in fifth group, fourteen 5-mers in second group and four 5-mers in third group. Once again, very few 5-mers fall in first and second group. More specifically, three 5-mers LIFYF, IFYFL, LDFKG, individual amino acids F and L of central regions of subsequence contributes the most in making accurate PPI predictions on *H.pylori* dataset.

These attention weight distribution patterns at the k-mer level and amino acid level are quite consistent across most sequences. Furthermore, this is quite consistent with our unique hypothesis of predicting PPIs using only the most discriminative subsequences. The starting or central k-mers distribution within subsequences gets the higher attention weight and serves as most influential regions and the surrounding k-mers distribution gets lower attention weights and exists as supportive and auxiliary information regions. In a nutshell, we validate the ADH-PPI suitability to discover useful patterns in protein sequences, their dependencies and explainable associations for PPI prediction.

8.6 Conclusion

This dissertation can be considered a huge milestone towards the accurate prediction of PPIs for a variety of species solely using raw sequences. First, unlike previous methods, it captures comprehensive amino acids order, occurrence and contextual information by generating k-mer of protein sequences, distributed representations of which are computed as the sum of their embeddings and the embeddings of their inherent amino acid sub-mers using FastText [47] approach. Second, instead of feeding entire protein sequences to deep learning models, it explores the discriminative

aptitude of multifarious regions of protein sequences to obtain highly informative amino acid distribution based subsequences. Third, it develops an attention based deep hybrid neural network which makes the best use of heterogeneous layers (LSTM, CNN, Attention) to make accurate and interpretable PPI predictions. A stringent benchmarking performance comparison of ADH-PPI with existing computational predictors proves that ADH-PPI outperforms existing machine and deep learning based PPI predictors by decent margin. A compelling future line of current would be to assess the performance potential of ADH-PPI approach for interaction prediction tasks related to other bio-molecules.

PROTEIN VIRUS INTERACTION PREDICTION

Viruses have a long history of posing threat to living organisms [126] as they have caused more than 300 million deaths worldwide [249]. A recent emanation of Severe Acquired Immunodeficiency Syndrome Coronavirus-2 (SARS-CoV-2) is an example of an acute virus that caused a global pandemic [316]. According to World Health Organization, SARS-CoV-2 has caused approximately more than 400 million infections and 6 million deaths across the globe [305]. Likewise, Ebola virus was also responsible for an epidemic that caused more than 11 thousand deaths in Africa [66].

Viruses are small microscopic particles that contain a genetic material (DNA or RNA) surrounded by a protein coat [126]. These particles are considered non-living because of their inability to reproduce or perform any other biological function since they lack specific proteins [399]. However, once they get a chance to enter inside host cell, they make interactions with available proteins in the cell and become capable to reproduce themselves [102]. Initially, to enter inside a host cell, the viruses interact with the host cell receptor proteins [108] and replicate themselves by injecting their genetic material in the cell's genome [280]. After the entrance into the cell, the aim of viruses is to interact with diverse types of proteins through which they can control the process of cell cycle, particle assembly, apoptosis and cell metabolism [102, 385]. The relationships between host and virus proteins are termed as virus-host (VH) protein-protein interactions (PPIs) [432].

To prevent viruses from interacting with host proteins, hosts have sophisticated mechanisms to recognize and confine the viruses, such as the dendritic and β - cells, T-cells and major histo-

⁰This chapter is an adapted version of the work presented in Asim et al., "LGCA-VHPPI: A Local-Global Residue Context Aware Viral-Host Protein-Protein Interaction Predictor", In *PLOS ONE* (2022) [21] and Asim et al., "MP-VHPPI: Meta Predictor For Viral Host Protein-Protein Interaction Prediction in Multiple Hosts and Viruses", Under review in *Frontiers in Medicine*

compatibility complex (MHC) [70]. Therefore, viruses tend to adapt in an efficient manner by interacting with specific host proteins and cellular pathways that prove to be substantial for evading or inactivating factors that are detrimental to viral growth [102]. Meanwhile, to enhance the immunity against viruses, it is difficult to develop efficient vaccines/drugs because of the poor understanding of different mechanisms that have been adapted by the viruses and their frequent transmissibility from cell-to-cell or species-to-species [337]. Consequently, analyses of virus-host PPIs is essential to explore their effects on diverse types of biological functions and to design antiviral strategies [329]. Furthermore, through such analyses essential viral proteins and viral dependencies on host proteins can be identified as drug targets to halt the replication process of viruses by pharmacological inhibition [278].

9.1 Related Work

Multiple experimental techniques have been utilized to identify virus-host protein-protein interactions (VHPPIs) such as protease assay [304], surface plasmon resonance (SPR) [347], Förster resonance energy (FRET) [427], Yeast two hybrid screening (Y2H) [56] and affinity purification mass spectrometry (AP-MS) [148]. Such conventional wet lab methods are expensive, time-consuming and error-prone, which impede inter and intra species large scale virus-host PPIs analyses. To empower the process of virus-host protein-protein interaction analyses, the development of computational approaches by utilizing the power of artificial intelligence is an active area of research [35, 112, 393]. With an aim to provide cheap, fast and accurate virus-host PPIs analyses to date around 13 AI-based predictors [9, 35, 105, 112, 115, 121, 216, 238, 276, 393, 433, 434, 473] have been proposed.

Recently, Yang et al., [434] proposed VHPPIs predictor by utilizing position-specific scoring matrices to statistically represent virus and host protein sequences that were further passed to Siamese convolutional neural network for VHPPI prediction. The predictor was evaluated on VHPPI data of human proteins and 8 different viruses. Another similar predictor namely, Deep Viral [276] used one hot vector encoding (OHE) for the discretization of sequences and convolutional neural network (CNN) architecture for VHPPI prediction. Deep Viral was evaluated on VHPPIs of human and 12 different viruses. Deep-VHPPI [238] predictor also used one hot vector encoding (OHE) and attention mechanism along with CNNs for VHPPIs prediction. The predictor was evaluated on VHPPIs data related to human and 4 different viruses.

Ding et al., [105] proposed a VHPPI predictor based on long short-term memory (LSTM) neural network. At preprocessing stage, they generated statistical representations of viral and host proteins by reaping the benefits of 3 different encoders namely, relative frequency of amino acid triplets (RFAT), frequency difference of amino acid triplets (FDAT) and amino acid composition (AC). The predictor [105] was evaluated on VHPPIs across proteins belonging to 137 different viruses and 13 hosts.

Denovo [121] used amino acids properties like dipoles and volumes of side chains to represent 20 amino acids (AAs) with only 7 cluster numbers to reduce the diversity of amino acids. The sequences were then encoded based on the normalized k-mer frequencies of 7 unique clusters. Denovo predictor used support vector machine (SVM) and was evaluated on the dataset of 10 viruses and human proteins. HOPITOR [36] used the similar encoding method as Denovo [121]. HOPITOR used SVM and was evaluated on 10 different viruses and human proteins. Yang et al., [433] proposed InterSPPI-HVPPI which utilized Doc2vec embeddings and random forest (RF) classifier for VHPPIs prediction. The predictor [433] was evaluated on data related to 12 viruses and human proteins. Karabulut et al., [216] proposed meta predictor (ML-AdVInfect) that reaped the benefits of 4 existing predictors namely HOPITOR [36], InterSPPI-HVPPI [433], VHPPI and Denovo [121]. Specifically authors passed the predictions of existing predictors to SVM classifier for final VHPPI prediction.

Barman et al., [35] proposed VHPPIs predictor that utilized RF classifier and statistical vectors generated through 4 different encoding methods namely, average domain-domain association score, virus methionine, virus seline and virus valine. The predictor was evaluated on VHPPIs data related to human proteins and 5 different viruses. Zhou et al., [473] used 7 sequence encoding methods i.e., RFAT, FDAT, AC, composition, transition and distribution of amino acid groups. The approach [473] used SVM for VHPPIs predictions across the proteins of 332 viruses and 29 hosts. Alguwaizani et al., [9] combined statistical vectors of 4 different encoders namely, amino acid repeats, the sum of squared length of single amino acid repeats (SARs), maximum of the sum of squared length of SARs in a window of 6 residues and composition of amino acids in 5 partitions of the protein sequence. The predictor used SVM classifier and experimentation was performed on VHPPI data related to 6 hosts and 5 viruses. Recently, we proposed LCGA-VHPPI predictor [21], that made use of local-global residue context aware sequence encoding scheme and a deep forest model. Proposed predictor was evaluated on data related to 23 viruses and human proteins.

Following the success of neural word embedding approaches in natural language processing and bioinformatics, Tsukiyama et al., proposed LSTM-PHV [393] that transformed viral host protein sequences to statistical vectors by learning statistical representation of k-mers in an unsupervised manner using Word2vec approach. The study [393] used bidirectional LSTM for VHPPI prediction and data of proteins belonging to 332 viruses and 29 hosts. Similarly, MTT [112] predictor utilized randomly initialized embeddings and LSTM based classifier. MTT predictor was evaluated on data related to 16 viruses and human proteins. Hangyu et al., [115] developed a VHPPI predictor based on Node2vec and Word2vec embeddings methods and a multilayer perceptron (MLP) classifier. Authors performed experimentation over 7 variants of SARS virus and 16 different hosts proteins.

The working paradigm of existing VHPPI predictors can be broadly categorized into two different stages. At first stage, raw sequences are transformed into statistical vectors where the

aim is to capture distributional information of 21 unique amino acids. In second stage, machine or deep learning classifier is utilized to discriminate interactive viral-host protein pairs from non-interactive ones. At first stage, while transforming raw sequences to statistical vectors, 2 predictors [238, 276], has made use of one hot vector encoding method which lacks information related to correlations of amino acids. Moreover, 3 predictors use word embedding generation approaches [112, 276, 393], that capture kmer-kmer associations but lack information related to distribution of amino acids. To capture distribution and various patterns of amino acids, other predictors utilized 10 different mathematical encoders [35, 105, 121, 433, 434] however, these encoders do not capture sequence order or amino acids (AAs) correlation information. Such information is crucial for the analyses of protein sequences as reported in the existing studies [91, 193, 375, 386] which include sequence encoders such as, amphiphilic pseudo-amino acid composition (APAAC) and Quasi sequence order (QS order). Despite the promising performance shown by APAAC and QS order encoders for subcellular location prediction [91], Cyclin protein classification [375] and protein-protein interaction prediction [193, 386] tasks, no researcher has explored their potential to effectively generate numerical representations of viral-host protein sequences.

At second stage, 4 predictors [105, 238, 276, 434] utilize convolutional neural networks (CNNs), 2 predictors [105, 393] make use of LSTM architecture and 8 predictors [9, 35, 112, 115, 121, 216, 433, 473] use traditional classifiers. As such predictors have shown better performances across limited hosts and viruses, therefore these predictors cannot be generalized across multiple hosts and viruses. For instance, LSTM-PHV is the most recent predictor which managed to produce better performance for human and SARS-CoV-2 related VHPPIs, but failed to produce similar performance over Zhou et al., [475] datasets which contain multiple hosts and viruses. To make a generic predictor capable to accurately predict interactions across multiple hosts and viruses, only one meta predictor [216] has been developed. However, this meta predictor relies on the predictions of 4 existing VHPPI predictors that have their own drawbacks at sequence encoding and classification level.

With an aim to develop more accurate and generic meta predictor, the contributions of this chapter are manifold, i) It makes use of two different physicochemical properties based sequence encoding methods namely, APAAC and QS order. In addition, unlike other protein sequence analysis tasks where numerical representations of protein sequences have been generated through these encoders by utilizing combination of different physicochemical properties, it proposes an effective way to generate numerical representations by using a precise subset of physicochemical properties. ii) Considering different physicochemical properties in both encoders extract some irrelevant and redundant features, to remove such features, it transforms original feature space into reduced and more discriminative feature space by utilizing dimensionality reduction method named feature agglomeration. iii) Using separate and combined statistical vectors generated through APAAC and Qsorder, it generates more effective and discriminative

probabilistic feature space by fusing the predictions of two different classifiers. Optimized probabilistic feature space is used to feed SVM classifier which makes final predictions. iv) Large-scale experimentation over 7 public benchmark datasets and performance comparison of proposed meta predictor with existing predictors is performed. v) To facilitate researchers and practitioners, web application based on proposed meta predictor is developed.

9.2 Materials and Methods

This section illustrates details of proposed meta predictor and Viral-host protein interaction prediction benchmark datasets. A comprehensive details of evaluation measures and classifiers used in propose meta predictor are describe in chapter 2.

9.2.1 Meta Predictor

Machine learning classifiers cannot directly operate on raw sequences due to their dependency over statistical representations. While transforming raw protein sequences into statistical vectors, the aim is to encode positional and discriminative information of amino acids. To represent viral and host protein sequences by extracting both types of information, proposed meta predictor makes use of two sequence encoders namely amphiphilic pseudo-amino acid composition (APAAC) and Quasi sequence-order (QS order). The statistical vectors generated by these methods depend on certain physicochemical properties. For example, APAAC [92] encoder contains three different physicochemical properties namely hydrophobicity, hydrophilicity and side chain mass whereas, QS order [91] has two content matrices namely, Schneider and Grantham. However, it is important to investigate which particular properties of both encoders are appropriate in order to generate more comprehensive statistical vectors, rather than utilizing all the available properties.

To fully utilize the potential of both encoders, a strategy similar to forward feature selection method is adopted to find out the most appropriate physicochemical properties. For instance, from 3 properties of APAAC encoder, first we generate statistical vectors by using one property and compute performance of RF classifier. Similarly, we repeat the same process for the second and third property in order to record the performance of RF classifier. On the basis of higher performance, we take the property-specific statistical vectors and combine them with the second best performing property vectors. This is followed by the evaluation on the basis of combined features, if this does not yield any performance gains then the iterative process stops and individual property-based statistical vectors with the highest performance are selected. In contrast, if there are any performance gains with such combinations then the combined encodings are retained and utilized further. Similar procedure is used to generate statistical representations using QS order.

The statistical vectors generated from the encoders may contain irrelevant and redundant features. In order to remove such features and retain only the most informative features, we utilize a dimensionality reduction algorithm named feature agglomeration [135]. While reducing dimensions of original feature space, it is important to find the target dimension of reduced feature space. To find an appropriate feature space, we reduce the dimension of original feature space from 40% to 95% with a step size of 5%. By utilizing RF classifier based on its performance, we chose the most appropriate feature space. It is noteworthy to mention that the process of property selection and appropriate reduced feature space selection is performed only using training data.

Furthermore, the training of meta predictor can be seen as a two-stage process. In the first stage, the statistical vectors generated for virus-host protein sequences using APAAC and QS order are separately passed through two machine learning classifiers i.e., RF and ET [12]. Then the prior representations are concatenated and passed again through the RF and ET classifiers, predictions of both classifiers using individual and combined encodings are utilized to create a new feature space on which SVM classifier is trained to make final predictions.

Figure 9.1 describes graphical illustration of the proposed meta predictor's workflow. More detailed working of the encoding methods is given in subsection 9.2.2. Dimensionality reduction method is explained in section 9.2.2.2. In addition, details about second stage classification is provided in subsection 9.2.3.

9.2.2 Protein Sequence Encoding

The following subsections briefly illustrate the working paradigm of APAAC and QS order sequence encoding methods.

9.2.2.1 Amphiphilic Pseudo-Amino Acid Composition (APAAC)

Chou et al. [92, 233] proposed APAAC encoder that makes use of pre-computed physicochemical values of hydrophobicity, hydrophilicity and side chain mass [92, 233]. Each physicochemical property contains 20 float values associated with 20 unique amino acids (supplementary file 2 Table 1). These values are computed based on diverse types of information related to protein folding and protein's interactions with the environment and other molecules. For each of the three quantitative properties, the values of its corresponding amino acids are normalized to zero mean and unit standard deviation through equation 9.1.

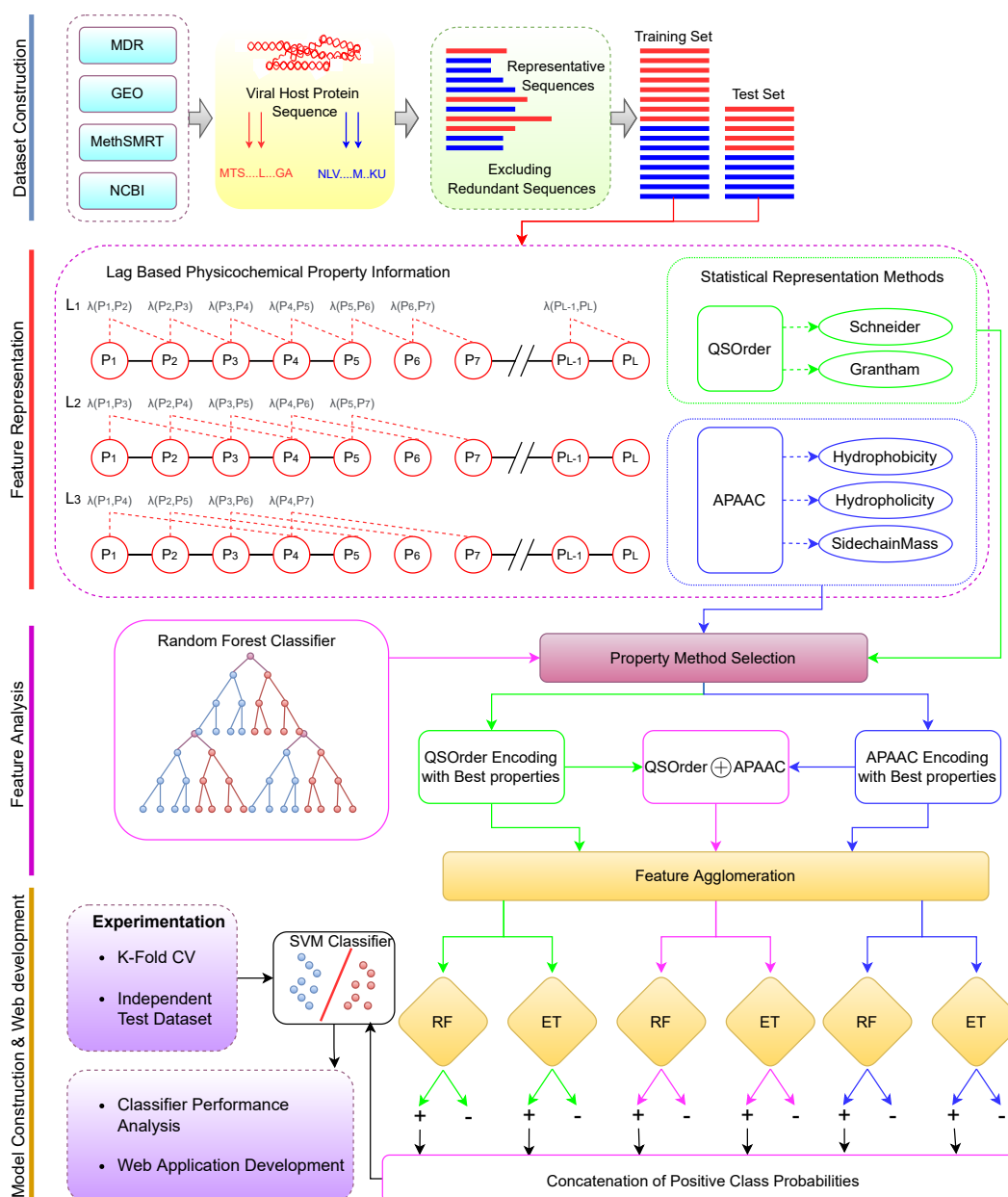


Figure 9.1: The overall working paradigm of the proposed VH-PPIs predictor. **Dataset Construction** To begin with, different datasets are collected from existing studies based on VH-PPIs from several databases such as, HPID, intact and VirusMentha. **Feature Representation** Obtained protein sequences are encoded on the basis of two physicochemical properties based protein sequence encoders i.e., QS order and APAAC. **Feature Analyses** Appropriate physicochemical properties are selected for the APAAC and QS order on the basis of feature analyses. **Model Construction** The VH-PPIs predictor is a SVM model formed on the basis of probabilistic vectors obtained from the RF and ET classifiers. Finally, a web server is established for fast and easy on-go analyses of VH-PPIs.

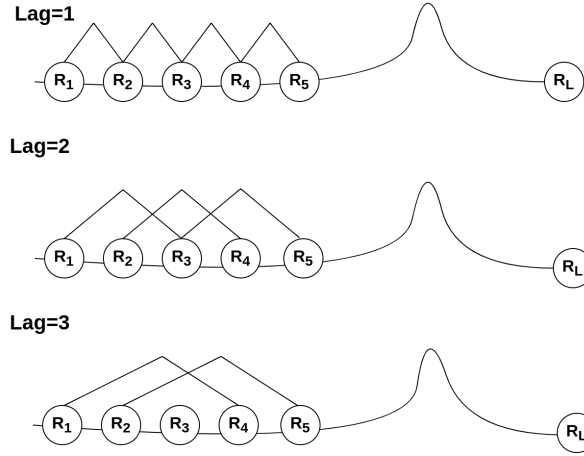


Figure 9.2: The process of computing amino acid combinations based on the lag values.

$$f(x) = \begin{cases} \text{Mean}[p_i] = \frac{\sum_{k=1}^{20} p_i[AA_k]}{20}, \\ S[p_i] = \sqrt{\frac{(\sum_{k=1}^{20} (p_i[AA_k] - \text{Mean}[p_i])^2)}{20}}, \\ P[p_i] = \frac{p_i[AA_k] - \text{Mean}[p_i]}{S[p_i]}, k \in \{1., 2, 3, \dots, 20\}, \\ p_i \in \{\text{hydrophobicity, hydrophilicity, side chain mass}\}. \end{cases} \quad (9.1)$$

whereas, p_i represents the physicochemical property based value of a amino acid (AA_k) which is either hydrophobicity, hydrophilicity or side chain mass. In equation 1, $\text{Mean}[p_i]$ is the mean of 20 amino acids in each property and $S[p_i]$ is the standard deviation, where both can computed using equation 9.1.

In each physicochemical property, using normalized values of all 20 amino acids, order of amino acids within host and viral protein sequences is captured using lag-based phenomenon.

For instance, we have a raw sequence $S=R_1, R_2, R_3, R_4, \dots, R_L$, where $R_{1, \dots, L}$ denotes 20 unique amino acids. If lag=1, then two most contiguous amino acids i.e., $S_{lag1} = R_1R_2, R_2R_3, R_3R_4, R_4R_5$, are taken, for lag=2, second-most contiguous amino acids, i.e., $S_{lag2} = R_1R_3, R_2R_4, R_3R_5$ are taken by skipping 1 amino acid and for lag=3, third-most contiguous amino acids are taken by skipping 2 amino acids i.e., $S_{lag3} = R_1R_4, R_2R_5$ and so on. After generating bigrams, from S_{lag1} , S_{lag2} , S_{lag3} , iteratively, bigrams are taken and in each bigram, physicochemical values of both amino acids are multiplied using a correlation function shown in equation 9.2.

$$P_i[B] = p_i(AA_j) \cdot p_i(AA_k), \quad (9.2)$$

$p_i \in \{\text{hydrophobicity, hydrophilicity, side chain mass}\}$

After computing the correlation functions, for a property, across N number of lags, a single float is computed by averaging property values across all the lag-based amino acid bigrams.

$$Enc[p_i] = \sum_{l=1}^{lag} \frac{P[B]}{seq\ len - lag_i}. \quad (9.3)$$

Furthermore, both types of sequence order and amino acid distributional information can be captured using equation 9.4.

$$Enc[AA] = \frac{frequency\ of\ AA\ in\ protein\ sequence}{1 + w \times Enc[p_i]}, \quad (9.4)$$

here, w is a weight parameter that varies from 0.1 to 1. Similarly, normalization is applied on the original sequence order information by using equation 9.5,

$$Enc[p_i]_{lag_i} = \frac{w \times Enc[p_i]_{lag_i}}{1 + w \times Enc[p_i]}. \quad (9.5)$$

Once the amino acid distribution and sequence order related information are encoded, the final statistical representation is obtained by concatenating the amino acid distributions and correlations among amino acids, that represent the sequence order information of a protein sequence.

$$Encoding[seq]P_i = Enc[AA] \parallel Enc[p_i]_{lag_i} \quad (9.6)$$

The dimension of the final statistical vector for a single physicochemical property is $20 + lag$ -D vector and for 3 physicochemical properties, the final statistical vector is $(20 + lag) \times 3$ dimensional vector. In which, first 20 numbers are the normalized amino acid frequencies and the next following discrete numbers reminisce the amphiphilic amino acid correlations along a protein chain.

9.2.2.2 Quasi-sequence (QS) Order

Owing to similar ideas like APAAC, QS order also encodes the sequence order and discriminative information based on different physicochemical properties [91]. To incorporate more significant sequence order information, QS order makes use of pre-computed values of 4 different physicochemical properties namely, hydrophobicity, hydrophilicity, polarity and side chain volume to compute the coupling factors among the amino acids of a protein sequence [91]. These physicochemical properties describe protein folding and its structural features, particularly surface physical chemistry. These pre-computed values have been averaged and on the basis of Manhattan distance, new values ($20 \times 20 = 400$) have been provided by Schneider et al., [356] and Grantham et al., [155] (for details see supplementary file 2 Table 2 and 3).

In QS order, first the bigrams of amino acids are generated on the basis of lag phenomenon as shown in Figure 9.2 and discussed earlier in APAAC. To compute a coupling factor $P[B]$, distance values between two amino acids are taken from the Tables 2 and 3 given in Supplementary File 2, with respect to bigrams generated via lag value. The coupling factor $P[B]$ can be written as;

$$P[B] = D_i^2(AA_k, AA_j), \quad (9.7)$$

$$D_i \in \{Schneider, Grantham\},$$

where, D is the distance value taken from the Schneider or Grantham's content matrices and B denotes a bigram of amino acids. Corresponding encoding value for a lag can be computed by averaging all the physicochemical distance values for bigrams,

$$Encoding [D_i]_{lag_i} = \frac{\sum_{k=1}^{len\ seq-i} (P[B]_k)}{len\ seq - 1}. \quad (9.8)$$

To get a single float value for the encoding, lag values are averaged depending on the size of lag. For example, for lag=3, first the bigrams are generated with lag=1,2,3, then the corresponding encodings for these bigrams are generated and averaged using following equation.

$$Encoding [D_i] = \sum_{i=1}^{lag} Encoding [D_i]_{lag_i} \quad (9.9)$$

These computed encoding values are normalized along with a weight factor w ,

$$Encoding [D_i]_{lag_i} = \frac{w \times Encoding [D_i]_{lag_i}}{1 + w \times Encoding [D_i]} \quad (9.10)$$

To incorporate the distribution of amino acids, normalized frequencies of 20 different amino acids are computed, according to the following equation,

$$Encoding [AA_k] = \frac{frequency\ of\ AA_k\ in\ protein\ sequence}{1 + w \times Encoding [D_i]}. \quad (9.11)$$

Finally, $(20+lag) \times 2$ dimensional statistical vector is formed by concatenating 20 amino acids distribution values and lag number of correlation factors referring to sequence order information with respect to distance values provided by Schneider and Grantham.

$$Encoding [seq] = Encoding [AA] \parallel Encoding [D_i]_{lag_i}, \quad (9.12)$$

where, $Encoding [AA]$ represents the normalized frequency values of 20 different amino acids and $Encoding [D_i]_{lag_i}$ refers to the sequence order information.

Dimensionality Reduction via Feature Agglomeration Clustering

Hierarchical clustering (HC) is a known group of clustering algorithms that construct clusters on the basis of similarities among the data samples. The end goal of HC is to compute clusters that are completely different from each other and data samples within a single cluster are similar to each other. Similar ideas are inherited by feature agglomeration, where the grouping is applied on the features of the data rather than the data samples. In feature agglomeration, two

steps are iteratively followed to achieve required dimensions of feature space namely, distance computation and pooling. First the distance among all the features are computed using Euclidean or Manhattan distance [351]. On the basis of the minimum distance, two features are combined together on the basis of a pooling function which can be the mean of respective features. This process is repeated unless the features are reduced to desired dimensions.

9.2.3 Iterative Representation Learning

Iterative representation learning is a crucial step for performance improvements of ML models, inspired by layer-wise training of deep learning models. In the current study, the proposed meta predictor works in a two-stage process based on iterative representation learning. In the first stage, the statistical vectors generated for virus-host protein sequences by APAAC and QS order are separately passed through two machine learning models i.e., RF and ET. Then the prior representations are concatenated and passed again through the RF and ET classifiers. As a result, for protein sequences, in total around 6 different positive class probabilities are obtained. In the second stage, these probabilistic values are concatenated with each other to form a new 6-D feature vector for protein sequences. This probabilistic feature representation of protein sequences is used as an input for a support vector machine classifier that provides results for the prediction of VHPPIs.

9.2.4 Benchmark Datasets

In order to develop and evaluate AI-based predictors for virus-host protein-protein interaction prediction, several datasets have been developed in the existing studies [9, 35, 393, 432, 434, 473]. We have collected 7 publicly available benchmark datasets from 4 different studies. These datasets have been extensively utilized in the development/evaluation of the most recent VHPPIs predictors [35, 121, 434, 473].

One dataset is taken from the study of Barman et al., [35], which contains VHPPIs across human and 4 viruses i.e., HIV-1, simian virus 40 (SV40), HBV, HCV, papilloma virus, these VHPPIs were downloaded from Virus-Mint database [71]. Whereas, negative samples were collected from Uniprot [97] based on their dissimilarity with the true VHPPIs.

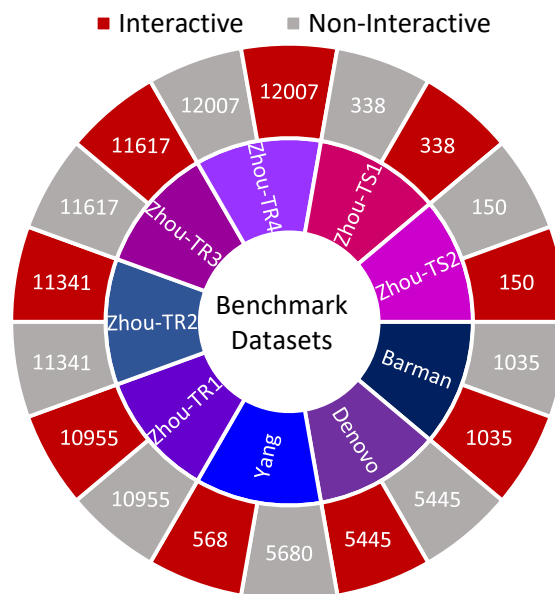


Figure 9.3: Distribution of sequences in interactive and non-interactive classes.

Similarly, another dataset is taken from Fatma et al., work [121], which contains VHPPIs of human and 173 viruses i.e., Paramyxoviridae, Filoviridae, Bunyaviridae, Flaviviridae, Adenoviridae, Orthomyxoviridae, Chordopoxviridae, Papillomaviridae, Herpesviridae, Retroviridae. These VHPPIs were collected from VirusMetha [62] and Uniprot [97]. Negative class samples were generated by random dissimilarity algorithm, which assumed the condition that two viral proteins comprised of similar amino acid sequences could not interact with the same host protein. The similarity between two proteins was decided through distance (dissimilarity) score based on normalized global alignment bit scores. Furthermore, once unique viral proteins were obtained, their interactions were decided based on the dissimilarity (distance) score > 0.8 with host proteins.

SARS-CoV-2 and human proteins related dataset is taken from Yang et al., work [434], where the interactions were collected from HPID [14], VirusHostNet [159], PHISTO [119] and PDB [376] databases. Moreover, negative samples were generated by dissimilarity-based negative sampling across the PPIs retrieved from Uniprot [97, 434].

To make predictor generic and capable to predict interactions over new viruses, we collected 4 datasets from Zhou et al. [473] study. These datasets contain interactions related to 29 different hosts and 332 different viruses. To collect raw sequences and interactions, authors utilized 5 different databases namely PSICQUIC [103], APID [11], IntAct [176], Mentha [62] and Uniprot [97]. Furthermore, for negative data, authors obtained protein sequences of 4 major hosts namely, human, non-human animal, plant and bacteria, from UniProt [97] and removed sequences with a sequence similarity higher than 80% to any positive data using CD-HIT-2D [141]. Moreover, in order to assess the applicability on new/unseen viruses, authors distributed VHPPIs of 29 hosts and 332 viruses into 4 different train and 2 test sets, the distribution of viruses and hosts in these datasets is given below,

TR1: PPIs between human and any virus except H1N1 virus.

TR2: PPIs between human and any virus except Ebola virus.

TR3: PPIs between any host and any virus except H1N1 virus.

TR4: PPIs between any host and any virus except Ebola virus.

TS1: PPIs between human and H1N1 virus.

TS2: PPIs between human and Ebola virus.

Furthermore, Figure 9.3 summarizes the statistics of datasets in terms of number of positive and negative samples. In order to perform experimentation, selected datasets are more appropriate due to multiple reasons such as, recent VHPPI predictors reported their performance scores, making it possible to compare our proposed VHPPIs predictor to existing predictors directly. These datasets contain sufficient VHPPIs which enable to train machine learning models in an optimal way. Furthermore, these datasets contain diverse VHPPIs across a broad selection of viruses and hosts which allows to test the generalizability of the model against multiple hosts and

viruses for the task of VHPPI prediction.

9.3 Evaluation Criteria

Following evaluation criteria of existing predictors, over Barman and SARS-CoV-2 datasets, we perform 5 fold cross validation based experimentation, whereas over Denovo, TR1-TS1, TR2-TS2, TR3-TS1 and TR4-TS2 datasets, we perform independent test based experimentation as their standard train test splits are available.

9.4 Results and Discussions

This section briefly describes the performance of proposed meta predictor at different levels of ensembling. Furthermore, it compares the performance of proposed meta predictor with existing predictors [9, 35, 112, 121, 393, 432, 434, 473] over 7 different benchmark datasets [35, 121, 434, 475].

9.4.1 Performance Analyses of Proposed Meta Predictor using Different Representations at Property Level and Encoder Level

The impact of different physicochemical properties and dimensionality reduction is explored by analyzing the performance of RF and SVM classifiers on the TR4-TS2 dataset. Table 9.1 shows 8 different evaluation measures based performance values produced by RF classifier using statistical representations generated through APAAC and Qsorder encoders using individual and combinations of properties. It also illustrates the performance values of classifier using combined statistical vectors of both encoders. To illustrate the performance impact of dimensionality reduction, it shows the performance of RF classifier using feature agglomeration method based generated comprehensive feature space of statistical vectors produced through individual encoders (APAAC, Qsorder) and combination of both encoders. To illustrate, the performance gains achieved through iterative representation learning of second stage classifier using first stage classifiers predicted probabilities, it shows the performance of SVM classifier.

In Table 9.1, for Qsorder encoder, p_1 represents Schneider-Wrede property and p_2 denotes Grantham property. Similarly for APAAC encoder, p_1 , p_2 and p_3 denotes hydrophobicity, hydrophilicity and side chain mass properties, respectively. RF classifier with statistical vectors generated through Qsorder using p_1 property produces 84.16% accuracy and 83.89% accuracy using p_2 property. It can be concluded that, RF classifier produces different performance when it is fed with two different statistical vectors generated through Qsorder encoder by using two different physicochemical properties p_1 and p_2 . This performance difference illustrates both properties extract and encode different types of information while generating statistical vectors. The performance of the classifier is improved when it is fed with combined statistical vectors

Table 9.1: Performance comparison of different statistical representations across 1st stage RF classifier and with iterative feature learning based 2nd stage SVM classifier.

Encoder	Properties	DR	Random Forest Classifier							
			ACC	PR	F1	SP	SN	AUPRCA	AUROC	MCC
QSOrder	p_1	no	84.16	85.46	84.01	84.06	91.01	98.02	97.55	69.91
	p_2	no	83.89	85.75	83.68	83.89	90.23	98.16	97.74	69.62
	p_1+p_2	no	84.23	85.99	84.03	84.23	91.77	98.24	97.89	70.20
	p_1+p_2	yes	85.23	86.72	85.08	85.23	92.30	97.90	97.49	71.94
APAAC	p_1	no	83.22	85.80	82.91	83.22	90.22	98.09	97.49	68.97
	p_2	no	82.89	85.05	82.62	82.89	89.45	98.04	97.35	67.90
	p_3	no	84.56	86.71	84.34	84.56	91.45	98.30	97.88	71.24
	$p_1+p_2+p_3$	no	85.23	87.17	85.04	85.23	92.23	98.16	97.63	72.38
	p_3+p_1	no	82.89	85.30	82.59	82.89	89.45	98.10	97.49	68.15
	p_3+p_2	no	85.57	87.65	85.37	85.57	92.59	98.24	97.75	73.19
	p_3+p_2	yes	86.24	88.36	86.05	86.24	92.98	98.26	97.96	74.57
APAAC+QSorder		no	86.24	87.88	86.09	86.24	92.90	98.10	97.49	74.10
		yes	86.24	87.88	86.09	86.24	92.91	98.24	97.80	74.10
2nd Stage Predictors			SVM Classifier							
Qsorder	+APAAC-DR-RF,									
Qsorder-DR-RF,	APAAC-DR-RF,		93.62	93.64	93.62	93.62	96.71	98.50	98.14	87.27
Qsorder+APAAC-DR-ET,	Qsorder-									
DR-ET, APAAC-DR-ET										

generated through both properties. Its performance gets further improved when it is fed with combined vectors of both properties reduced through the feature agglomeration method. This performance improvement validates, that both properties extract some redundant features that when eradicated in the newly generated feature space, the performance gets improved.

Similarly, for APAAC encoder among 3 statistical vectors generated through 3 different properties, RF classifier produces better performance with p_3 property and produces the lowest performance with p_2 property. So, according to the working paradigm of the proposed property selection method, top-performing property p_3 vectors will combine with p_1 and p_2 properties vectors iteratively. From the concatenation of p_3 property vector with p_1 and p_2 property vectors, classifier achieves slight performance gain with p_3 and p_2 concatenation. Furthermore, when p_3 and p_2 properties vectors combined with the p_1 property, the performance of the classifier decreased as compared to its performance with p_2 and p_3 properties combinations and the property selection method selected p_2 and p_3 as two optimal properties. These results reveal that to fully utilize the potential of the APAAC encoder, it is essential to utilize the best combination of properties. Furthermore, concatenation of statistical vectors generated through selected best

properties based APAAC and Qsorder encoders fail to improve the performance of the RF classifier as compared to its performance on individual statistical representations.

Dimensionality reduction along with individual encoders has improved the performance of RF classifier as compared to its performance on the same encoders without applying dimensionality reduction. However, it produces almost similar performance with and without dimensionality reduction on combined vectors of APAAC and Qsorder encoders.

To gain further performance enhancement, at the second stage we utilize positive class probabilities predicted by ET and RF classifiers using feature agglomeration based optimized statistical vectors of individual APAAC and Qsorder encoders and both encoders combined vectors. SVM classifier is trained on newly generated probabilistic 6D feature space where it achieves higher performance as compared to the performance values of RF and ET classifiers. In comparison to the performance of RF classifier with sequence representations generated through (APAAC+Qsorder, DR=yes), it achieves performance improvements of 7.38% in accuracy, 5.76% in precision, 7.53% in F1-score, 7.38% in specificity, 3.1% in sensitivity, 0.26% in AUPRC, 0.34% in AUROC and 13.17% in MCC. In comparison to performance of RF classifier with sequence representations generated through (p_1+p_2 , DR=yes) of Qsorder and (p_3+p_2 , DR=yes) of APAAC, it achieves performance improvements with an average margin of 6.10% across all the evaluation measures. Therefore, it is inferred that the SVM classifier along with the iterative representation learning leads to the highest performance for virus-host protein-protein interaction prediction.

9.4.2 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on Barman's Dataset

Table 9.2 shows the performance values of 7 different evaluation measures of the proposed meta predictor and 6 existing VHPPI predictors [9, 21, 35, 433, 473] on Barman's dataset, [35]. From 6 existing predictors, our LGCA-VHPPI [21] predictor achieves better performance in terms of accuracy 82%, specificity 89.37%, f1-score 81.47%, MCC 63.99% and AUROC 88%. Whereas, Zhou et al., [473] predictor produces better performance in terms of precision 82.46%. Among 7 different evaluation measures, Barman et al., predictor [35] only managed to produce the highest sensitivity 89.08% as compared to the sensitivity of 5 other predictors. Comparatively, the proposed meta predictor outperforms 6 previously mentioned predictors [9, 21, 35, 433, 473] in terms of 6 distinct evaluation measures. Overall, in terms of accuracy, the proposed meta predictor achieves an improvement of 0.9%, 1.79% in sensitivity 1.62% increase in precision, 1.27% increase in F1-score, 2.97% in MCC and 0.17% in terms of AUROC.

In terms of robustness on Barman's dataset, the proposed and existing predictors fall into two different categories based on the differences between their specificity and sensitivity scores, i.e., less biased, predictors with a small difference in specificity and sensitivity scores and more biased predictors with a large difference in specificity and sensitivity scores. Individually there are sensitivity and specificity differences of 5.4%, 9.76%, 7%, 33.42%, 7.63%, 7.37% and 7.97% for

Table 9.2: Performance comparison of proposed MP-VHPPI with existing viral-host PPI predictors over a benchmark Barman dataset in terms of 7 different evaluation measures. Performance figures of Barman et al. SVM [35], Barman et al. RF [35], Alguwzizani et al. SVM [9] and Yang et al. RF [433] are taken from Yang et al. [433] work.

Approach	ACC	SN	SP	PR	F1	MCC	AUROC
Yang et al. RF [433]	79.17	81.85	76.45	77.83	79.79	58.40	87.1
Alguwzizani et al. SVM [9]	78.6	73.72	83.48	81.69	77.50	57.50	84.70
Barman et al. SVM [35]	71.00	67.00	74.00	72.00	69.41	44.0	73.00
Barman et al. RF [35]	72.41	89.08	55.66	82.26	66.39	48.00	76.00
Zhou et al. SVM [473]	79.95	76.14	83.77	82.46	79.17	60.1	85.8
Our LGCA-VHPPI [21]	82.00	82.00	89.37	82.40	81.47	63.99	88.00
Proposed MP-VHPPI	82.90	90.87	82.90	84.08	82.74	66.96	88.17

Yang’s RF [433], Alguwzizani et al., SVM [9], Barman et al., SVM [35] and RF [35], Zhou et al., SVM [473], our LGCA-VHPPI [21] predictor and the proposed meta predictor, respectively. On the basis of these difference values, among all predictors, Yang’s RF [433], Barman et al., SVM, Zhou et al., SVM [473], our LGCA-VHPPI [21] and proposed meta predictor can be considered less biased as they have small difference ($<8\%$) in terms of their specificity and sensitivity scores. Contrarily, the other two predictors Barman’s RF [35] and Alguwzizani et al., SVM [9], have large differences between sensitivity and specificity scores and are biased towards either type I or type II error. Type I error arises when a predictor is prone towards the false positive predictions due to low specificity and high sensitivity scores ($T_{IE} = 1 - SP$) and in type II error the predictor is prone to false negative predictions due to low sensitivity and high specificity scores ($T_{II}E = 1 - SN$). Barman’s RF [35] is more prone to type I error due to high sensitivity and lower specificity scores, whereas Alguwzizani et al., SVM [9] is more prone to type II error due to higher specificity and lower sensitivity scores.

9.4.3 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on Denovo’s Dataset

Table 9.3 illustrates performance values of 7 different evaluation measures of the proposed meta predictor and 7 existing VHPPI predictors Yang et al., RF [433], Alguwzizani et al., SVM [9], Fatma et al., SVM [121], Yang et al., CNN [434], Zhou et al., SVM [473], Dong et al., LSTM [112] and our LCGA-VHPPI on Denovo dataset [121].

From 7 existing predictors, our LGCA-VHPPI predictor [21] achieves better performance in terms of accuracy 94.24%, sensitivity 94.24%, f1-score 94.23%, MCC 88.56% and AUROC 98.49%. Whereas, Yang et al., predictor [434] achieves the highest performance values in terms of specificity 97.41% and precision 97.23%. Among all existing predictors, Fatma et al., predictor [121] shows the least performance. In comparison to these predictors, the proposed meta predictor offers performance improvements across 4 different evaluation measures. It achieves a

performance gain of 0.35% in both accuracy and f1-score, 2.99% increment in sensitivity and 0.76% increment in MCC.

Table 9.3: Performance comparison of proposed MP-VHPPI with existing viral-host PPI predictors over benchmark DeNovo dataset [121] in terms of 7 different evaluation measures. Performance figures of DeNovo SVM [121], Alguwzizani et al. SVM [9] and Yang et al. RF on DeNovo dataset [121] are taken from Yang et al. work [433].

Approach	ACC	SN	SP	PR	F1	MCC	AUROC
Yang et al. RF [433]	93.23	90.33	96.17	95.99	93.07	86.60	98.10
Alguwzizani et al. SVM [9]	86.47	86.35	86.59	86.56	86.46	72.90	92.60
Fatma et al. SVM [121]	81.90	80.71	83.06	–	–	–	–
Yang et al. CNN [434]	94.12	90.82	97.41	97.23	93.92	–	–
Zhou et al. SVM [473]	84.47	80.00	88.94	87.86	–	62.92	89.7
Dong et al. LSTM [112]	–	84.12	–	83.92	84.02	–	92.21
Our LGCA-VHPPI [21]	94.24	94.24	96.47	94.32	94.23	88.56	98.49
Proposed MP-VHPPI	94.59	97.23	94.59	94.73	94.58	89.32	98.16

The predictors on the Denovo dataset can be seen in two different categories as done previously in terms of Barman’s dataset on the basis of specificity and sensitivity differences. Individually there exist differences of 5.84 %, 6.59%, 2.35%, 2.23% , 2.64% across Yang et al., predictor [433], Yang et al., CNN [434], Fatma et al., [121], our LGCA-VHPPI [21] and proposed meta predictor. Due to less difference (<3%) in the specificity and sensitivity scores, Alguwzizani et al., [9], Fatma et al., [121], our LGCA-VHPPI [21] and proposed meta predictor can be considered less biased towards type I and type II errors as compared to other two predictors i.e., Yang et al., RF [433] and Yang et al., CNN [434] that are more biased towards type II error due to high specificity and low sensitivity scores.

9.4.4 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on SARS-CoV-2 Dataset

Due to a recent pandemic of SARS-CoV-2, it is important to analyze the performance of a predictor on SARS-CoV-2 and human proteins. Table 9.4 shows performance values of proposed meta predictor, Yang et al., CNN [434] and our LGCA-VHPPI [21], across SARS-CoV-2 and human proteins dataset [434], in terms of 8 distinct evaluation measures.

Out of two existing predictors, Yang et al., predictor based on CNN achieves better accuracy 90.64%. Whereas, our LGCA-VHPPI predictor [21] shows better performance in terms of, sensitivity 93.6%, precision 85.67%, AUPRC 38.01% and f1-score 85.07%. Due to the highly imbalance number of samples for interactive and non-interactive classes in SARS-CoV-2 dataset, Yang et al., predictor [434] performs poorly as evident from its extremely low sensitivity, precision, F1 and AUPRC scores. The proposed meta predictor outperforms existing predictors in terms of accuracy

Table 9.4: Performance comparison of the proposed predictor with existing Yang et al. predictor [434] over the SARS-CoV-2 dataset.

Approach	ACC	SN	SP	PR	F1	MCC	AUPRC	AUROC
Yang et al., CNN [434]	90.64	16.37	98.06	45.81	24.12	-	32.9	-
our LGCA-VHPPI [21]	90.11	93.6	50.04	85.67	85.07	22.21	38.01	80.0
Proposed MP-VHPPI	91.18	95.58	51.74	86.01	87.27	10.08	47.07	82.95

by a margin of 0.54%, 1.98% in sensitivity, 0.34% in precision, 2.2% in f1-score, 9.06% in terms of AUPRC and 2.95% in AUROC.

Individually, there exist differences of 81.69%, 43.56% and 43.84% in specificity and sensitivity scores for Yang et al., predictor [434], our LGCA-VHPPI predictor [21] and the proposed meta predictor. On the basis of that, it can be inferred that the proposed meta predictor and our LGCA-VHPPI predictor [21] are less biased towards type I and type II errors. Whereas, Yang et al., predictor [434] is biased towards type II error due to high specificity and low sensitivity scores.

9.4.5 Proposed MP-VHPPI Predictor Performance Comparison with Existing Predictors on Unseen Viruses Test Sets

To assess the applicability of the VHPPI predictors on unseen viruses where predictors are trained on different types of viruses and evaluation is performed on the test sets that contain viruses (Influenza A virus subtype H1N1 and Ebola virus EBV) which are not part of the training sets. Table 9.5 compares the performance values of the proposed meta predictor with 4 existing predictors i.e., Zhou et al., SVM [473], Tsukiyama et al., LSTM-PHV [393], Dong et al., predictor [112] and our LGCA-VHPPI [21].

Over TR1-TS1 dataset, out of 4 existing predictors Tsukiyama et al., LSTM-PHV [393] performs better in terms of accuracy 86.7% and MCC 73.7%, Dong et al. predictor [112] shows the highest precision 86.28%, f1-score 86.40% and AUROC 94.61%. our LCGA-VHPPI shows the highest performance in terms of specificity and sensitivity i.e., 83.82% and 91.48%. Whereas Zhou et al. predictor [473] shows the least performance across all evaluation measures except sensitivity. In comparison to the existing predictors, the proposed meta predictor outperforms existing predictors across 7 evaluation measures. It achieves an increase of 3.56% in accuracy, 6.44% in specificity, 3.58% in sensitivity, 5.16% in precision, 3.79% in F1-score, 7.99% in MCC and 2.09% in AUROC. Three out of 4 existing predictors, Tsukiyama et al., LSTM-PHV [393], Zhou et al., SVM [473] and our LGCA-VHPPI [21], are biased towards type 1 error due to lower specificity (82.9%, 66.14%, 83.82%) and higher sensitivity scores (90.6%, 89.76%, 91.48%) with differences of 7.7%, 23.62% and 7.66%. In comparison, the proposed meta predictor is robust and generalizable due to the small difference between specificity and sensitivity scores i.e., 4.8% and

Table 9.5: Performance comparison of the proposed MP-VHPPI with existing virus-Host PPI predictors over 4 datasets developed by Zhou et al., [473], to assess the applicability on the unseen viruses. The performance values of the existing approaches i.e., Zhou et al., [473] and Tsukiyama et al. (LSTM-PHV) [393] are taken from their corresponding studies [393, 473]

Dataset	Approach	ACC	SN	SP	PR	F1	MCC	AUROC
TR1-TS1	Zhou et al. (SVM) [473]	77.95	89.76	66.14	72.61	-	57.5	88.6
	Tsukiyama et al. LSTM-PHV [393]	86.7	90.6	82.9	84.1	-	73.7	91.2
	Dong et al., LSTM [112]	-	86.51	-	86.28	86.40	-	94.61
	our LGCA-VHPPI [21]	83.82	91.48	83.82	85.34	83.64	69.14	94.0
	Proposed MP-VHPPI	90.26	95.06	90.26	91.44	90.19	81.69	96.70
TR2-TS2	Zhou et al. (SVM) [473]	78.00	90.67	65.33	72.34	-	57.9	86.7
	Tsukiyama et al. LSTM-PHV [393]	84.0	93.3	74.7	78.7	-	69.2	94.1
	Dong et al., LSTM [112]	-	92.53	-	90.93	91.23	-	96.80
	our LGCA-VHPPI [21]	86.58	93.11	86.57	88.35	86.42	74.9	96.0
	Proposed MP-VHPPI	94.30	97.07	94.30	94.39	94.29	88.69	97.77
TR3-TS1	Zhou et al. (SVM) [473]	77.43	88.98	65.88	72.28	-	56.4	88.4
	Tsukiyama et al. LSTM-PHV [393]	85.7	89.2	82.2	83.3	-	71.6	92.1
	our LGCA-VHPPI [21]	83.29	91.2	83.28	85.31	83.05	68.57	94.0
	Proposed MP-VHPPI	90.53	95.06	90.53	90.78	90.51	81.31	95.98
TR4-TS2	Zhou et al. (SVM) [473]	81.67	94.67	68.67	75.13	-	65.6	89.0
	Tsukiyama et al. LSTM-PHV [393]	90.0	91.3	88.7	89.0	-	80.0	95.6
	our LGCA-VHPPI [21]	85.57	92.59	85.57	87.65	85.37	73.19	96.0
	Proposed MP-VHPPI	93.62	96.71	93.62	93.64	93.62	87.27	98.14

overall higher sensitivity, specificity, AUROC, accuracy and MCC scores.

Over TR2-TS2 dataset, out of four existing predictors Tsukiyama et al., LSTM-PHV performs better in terms of sensitivity 93.3%, whereas Dong et al. [112] predictor performs better in terms of precision 90.93%, f1-score 91.23% and AUROC 96.80%. our LGCA-VHPPI [21] predictor performs better in terms of accuracy 86.58%, specificity 86.57% and MCC 74.9%. Zhou et al. predictor [473], shows the least performance across all the evaluation metrics except sensitivity 90.67%. The proposed meta predictor outperforms existing predictors across all of the evaluation measures. Overall, the proposed meta predictor achieves a gain of 7.72% in accuracy, 3.77% increase in sensitivity, 7.73% in specificity, 3.46% in precision, 3.06% in F1, 13.79% in MCC and 0.97% in AUROC. Among these predictors, the predictors of Tsukiyama [393], Zhou et al., [473] and our LGCA-VHPPI [21], are prone to type 1 error due to high sensitivity and low specificity scores. For instance, the difference in specificity and sensitivity scores of Zhou et al. predictor is 25.34%, 18.6% for Tsukiyama et al., LSTM-PHV [393] and 6.54% for our LGCA-VHPPI predictor [21]. Due to these big differences, these predictors do not generalize well against the human and Ebola virus protein data. Whereas, the proposed meta predictor has a smaller difference of 2.77% between specificity and sensitivity values, which makes it more generalizable than existing

predictors.

Out of three existing predictors, LSTM-PHV predictor performs better across TR3-TS1 in terms of 2 different evaluation metrics i.e., 85.7%, 71.6%, for accuracy and MCC. Similarly, our LGCA-VHPPI predictor [21] shows better performance in terms of sensitivity 91.2%, specificity 83.28%, precision 85.31% and AUROC 94.0%. On the other hand, the proposed meta predictor outperforms existing predictors on 7 different evaluation measures by significant margins. The proposed meta predictor achieves a raise of 4.83% in accuracy, 3.86% in sensitivity, 7.25% in specificity, 5.47% in precision, 9.71% in MCC, 7.46% in f1 and 1.98% in AUROC. Similar to the previous cases, existing predictors are again prone to type 1 errors due to high sensitivity and low specificity scores with differences of 23.1%, 7% and 7.92% for Zhou et al. [473], LSTM-PHV [393] and LGCA-VHPPI [21] predictors. Comparatively, the proposed meta predictor has a smaller difference of 4.53% between specificity and sensitivity scores, which makes the proposed meta predictor more suitable for VHPPI prediction.

Over TR4-TS2 dataset out of three existing predictors, LSTM-PHV [393] achieves better results across 4 evaluation measures i.e., 90.0%, 88.7%, 89.0%, 80.0%, in terms of accuracy, specificity, precision and MCC. LGCA-VHPPI [21] excels in terms of AUROC 96.0%. Whereas, Zhou et al., SVM [473] shows better sensitivity score 94.67%. The proposed predictor achieves performance gains of 3.62% in accuracy, 2.04% in sensitivity, 4.92% in specificity, 4.64% in precision, 8.25% in f1-score, 7.27% in MCC and 2.14% in AUROC. There exists a difference in the specificity and sensitivity scores of these predictors which are 26% for Zhou et al. predictor and 7.02% for our LGCA-VHPPI [21], which makes them more biased towards type I error due to high sensitivity and lower specificity scores. Comparatively, LSTM-PHV and the proposed meta predictor have a lower difference in specificity and sensitivity scores of (<3.1%), which suggests that for TR4-TS2 dataset, both of the predictors are able to generalize well over positive and negative class samples.

9.4.6 Discussion

Since last decade, the development of machine and deep learning-based computational approaches for virus-host protein-protein interaction prediction has been an active area of research [35, 393]. In the marathon of developing robust computational VHPPI predictors, the aim of each newly developed predictor has been to utilize raw virus-host protein sequences and precisely discriminate interactive viral-host protein sequences from non-interactive ones. However, most predictors have been evaluated on a limited type of viruses and hosts, such as 6 different predictors have been evaluated on Barman dataset that contains 5 different viruses and human proteins as host. Seven predictors are evaluated on Denovo dataset that is comprised of 10 different viruses and human proteins as host and 2 predictors are evaluated on SARS-CoV-2 virus. Only 4 predictors are evaluated on the Zhou et al., [473] dataset, that consists of 332 viruses and 29 hosts proteins. These datasets are more suitable to evaluate the robustness,

generalizability and predictive performance of a computational predictor. These datasets were developed with an objective to train models on different types of viruses and evaluate them on the particular viruses which were not part of the training set.

Over unseen virus host protein-protein interaction prediction datasets, the performance of existing predictors is comparably low, as compared to their performance on Barman and Denovo datasets. Recently, we developed a machine learning-based predictor namely LGCA-VHPPI [21], which produced state-of-the-art performance on both Barman and Denovo datasets. We evaluated our predictor on Zhou et al., [473] datasets, where it showed relatively lower performance as compared to its performance on Barman and Denovo datasets. This motivated us to develop an improved predictor that performs better not only on Barman and Denovo datasets but also produces similar performance for unseen viral-host protein-protein interaction predictions.

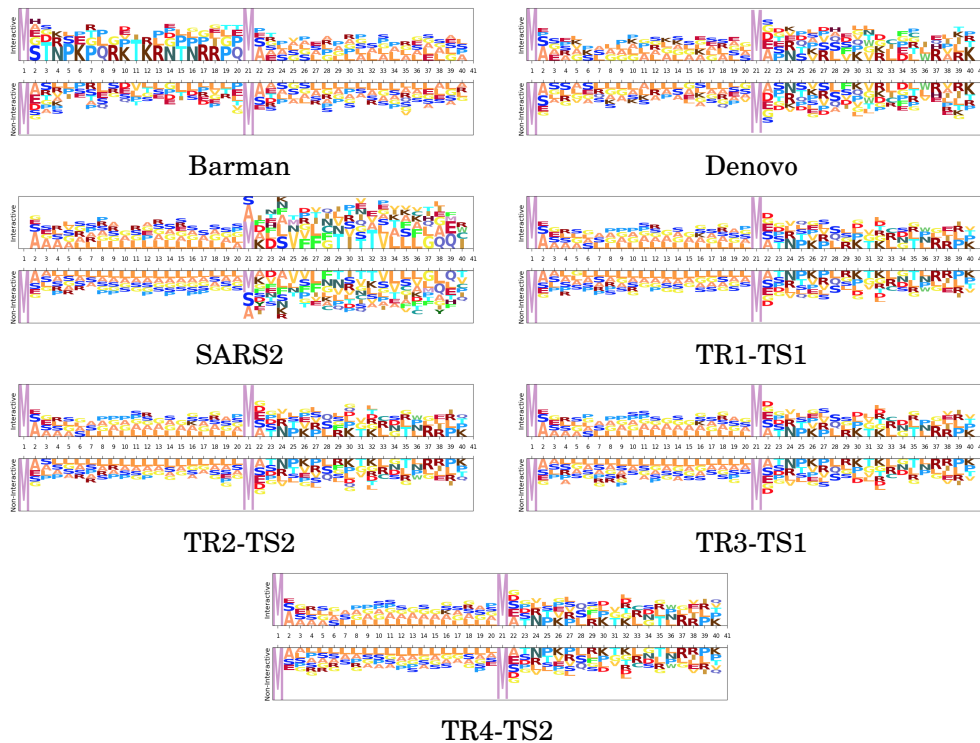


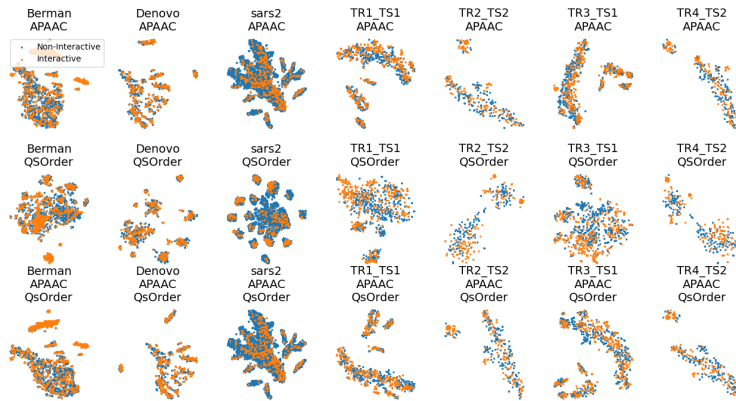
Figure 9.4: Distribution of amino acids in 7 different datasets. For each dataset the distribution of amino acids is shown across interactive and non-interactive protein samples

As discussed in section 9.4, most of the existing predictors are biased towards type I or type II error, this is mainly because in viral and host protein sequences, distribution of amino acids is almost similar for interactive and non-interactive classes. To illustrate this phenomenon, we perform amino acids distribution analysis across both classes with the help of Two Sample Logo [384]. As viral host protein sequences are highly variable in length, so to perform position-aware distribution analysis, we take 20 amino acids from the start of host proteins and discard others and similarly, we take 20 amino acids from the start of viral protein sequences. Figure 9.4

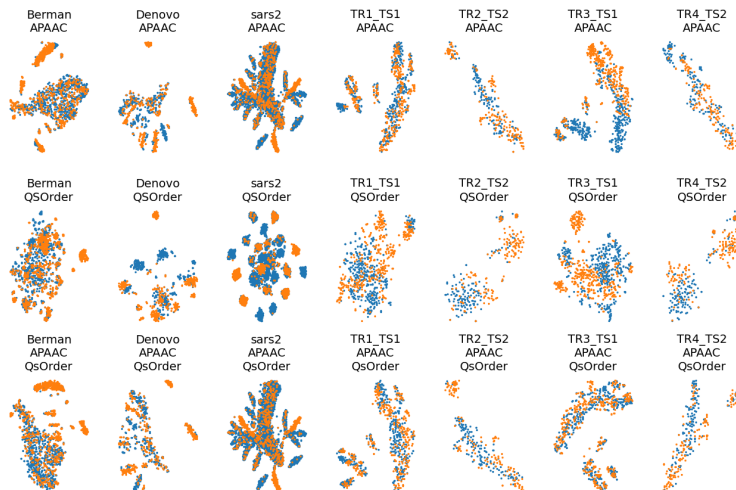
illustrates the distribution of amino acids in interactive and non-interactive classes for 7 different data sets. It can be seen that the distribution of amino acids is approximately similar in interactive and non-interactive classes. Considering Barman's dataset as an example (Figure 9.4:Barman), in interactive and non-interactive samples, there are overlapping amino acids at every position i.e., for position 2, interactive samples contain one of the following amino acid, H, A, E, G, S whereas, non-interactive samples also contain one of the following amino acid, A, E, G, S. In both classes occurrence of 4 amino acids is the same while few samples of the interactive class contain amino acid H, a similar trend exists at other locations as well. Furthermore, other datasets also contain a similar distribution of amino acids as in Barman dataset. It can be concluded that, across all 7 datasets, we observe limited discriminative distribution of amino acids and because of that existing predictors lack in performance due to the utilization of suboptimal sequence encoding methods that generate statistical vectors by neglecting most of the discriminative features about the distribution of amino acids in interactive and non-interactive classes.

It is important to mention that all the amino acids are either polar or non-polar in nature and can carry charges, such as out of 21 unique amino acids, 11 amino acids are polar in nature, 4 AAs carry a positive charge (R, D, H, K), 2 AAs carry a negative charge (D, E) and 5 AAs are neutral (C, Q, S, T, Y). Whereas, 10 amino acids are non-polar in nature (A, G, I, L, M, F, P, W, Y, V). Irrespective of positions aware occurrences, considering the overall distribution of amino acids in the protein sequence, charges can be computed by utilizing the physicochemical properties. Overall charge information of amino acids along with their distribution information can extract and encode more discriminative patterns.

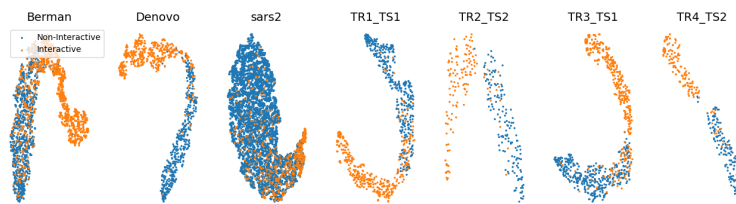
Figure 9.5 shows different clusters of 7 benchmark datasets for the intrinsic analyses of the statistical vectors generated through APAAC and Qsorder sequence encoders. These clusters are computed by first reducing the dimensions of statistical vectors through principal component analysis (PCA) and then by t-distributed stochastic neighbor embedding (TSNE). In Figure 9.5(A) and 9.5(B), rows represent clusters of interactive and non-interactive classes based on statistical vectors generated through individual encoders (APAAC, Qsorder) and a combination of both encoders. Whereas, the columns represent 7 different benchmark datasets namely, Barman, Denovo, SARS-CoV-2, TR1-TS1, TR2-TS2, TR3-TS1 and TR4-TS2. Overall, statistical vectors from APAAC and Qsorder without dimensionality reduction lead to the formation of overlapping clusters for interactive and non-interactive classes. This overlapping reveals that generated statistical vectors are almost similar and contain less discriminative information about interactive and non-interactive classes, as shown in Figure 9.5(A). Furthermore, this overlapping behavior among clusters exists due to the extraction of some irrelevant and redundant features by different physicochemical properties. To eradicate such type of information, we utilize the feature agglomeration method with an objective to transform generated statistical vectors into a more informative and discriminative feature space. Comparatively, statistical representations of APAAC and Qsorder with dimensionality reduction lead to the formation of slightly unique yet



A. Feature space without Dimensionality reduction



B. Feature space with Dimensionality reduction



C. Probabilistic feature space

Figure 9.5: Clusters formation with representations of protein sequences based on APAAC and Qsorder without dimensionality reduction (a), with dimensionality reduction (b) and 6D representations from ET and RF classifier (c).

heavily dependent clusters as shown in Figure 9.5(B). Though these encodings could be used for classification purposes, however, still the performance would not be very promising. In addition, the clusters do not seem independent because a single human protein that interacts with some viral proteins, might not interact with some other viral proteins. This means that positive and negative class samples can have very similar representations due to the presence of such proteins. Although dimensionality reduction produces better feature space, however, still clusters are not very much separable. To further improve the performance of the predictor, we perform iterative representation learning, where we pass 3 different statistical representations separately to RF and ET classifiers and take their predicted class probabilities to develop a new feature space. The generated feature space lead to the formation of unique and independent clusters as shown in Figure 9.5(C), which suggest the presence of comprehensive discriminatory features for interactive and non-interactive VHPPI pairs. Due to the discriminative and informative nature of newly generated feature space, we utilize this feature space to train SVM classifier for virus-host protein-protein interaction prediction.

Overall, as compared to state-of-the-art predictors, the proposed predictor has shown a slight performance improvement on Barman and Denovo datasets and significant performance improvements on Sars-CoV-2 datasets and other 4 datasets namely, TR1-TS1, TR2-TS2, TR3-TS1 and TR4-TS2. We believe that the performance of proposed predictor can be further improved by incorporating representations learned through diverse types of language models such as BERT and XLNET.

This work can be considered another step up in terms of designing a robust tool for VHH-PPI prediction, which in spite of performing quite precisely can be further improved. In the future, three main paradigms can be opted for further performance improvements, first using physicochemical properties based encoders deep learning architectures such as convolution and recurrent neural networks can be tested, secondly, Bayesian optimization can be used to get performance enhancements by optimizing ensembling strategy and lastly, language models such as BERT, XLNET and pre-trained protein language-based models such as ProtTrans, TAPE and ProtBert can be used to design better VHPPI predictors.

9.5 Conclusion

The prime objective of this research is the development of a robust machine learning-based computational framework capable of precisely predicting viral host protein-protein interactions across a wide range of hosts and viruses. Proposed meta predictor makes use of APAAC and QS order sequence encoders for statistical representation generation and feature agglomeration method to refine feature space. Furthermore, meta predictor utilizes the predictions of random forest and extra tree classifiers to feed SVM classifier that makes final predictions. Experimental results reveal the competence of APAAC and QS order encoders for most effectively generating

numerical representations of sequences by capturing amino acids sequence order and distributional information. We have observed dimensionality reduction method removes irrelevant and redundant information which slightly improves the performance of classifiers. The process of iterative representation learning in which predictions of RF and ET classifiers are passed to SVM classifier, significantly improves the accuracy of interactions predictions. The proposed meta predictor is evaluated over 7 benchmark datasets where it outperforms existing predictors with a significant margin of 3.07%, 6.07%, 2.95% and 2.85%, in terms of accuracy, MCC, precision and sensitivity, respectively. We believe that deployment of proposed meta predictor as a web interface will assist researchers and practitioners in analyzing the complex phenomenon of VHPPIs at a larger scale to unravel substantial drug targets and optimize antiviral strategies.

CONCLUSION AND FUTURE WORK

This chapter compiles the conclusive remarks drawn from the problems that are considered in this dissertation and the solutions presented to tackle them. Possible limitations of the presented solutions are discussed along with future research work that would mitigate these limitations.

10.1 Conclusions

The aim of this dissertation is to empower the process of biological sequence analysis using the powers of Artificial Intelligence (AI). Biological sequences of different biomolecules (e.g., DNA, RNA, protein) contain diverse types of information such as, a set of instructions to produce distinct proteins in different amounts, genetic diseases, potential biological pathways to design therapies for various diseases, the ways in which viruses hijack cellular processes, strategies for controlling their propagation and procedures for promoting certain immunity responses. Key idea is to use AI algorithms to explore such information while using only raw biological sequences. Raw sequences of DNA/RNA and proteins are composed of repetitive patterns of 4 unique nucleic acids and 20 unique amino acids, respectively. Therefore, distribution of distinct nucleic and amino acids in raw sequences represents diverse information related to biological processes and treatment of different diseases. It is essential to develop automatic methods which can precisely extract distributional information of nucleic and amino acids and use this information to perform different types of analyses to unlock the hidden potential of biological sequences. The main motivation for this work comes from the author's observation that there is a very limited availability of robust computational frameworks for Genomics and Proteomics sequence analysis despite their immense need in research and industry.

The main contribution of this dissertation is the conceptualization and implementation of a computational framework by using the powers of AI approaches. The generic nature of the

presented framework makes it applicable to a variety of Genomics (DNA, RNA) and Proteomics (protein) sequence analysis tasks. A highlight of the presented framework is that apart from newly developed methodologies, it contains existing sequence encoders and the most widely used predictors that will facilitate researchers to reproduce the performance of existing methodologies for diverse types of tasks and will enable them to develop meta-predictors and compare the performance of their novel approaches with existing approaches on new benchmarks.

In the area of Genomics (DNA, RNA) sequence analysis, proposed framework is used to solve 5 different problems where it produces state-of-the-art performances. To supplement the process of genomic sequence analysis, a novel sequence encoder is presented in this dissertation. Proposed encoder is competent in capturing position specific distributional information of nucleic acids in the DNA sequences and encode such information into statistical vectors. Using statistical vectors generated through proposed sequence encoder, random forest classifier manages to outperform existing computational approaches for the task of DNA modification predictions across multiple species. Another contribution to Genomics sequence analysis is the development of a novel deep learning classifier for histone occupancy and modification prediction. Proposed predictor is also evaluated in another similar application area, namely enhancer identification and strength prediction, where it also produces state-of-the-art performance.

Third contribution is the development of a novel classifier based on DenseNet architecture which is capable of more precisely discriminating 13 different classes of small non-coding RNAs. To more precisely analyze the impact of utilizing alternative paths for the flow of gradient, two different depth based ResNet architectures are adapted which also produces decent performances for small non-coding RNA classification. Fourth contribution is the development of a novel autoencoder and convolutional neural network based predictor for accurate identification of circular RNAs. A comprehensive experimentation is performed using proposed predictor with an aim to find appropriate regions of genome that contain more comprehensive information about circular RNAs. Fifth contribution is the development of an explainable classifier for predicting multi-compartment subcellular localizations of four different RNA types across multiple species. This classifier utilizes a unique graph based encoding method to capture comprehensive local and global interaction patterns and translational invariances of nucleotides which are difficult to capture in traditional sequence encoding methods due to their focus on occurrence and physicochemical properties of nucleic acids. Proposed predictor is capable of highlighting unique patterns of nucleic acids in the RNA sequences associated with particular subcellular localization.

In the area of Proteomics sequence analysis, proposed framework is used to solve 2 different problems namely: host protein-protein interaction prediction and viral host protein-protein interaction prediction. A novel predictor based on hybrid architecture is developed that makes use of LSTM, CNN and Attention layers to more precisely distinguish interactive protein pairs from non-interactive ones across multiple species. Proposed approach uses FastText embedding generation method for the comprehensive characterization of protein sequences. A generic

meta predictor (MP-VHPPI) is proposed that can more accurately predict viral-host protein-protein interactions across multiple hosts and viruses. To generate the most effective numerical representations of viral-host protein sequences, meta predictor reaps the benefits of two different sequence encoding methods that are competent in capturing amino acids sequence order and distributional information. Furthermore, it takes advantage of dimensionality reduction to transform original feature space into more informative feature space. It generates a new feature space that contains predicted probabilities of two tree-based classifiers by feeding them with optimized feature spaces of individual encoders and their combined encodings. The probabilistic feature space is fed to SVM classifier that makes final predictions.

10.2 Limitations

This section summarizes the limitations of the different methods presented in this dissertation.

A novel encoder is proposed for the characterization of DNA sequences that produces good performance with different machine learning classifiers for the task of predicting 3 different DNA modifications. However, the working paradigm of the proposed encoder relies on the assumption that, for a particular dataset, distribution of nucleic acids remains somewhat similar between the sequences of the same class while differs among the sequences of different classes. Based on this assumption, proposed encoder uses class labels of training sequences to learn class conditional densities based distribution of nucleic acids and maps these distributions to generate statistical representation of test sequences. A comprehensive empirical evaluation on large number of benchmark datasets related to multiple species thoroughly validates the assumption. However, this assumption proves more effective when statistical representation of test sequences is generated using the class conditional densities based distribution of nucleic acids learned on large training dataset. If the distribution of nucleic acids is learned on small training data, then the proposed approach will incorrectly characterize the distribution of nucleic acids within test sequences that are not seen during the training phase.

Furthermore, decisions of deep learning predictors are hard to interpret because of their black box working paradigms. Recently, in the domain of Natural Language Processing (NLP), significant efforts have been made to develop explainable predictors. However, the domain of Genomics (DNA, RNA) and Proteomics (protein) sequence analysis is lagging in this regard due to the scarcity of explainable deep learning predictors. In this dissertation, 5 deep learning predictors are developed however, only 2 predictors can explain their decisions. Integration of Attention layer in other three predictors will not only make the predictors explainable but may also increase the predictive performance of these predictors as it assists the predictors to focus on the most discriminative features.

Deep learning predictors produce better performance when trained on large datasets and their performance reduces on account of small datasets. To solve this problem, pre-trained k-

mer embeddings are generated that facilitate deep learning predictors to perform better even when they are trained on small datasets. We utilize a similar strategy to develop two predictive approaches for RNA subcellular location prediction and host protein-protein interaction prediction tasks, respectively. However, using pre-trained k-mer embeddings, only embedding layer of deep learning predictors gets pre-trained weights while other layers are still randomly initialized. To solve this problem, in the NLP domain, several language models have been proposed. Hence, by adapting these models or developing paradigms similar to these models, performance of both proposed predictors can be further improved.

10.3 Future Work

This section provides an overview of various compelling research directions which will be considered in future.

The presented framework is generic and contains diverse types of feature selection and dimensionality reduction algorithms. However, potential of these approaches is not utilized in the development of end-to-end pipelines developed for multiple applications in this dissertation. For instance, using proposed encoder and random forest classifier, a web based application is developed for DNA modification prediction for multiple species. Incorporation of an appropriate feature selection or dimensionality reduction algorithm may improve the performance of proposed predictor. Furthermore, while evaluating the performance of proposed encoder with multiple classifiers, we note that different classifiers produce the best performance for various species data and different types of DNA modification prediction. On average the performance of random forest classifier is better, hence, we use a random forest classifier with proposed encoder to construct final predictor. By reaping the benefits of multiple classifiers, a meta predictor can be developed which may further improve the performance for DNA modification prediction in multiple species.

DenseNet based predictor is proposed and evaluated only for small non-coding RNA classification. Proposed predictor can be utilized for other DNA, RNA and protein sequence analysis tasks where it may produce state-of-the-art performance similar to small non-coding RNA classification. Furthermore, incorporation of attention layer in the proposed predictor may slightly improve its performance and could make predictor decisions explainable. While developing protein-protein interaction predictor, a unique idea is proposed to generate fixed-length sequences by taking most informative regions. A similar idea can be used to develop better predictors for circular RNA and protein interaction prediction and viral host protein-protein interaction prediction tasks.

In order to reap the benefits of multiple sequence encoders, proposed generic framework contains a unique approach called meta sequence descriptor that generates a new feature space by combining the weighted feature spaces of a certain number of homogeneous or heterogeneous sequence encoders. Meta sequence descriptor approach is not evaluated extensively, in future, this particular method can be utilized to develop more appropriate statistical representations of

biological sequences which will help the classifiers to achieve more promising performance for different sequence analysis tasks. Proposed framework contains 12 different types of embedding generation approaches, however, in this dissertation, only two embedding generation methodologies are utilized. To practically analyze which embedding method is better for multiple sequence analysis tasks, a detailed comparative study using multiple types of deep learning predictors needs to be performed. Following the success of diverse types of language models in the domain of NLP, there is also a need of investigating the efficacy of language models in the domain of Genomics and Proteomics sequence analysis.

BIBLIOGRAPHY

- [1] Z. ABBAS, H. TAYARA, AND K. CHONG, *Zayyunet a unified deep learning model for the identification of epigenetic modifications using raw genomic sequences*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, (2021).
- [2] Z. ABBAS, H. TAYARA, AND K. TO CHONG, *Spinenet-6ma: A novel deep learning tool for predicting dna n6-methyladenine sites in genomes*, IEEE Access, 8 (2020), pp. 201450–201457.
- [3] H. ABDI, *Singular value decomposition (svd) and generalized singular value decomposition*, Encyclopedia of measurement and statistics, (2007), pp. 907–912.
- [4] A. AHMAD, H. LIN, AND S. SHATABDA, *Locate-r: Subcellular localization of long non-coding rnas using nucleotide compositions*, Genomics, 112 (2020), pp. 2583–2589.
- [5] B. AL-SALEMI, S. A. M. NOAH, AND M. J. AB AZIZ, *Rfboost: an improved multi-label boosting algorithm and its application to text categorisation*, Knowledge-Based Systems, 103 (2016), pp. 104–117.
- [6] W. ALAM, H. TAYARA, AND K. T. CHONG, *i4mc-deep: An intelligent predictor of n4-methylcytosine sites using a deep learning approach with chemical properties*, Genes, 12 (2021), p. 1117.
- [7] B. ALBERTS, *The cell as a collection of protein machines: preparing the next generation of molecular biologists*, cell, 92 (1998), pp. 291–294.
- [8] W. ALGHAMDI, E. ALZHRANI, M. Z. ULLAH, AND Y. D. KHAN, *4mc-rf: improving the prediction of 4mc sites using composition and position relative features and statistical moment*, Analytical Biochemistry, 633 (2021), p. 114385.
- [9] S. ALGUWAIZANI, B. PARK, X. ZHOU, D.-S. HUANG, AND K. HAN, *Predicting interactions between virus and host proteins using repeat patterns and composition of amino acids*, Journal of healthcare engineering, 2018 (2018).
- [10] B. A. ALHAMWE, R. KHALAILA, J. WOLF, V. VON BÜLOW, H. HARB, F. ALHAMDAN, C. S. HUI, S. L. PRESCOTT, A. FERRANTE, H. RENZ, ET AL., *Histone modifications*

BIBLIOGRAPHY

- and their role in epigenetics of atopy and allergic diseases*, *Allergy, Asthma & Clinical Immunology*, 14 (2018), pp. 1–16.
- [11] D. ALONSO-LOPEZ, M. A. GUTIÉRREZ, K. P. LOPES, C. PRIETO, R. SANTAMARÍA, AND J. DE LAS RIVAS, *Apid interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks*, *Nucleic acids research*, 44 (2016), pp. W529–W535.
- [12] E. ALPAYDIN, *Machine learning*, MIT Press, 2021.
- [13] N. AMIN, A. MCGRATH, AND Y.-P. P. CHEN, *Evaluation of deep learning in non-coding rna classification*, *Nature Machine Intelligence*, 1 (2019), p. 246.
- [14] M. G. AMMARI, C. R. GRESHAM, F. M. MCCARTHY, AND B. NANDURI, *Hpidb 2.0: a curated database for host–pathogen interactions*, *Database*, 2016 (2016).
- [15] D. AMMONS, J. RAMPERSAD, AND G. E. FOX, *5s rrna gene deletions cause an unexpectedly high fitness loss in escherichia coli*, *Nucleic acids research*, 27 (1999), pp. 637–642.
- [16] S. A. ANDREI, E. SIJBESMA, M. HANN, J. DAVIS, G. O’MAHONY, M. W. PERRY, A. KARAWAJCZYK, J. EICKHOFF, L. BRUNSVELD, R. G. DOVESTON, ET AL., *Stabilization of protein-protein interactions in drug discovery*, *Expert Opinion on Drug Discovery*, 12 (2017), pp. 925–940.
- [17] W. J. ANSORGE, *Next-generation dna sequencing techniques*, *New biotechnology*, 25 (2009), pp. 195–203.
- [18] E. ASGARI AND M. R. MOFRAD, *Continuous distributed representation of biological sequences for deep proteomics and genomics*, *PloS one*, 10 (2015), p. e0141287.
- [19] M. N. ASIM, M. A. IBRAHIM, M. IMRAN MALIK, A. DENGEL, AND S. AHMED, *Advances in computational methodologies for classification and sub-cellular locality prediction of non-coding rnas*, *International Journal of Molecular Sciences*, 22 (2021), p. 8719.
- [20] M. N. ASIM, M. A. IBRAHIM, M. I. MALIK, A. DENGEL, AND S. AHMED, *Enhancer-dsnet: A supervisedly prepared enriched sequence representation for the identification of enhancers and their strength*, in *International Conference on Neural Information Processing*, Springer, 2020, pp. 38–48.
- [21] M. N. ASIM, M. A. IBRAHIM, M. I. MALIK, A. DENGEL, AND S. AHMED, *Lgca-vhppi: A local-global residue context aware viral-host protein-protein interaction predictor*, *PloS one*, 17 (2022), p. e0270275.

-
- [22] M. N. ASIM, M. A. IBRAHIM, M. I. MALIK, I. RAZZAK, A. DENGEL, AND S. AHMED, *Histone-net: a multi-paradigm computational framework for histone occupancy and modification prediction*, *Complex & Intelligent Systems*, (2022), pp. 1–21.
- [23] M. N. ASIM, M. A. IBRAHIM, M. I. MALIK, C. ZEHE, O. CLOAREC, J. TRYGG, A. DENGEL, AND S. AHMED, *El-rmlocnet: An explainable lstm network for rna-associated multi-compartment localization prediction*, *Computational and Structural Biotechnology Journal*, (2022).
- [24] M. N. ASIM, M. A. IBRAHIM, C. ZEHE, O. CLOAREC, R. SJOGREN, J. TRYGG, A. DENGEL, AND S. AHMED, *L2s-mirloc: a lightweight two stage mirna sub-cellular localization prediction framework*, in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.
- [25] M. N. ASIM, M. A. IBRAHIM, C. ZEHE, J. TRYGG, A. DENGEL, AND S. AHMED, *Bot-net: a lightweight bag of tricks-based neural network for efficient lncrna–mirna interaction prediction*, *Interdisciplinary Sciences: Computational Life Sciences*, (2022), pp. 1–22.
- [26] M. N. ASIM, M. I. MALIK, A. DENGEL, AND S. AHMED, *K-mer neural embedding performance analysis using amino acid codons*, in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.
- [27] M. N. ASIM, M. I. MALIK, C. ZEHE, J. TRYGG, A. DENGEL, AND S. AHMED, *Mirlocpredictor: a convnet-based multi-label microrna subcellular localization predictor by incorporating k-mer positional information*, *Genes*, 11 (2020), p. 1475.
- [28] M. N. ASIM, M. I. MALIK, C. ZEHE, J. TRYGG, A. DENGEL, AND S. AHMED, *A robust and precise convnet for small non-coding rna classification (rpc-snrc)*, *IEEE Access*, 9 (2020), pp. 19379–19390.
- [29] Ž. AVSEC, R. KREUZHUBER, J. ISRAELI, N. XU, J. CHENG, A. SHRIKUMAR, A. BANERJEE, D. S. KIM, L. URBAN, A. KUNDAJE, ET AL., *Kipoi: accelerating the community exchange and reuse of predictive models for genomics*, *BioRxiv*, (2018), p. 375345.
- [30] N. S. BABAIHA, R. AGHDAM, AND C. ESLAHCHI, *Nn-rnalloc: neural network-based model for prediction of mrna sub-cellular localization using distance-based sub-sequence profiles*, *bioRxiv*, (2021).
- [31] A. BACHMAYR-HEYDA, A. T. REINER, K. AUER, N. SUKHBAATAR, S. AUST, T. BACHLEITNER-HOFMANN, I. MESTERI, T. W. GRUNT, R. ZEILLINGER, AND D. PILS, *Correlation of circular rna abundance with proliferation—exemplified with colorectal and ovarian cancer, idiopathic lung fibrosis and normal human tissues*, *Scientific reports*, 5 (2015), pp. 1–10.

BIBLIOGRAPHY

- [32] A. BAIROCH AND R. APWEILER, *The swiss-prot protein sequence data bank and its new supplement trembl*, *Nucleic acids research*, 24 (1996), pp. 21–25.
- [33] D. R. BAISYA AND S. LONARDI, *Prediction of histone post-translational modifications using deep learning*, *Bioinformatics*, 36 (2020), pp. 5610–5617.
- [34] A. J. BANNISTER AND T. KOUZARIDES, *Regulation of chromatin by histone modifications*, *Cell research*, 21 (2011), pp. 381–395.
- [35] R. K. BARMAN, S. SAHA, AND S. DAS, *Prediction of interactions between viral and host proteins using supervised machine learning methods*, *PloS one*, 9 (2014), p. e112034.
- [36] A. H. BASIT, W. A. ABBASI, A. ASIF, S. GULL, AND F. U. A. A. MINHAS, *Training host-pathogen protein–protein interaction predictors*, *Journal of bioinformatics and computational biology*, 16 (2018), p. 1850014.
- [37] S. BASITH, B. MANAVALAN, T. H. SHIN, AND G. LEE, *Sdm6a: a web-based integrative machine-learning framework for predicting 6ma sites in the rice genome*, *Molecular Therapy-Nucleic Acids*, 18 (2019), pp. 131–141.
- [38] F. BENITES AND E. SAPOZHNIKOVA, *Haram: A hierarchical aram neural network for large-scale text classification*, in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Nov 2015, pp. 847–854.
- [39] D. BENVENISTE, H.-J. SONNTAG, G. SANGUINETTI, AND D. SPROUL, *Transcription factor binding predicts histone modifications in human cell lines*, *Proceedings of the National Academy of Sciences*, 111 (2014), pp. 13367–13372.
- [40] T. BERGGÅRD, S. LINSE, AND P. JAMES, *Methods for the detection and analysis of protein–protein interactions*, *Proteomics*, 7 (2007), pp. 2833–2842.
- [41] B. E. BERNSTEIN, C. L. LIU, E. L. HUMPHREY, E. O. PERLSTEIN, AND S. L. SCHREIBER, *Global nucleosome occupancy in yeast*, *Genome biology*, 5 (2004), pp. 1–11.
- [42] M. BHASIN AND G. P. RAGHAVA, *Classification of nuclear receptors based on amino acid composition and dipeptide composition*, *Journal of Biological Chemistry*, 279 (2004), pp. 23262–23266.
- [43] H. BHASKAR, D. C. HOYLE, AND S. SINGH, *Machine learning in bioinformatics: A brief survey and recommendations for practitioners*, *Computers in biology and medicine*, 36 (2006), pp. 1104–1125.
- [44] A. A. BHAT, S. N. YOUNES, S. S. RAZA, L. ZARIF, S. NISAR, I. AHMED, R. MIR, S. KUMAR, S. K. SHARAWAT, S. HASHEM, ET AL., *Role of non-coding rna networks in leukemia progression, metastasis and drug resistance*, *Molecular Cancer*, 19 (2020), pp. 1–21.

-
- [45] L. BIN, *Bioseq-analysis: a platform for dna, rna and protein sequence analysis based on machine learning approaches*, Briefings in bioinformatics, 20 (2019), pp. 1280–1294.
- [46] H. BINDER, L. STEINER, J. PRZYBILLA, T. ROHLF, S. PROHASKA, AND J. GALLE, *Transcriptional regulation by histone modifications: towards a theory of chromatin re-organization during stem cell differentiation*, Physical biology, 10 (2013), p. 026006.
- [47] P. BOJANOWSKI, E. GRAVE, A. JOULIN, AND T. MIKOLOV, *Enriching word vectors with subword information*, Transactions of the Association for Computational Linguistics, 5 (2017), pp. 135–146.
- [48] R. P. BONIDIA, D. S. DOMINGUES, D. S. SANCHES, AND A. C. DE CARVALHO, *Mathfeature: feature extraction package for dna, rna and protein sequences based on mathematical descriptors*, Briefings in Bioinformatics, 23 (2022), p. bbab434.
- [49] R. P. BONIDIA, L. D. SAMPAIO, D. S. DOMINGUES, A. R. PASCHOAL, F. M. LOPES, A. C. DE CARVALHO, AND D. S. SANCHES, *Feature extraction approaches for biological sequences: a comparative study of mathematical features*, Briefings in Bioinformatics, 22 (2021), p. bbab011.
- [50] R. P. BONIDIA, L. D. H. SAMPAIO, D. S. DOMINGUES, A. R. PASCHOAL, F. M. LOPES, A. C. P. DE LEON FERREIRA, D. S. SANCHES, ET AL., *Feature extraction approaches for biological sequences: A comparative study of mathematical models*, bioRxiv, (2020).
- [51] R. P. BONIDIA, D. S. SANCHES, AND A. C. DE CARVALHO, *Mathfeature: feature extraction package for biological sequences based on mathematical descriptors*, bioRxiv, (2020).
- [52] A. BOTCHKAREV, *Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology*, arXiv preprint arXiv:1809.03006, (2018).
- [53] M. R. BOUTELL, J. LUO, X. SHEN, AND C. M. BROWN, *Learning multi-label scene classification*, Pattern recognition, 37 (2004), pp. 1757–1771.
- [54] M. BREHOVE, T. WANG, J. NORTH, Y. LUO, S. J. DREHER, J. C. SHIMKO, J. J. OTTESEN, K. LUGER, AND M. G. POIRIER, *Histone core phosphorylation regulates dna accessibility*, Journal of Biological Chemistry, 290 (2015), pp. 22612–22621.
- [55] M. C. BRIDGES, A. C. DAULAGALA, AND A. KOURTIDIS, *Lnccation: lncrna localization and function*, Journal of Cell Biology, 220 (2021), p. e202009045.
- [56] A. BRÜCKNER, C. POLGE, N. LENTZE, D. AUERBACH, AND U. SCHLATTNER, *Yeast two-hybrid, a powerful tool for systems biology*, International journal of molecular sciences, 10 (2009), pp. 2763–2788.

BIBLIOGRAPHY

- [57] M. BUSTA, L. NEUMANN, AND J. MATAS, *Fasttext: Efficient unconstrained scene text detector*, in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1206–1214.
- [58] E. BYVATOV AND G. SCHNEIDER, *Support vector machine applications in bioinformatics.*, Applied bioinformatics, 2 (2003), pp. 67–77.
- [59] C. CAI, L. HAN, Z. JI, AND Y. CHEN, *Enzyme family classification by support vector machines*, Proteins: Structure, Function, and Bioinformatics, 55 (2004), pp. 66–76.
- [60] C. CAI, L. HAN, Z. L. JI, X. CHEN, AND Y. Z. CHEN, *Sum-prot: web-based support vector machine software for functional classification of a protein from its primary sequence*, Nucleic acids research, 31 (2003), pp. 3692–3697.
- [61] J. CAI, D. WANG, R. CHEN, Y. NIU, X. YE, R. SU, G. XIAO, AND L. WEI, *A bioinformatics tool for the prediction of dna n6-methyladenine modifications based on feature fusion and optimization protocol*, Frontiers in bioengineering and biotechnology, 8 (2020), p. 502.
- [62] A. CALDERONE AND G. CESARENI, *Mentha: the interactome browser*, EMBnet. journal, 18 (2012), p. 128.
- [63] S. CAO, W. LU, AND Q. XU, *Grarep: Learning graph representations with global structural information*, in Proceedings of the 24th ACM international on conference on information and knowledge management, 2015, pp. 891–900.
- [64] Y. CAO, S. FANNING, S. PROOS, K. JORDAN, AND S. SRIKUMAR, *A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies*, Frontiers in Microbiology, (2017), p. 1829.
- [65] Z. CAO, X. PAN, Y. YANG, Y. HUANG, AND H.-B. SHEN, *The Inclocator: a subcellular localization predictor for long non-coding rnas based on a stacked ensemble classifier*, Bioinformatics, 34 (2018), pp. 2185–2194.
- [66] M. W. CARROLL, D. A. MATTHEWS, J. A. HISCOX, M. J. ELMORE, G. POLLAKIS, A. RAMBAUT, R. HEWSON, I. GARCÍA-DORIVAL, J. A. BORE, R. KOUNDOUNO, ET AL., *Temporal and spatial analysis of the 2014–2015 ebola virus outbreak in west africa*, Nature, 524 (2015), pp. 97–101.
- [67] M. CHAABANE, *End-to-end learning framework for circular rna classification from other long non-coding rnas using multi-modal deep learning.*, (2018).
- [68] M. CHAABANE, R. M. WILLIAMS, A. T. STEPHENS, AND J. W. PARK, *circdeep: deep learning approach for circular rna classification from other long non-coding rna*, Bioinformatics, 36 (2020), pp. 73–80.

- [69] G. CHAO, Y. LUO, AND W. DING, *Recent advances in supervised dimension reduction: A survey*, Machine learning and knowledge extraction, 1 (2019), pp. 341–358.
- [70] D. D. CHAPLIN, *1. overview of the human immune response*, Journal of allergy and clinical immunology, 117 (2006), pp. S430–S435.
- [71] A. CHATR-ARYAMONTRI, A. CEOL, D. PELUSO, A. NARDOZZA, S. PANNI, F. SACCO, M. TINTI, A. SMOLYAR, L. CASTAGNOLI, M. VIDAL, ET AL., *Virusmint: a viral protein interaction database*, Nucleic acids research, 37 (2009), pp. D669–D673.
- [72] J. CHEN, Q. ZOU, AND J. LI, *Deepm6baseq-el: prediction of human n6-methyladenosine (m6a) sites with lstm and ensemble learning*, Frontiers of Computer Science, 16 (2022), pp. 1–7.
- [73] K. CHEN, B. S. ZHAO, AND C. HE, *Nucleic acid modifications in regulation of gene expression*, Cell chemical biology, 23 (2016), pp. 74–85.
- [74] K. M. CHEN, E. M. COFER, J. ZHOU, AND O. G. TROYANSKAYA, *Selene: a pytorch-based deep learning library for sequence data*, Nature methods, 16 (2019), pp. 315–318.
- [75] L. CHEN, Y.-H. ZHANG, G. HUANG, X. PAN, S. WANG, T. HUANG, AND Y.-D. CAI, *Discriminating cirrnas from other lncrnas using a hierarchical extreme learning machine (h-elm) algorithm with feature selection*, Molecular genetics and genomics, 293 (2018), pp. 137–149.
- [76] W. CHEN, P.-M. FENG, H. LIN, AND K.-C. CHOU, *irspot-psednc: identify recombination spots with pseudo dinucleotide composition*, Nucleic acids research, 41 (2013), pp. e68–e68.
- [77] W. CHEN, T.-Y. LEI, D.-C. JIN, H. LIN, AND K.-C. CHOU, *Pseknc: a flexible web server for generating pseudo k-tuple nucleotide composition*, Analytical biochemistry, 456 (2014), pp. 53–60.
- [78] W. CHEN, X. SONG, H. LV, AND H. LIN, *irna-m2g: identifying n2-methylguanosine sites based on sequence-derived information*, Molecular Therapy-Nucleic Acids, 18 (2019), pp. 253–258.
- [79] W. CHEN, H. TRAN, Z. LIANG, H. LIN, AND L. ZHANG, *Identification and analysis of the n6-methyladenosine in the saccharomyces cerevisiae transcriptome*, Scientific reports, 5 (2015), pp. 1–8.
- [80] W. CHEN, H. YANG, P. FENG, H. DING, AND H. LIN, *idna4mc: identifying dna n4-methylcytosine sites based on nucleotide chemical properties*, Bioinformatics, 33 (2017), pp. 3518–3523.

BIBLIOGRAPHY

- [81] W.-J. CHEN, Y.-H. SHAO, C.-N. LI, AND N.-Y. DENG, *Mltsvm: a novel twin support vector machine to multi-label learning*, Pattern Recognition, 52 (2016), pp. 61–74.
- [82] X. CHEN, P. HAN, T. ZHOU, X. GUO, X. SONG, AND Y. LI, *circrnadb: a comprehensive database for human circular rnas with protein-coding annotations*, Scientific reports, 6 (2016), pp. 1–6.
- [83] Y.-Z. CHEN, Z. CHEN, Y.-A. GONG, AND G. YING, *Sumohydro: a novel method for the prediction of sumoylation sites based on hydrophobic properties*, PloS one, 7 (2012), p. e39195.
- [84] Z. CHEN, Y.-Z. CHEN, X.-F. WANG, C. WANG, R.-X. YAN, AND Z. ZHANG, *Prediction of ubiquitination sites by using the composition of k-spaced amino acid pairs*, PloS one, 6 (2011), p. e22930.
- [85] Z. CHEN, P. ZHAO, C. LI, F. LI, D. XIANG, Y.-Z. CHEN, T. AKUTSU, R. J. DALY, G. I. WEBB, Q. ZHAO, ET AL., *ilearnplus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization*, Nucleic acids research, 49 (2021), pp. e60–e60.
- [86] Z. CHEN, P. ZHAO, F. LI, A. LEIER, T. T. MARQUEZ-LAGO, Y. WANG, G. I. WEBB, A. I. SMITH, R. J. DALY, K.-C. CHOU, ET AL., *ifeature: a python package and web server for features extraction and selection from protein and peptide sequences*, Bioinformatics, 34 (2018), pp. 2499–2502.
- [87] Z. CHEN, P. ZHAO, F. LI, T. T. MARQUEZ-LAGO, A. LEIER, J. REVOTE, Y. ZHU, D. R. POWELL, T. AKUTSU, G. I. WEBB, ET AL., *ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of dna, rna and protein sequence data*, Briefings in bioinformatics, 21 (2020), pp. 1047–1057.
- [88] Z. CHEN, Y. ZHOU, J. SONG, AND Z. ZHANG, *hcksaap_ubsite: improved prediction of human ubiquitination sites by exploiting amino acid pattern and properties*, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1834 (2013), pp. 1461–1467.
- [89] L. CHENG AND K.-S. LEUNG, *Quantification of non-coding rna target localization diversity and its application in cancers*, Journal of molecular cell biology, 10 (2018), pp. 130–138.
- [90] P. CHEUNG, C. D. ALLIS, AND P. SASSONE-CORSI, *Signaling to chromatin through histone modifications*, Cell, 103 (2000), pp. 263–271.
- [91] K.-C. CHOU, *Prediction of protein subcellular locations by incorporating quasi-sequence-order effect*, Biochemical and biophysical research communications, 278 (2000), pp. 477–483.

-
- [92] K.-C. CHOU, *Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes*, *Bioinformatics*, 21 (2005), pp. 10–19.
- [93] K.-C. CHOU AND Y.-D. CAI, *Prediction of protein subcellular locations by go–fund–pseaa predictor*, *Biochemical and Biophysical Research Communications*, 320 (2004), pp. 1236–1239.
- [94] K. W. CHURCH, *Word2vec*, *Natural Language Engineering*, 23 (2017), pp. 155–162.
- [95] B. M. M. CONQUE, A. Y. KASHIWABARA, AND F. M. LOPES, *Feature extraction from complex networks: A case of study in genomic sequences classification*, arXiv preprint arXiv:1412.5627, (2014).
- [96] M. G. S. CONSORTIUM, R. WATERSTON, K. LINDBLAD-TOH, E. BIRNEY, AND J. ROGERS, *Initial sequencing and comparative analysis of the mouse genome*, *Nature*, 420 (2002), pp. 520–562.
- [97] U. CONSORTIUM, *Uniprot: a worldwide hub of protein knowledge*, *Nucleic acids research*, 47 (2019), pp. D506–D515.
- [98] J. CURSONS, K. A. PILLMAN, K. G. SCHEER, P. A. GREGORY, M. FOROUTAN, S. HEDIYEH-ZADEH, J. TOUBIA, E. J. CRAMPIN, G. J. GOODALL, C. P. BRACKEN, ET AL., *Combinatorial targeting by micrnas co-ordinates post-transcriptional control of emt*, *Cell systems*, 7 (2018), pp. 77–91.
- [99] M. CUSACK, H. W. KING, P. SPINGARDI, B. M. KESSLER, R. J. KLOSE, AND S. KRIAUCIUNIS, *Distinct contributions of dna methylation and histone acetylation to the genomic occupancy of transcription factors*, *Genome research*, 30 (2020), pp. 1393–1406.
- [100] H. S. S. DANIEL D. LEE, *Algorithms for non-negative matrix factorization*, *Neural computation*, (2000).
- [101] S. DASGUPTA, *Experiments with random projection*, arXiv preprint arXiv:1301.3849, (2013).
- [102] N. E. DAVEY, G. TRAVÉ, AND T. J. GIBSON, *How viruses hijack cell regulation*, *Trends in biochemical sciences*, 36 (2011), pp. 159–169.
- [103] N. DEL TORO, M. DUMOUSSEAU, S. ORCHARD, R. C. JIMENEZ, E. GALEOTA, G. LAUNAY, J. GOLL, K. BREUER, K. ONO, L. SALWINSKI, ET AL., *A new reference implementation of the psicquic web service*, *Nucleic acids research*, 41 (2013), pp. W601–W606.
- [104] F. DEL VECCHIO, V. MASTROIACO, A. DI MARCO, C. COMPAGNONI, D. CAPECE, F. ZAZZERONI, C. CAPALBO, E. ALESSE, AND A. TESSITORE, *Next-generation sequencing: Recent*

BIBLIOGRAPHY

- applications to the analysis of colorectal cancer*, Journal of translational medicine, 15 (2017), pp. 1–19.
- [105] L. DENG, J. ZHAO, AND J. ZHANG, *Predict the protein-protein interaction between virus and host through hybrid deep neural network*, in 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), IEEE, 2020, pp. 11–16.
- [106] R. DIAS AND A. TORKAMANI, *Artificial intelligence in clinical and genomic diagnostics*, Genome medicine, 11 (2019), pp. 1–12.
- [107] M.-C. DIDOT, C. M. FERGUSON, S. LY, A. H. COLES, A. O. SMITH, A. A. BICKNELL, L. M. HALL, E. SAPP, D. ECHEVERRIA, A. A. PAI, ET AL., *Nuclear localization of huntingtin mrna is specific to cells of neuronal origin*, Cell reports, 24 (2018), pp. 2553–2560.
- [108] D. S. DIMITROV, *Virus entry: molecular mechanisms and biomedical applications*, Nature Reviews Microbiology, 2 (2004), pp. 109–122.
- [109] J. DING, S. ZHOU, AND J. GUAN, *mirfam: an effective automatic mirna classification method based on n-grams and a multiclass svm*, BMC bioinformatics, 12 (2011), pp. 1–11.
- [110] Y. DING, J. TANG, AND F. GUO, *Predicting protein-protein interactions via multivariate mutual information of protein sequences*, BMC bioinformatics, 17 (2016), pp. 1–13.
- [111] Q. DONG, S. ZHOU, AND J. GUAN, *A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation*, Bioinformatics, 25 (2009), pp. 2655–2662.
- [112] T. N. DONG, G. BROGDEN, G. GEROLD, AND M. KHOSLA, *A multitask transfer learning framework for the prediction of virus-human protein-protein interactions*, BMC bioinformatics, 22 (2021), pp. 1–24.
- [113] E. S. DORN AND J. G. COOK, *Nucleosomes in the neighborhood: new roles for chromatin modifications in replication origin control*, Epigenetics, 6 (2011), pp. 552–559.
- [114] L. DOU, X. LI, H. DING, L. XU, AND H. XIANG, *Prediction of m5c modifications in rna sequences by combining multiple sequence features*, Molecular Therapy-Nucleic Acids, 21 (2020), pp. 332–342.
- [115] H. DU, F. CHEN, H. LIU, AND P. HONG, *Network-based virus-host interaction prediction with application to sars-cov-2*, Patterns, 2 (2021), p. 100242.
- [116] X. DU, S. SUN, C. HU, Y. YAO, Y. YAN, AND Y. ZHANG, *Deepppi: boosting prediction of protein-protein interactions with deep neural networks*, Journal of chemical information and modeling, 57 (2017), pp. 1499–1510.

- [117] I. DUBCHAK, I. MUCHNIK, S. R. HOLBROOK, AND S.-H. KIM, *Prediction of protein folding class using global description of amino acid sequence*, Proceedings of the National Academy of Sciences, 92 (1995), pp. 8700–8704.
- [118] I. DUBCHAK, I. MUCHNIK, C. MAYOR, I. DRALYUK, AND S.-H. KIM, *Recognition of a protein fold in the context of the scop classification*, Proteins: Structure, Function, and Bioinformatics, 35 (1999), pp. 401–407.
- [119] S. DURMUŞ TEKİR, T. ÇAKIR, E. ARDIÇ, A. S. SAYILIRBAŞ, G. KONUK, M. KONUK, H. SARIYER, A. UĞURLU, İ. KARADENİZ, A. ÖZGÜR, ET AL., *Phisto: pathogen–host interaction search tool*, Bioinformatics, 29 (2013), pp. 1357–1358.
- [120] V. D'ARGENIO, *The high-throughput analyses era: are we ready for the data struggle?*, High-throughput, 7 (2018), p. 8.
- [121] F.-E. EID, M. ELHEFNAWI, AND L. S. HEATH, *Denovo: virus-host sequence-based protein–protein interaction prediction*, Bioinformatics, 32 (2016), pp. 1144–1150.
- [122] S. A. ELELA AND R. N. NAZAR, *Role of the 5.8 s rrna in ribosome translocation*, Nucleic acids research, 25 (1997), pp. 1788–1794.
- [123] J. ESPADALER, O. ROMERO-ISART, R. M. JACKSON, AND B. OLIVA, *Prediction of protein–protein interactions using distant conservation of sequence patterns and structure relationships*, Bioinformatics, 21 (2005), pp. 3360–3368.
- [124] M. ESTELLER, *Non-coding rnas in human disease*, Nature reviews genetics, 12 (2011), p. 861.
- [125] A. ESULI, T. FAGNI, AND F. SEBASTIANI, *Mp-boost: A multiple-pivot boosting algorithm and its application to text categorization*, in International Symposium on String Processing and Information Retrieval, Springer, 2006, pp. 1–12.
- [126] H. EVANS AND M. SHAPIRO, *Viruses*, in Manual of techniques in insect pathology, Elsevier, 1997, pp. 17–53.
- [127] Y. FAN, M. CHEN, AND Q. ZHU, *Inclocpred: predicting lncrna subcellular localization using multiple sequence feature information*, IEEE Access, 8 (2020), pp. 124702–124711.
- [128] G. FANG, F. ZENG, X. LI, AND L. YAO, *Word2vec based deep learning network for dna n4-methylcytosine sites identification*, Procedia Computer Science, 187 (2021), pp. 270–277.
- [129] Y. FANG AND M. J. FULLWOOD, *Roles, functions, and mechanisms of long non-coding rnas in cancer*, Genomics, proteomics & bioinformatics, 14 (2016), pp. 42–54.

- [130] P. FENG, H. YANG, H. DING, H. LIN, W. CHEN, AND K.-C. CHOU, *idna6ma-pseknc: Identifying dna n6-methyladenosine sites by incorporating nucleotide physicochemical properties into pseknc*, *Genomics*, 111 (2019), pp. 96–102.
- [131] P. FENG, H. YANG, H. DING, H. LIN, W. CHEN, AND K.-C. CHOU, *iDNA6mA-PseKNC: Identifying DNA n6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC*, *Genomics*, 111 (2019), pp. 96–102.
- [132] P. FENG, J. ZHANG, H. TANG, W. CHEN, AND H. LIN, *Predicting the organelle location of noncoding rnas using pseudo nucleotide compositions*, *Interdisciplinary Sciences: Computational Life Sciences*, 9 (2017), pp. 540–544.
- [133] S. FENG, Y. LIANG, W. DU, W. LV, AND Y. LI, *Lnclocation: efficient subcellular location prediction of long non-coding rna-based multi-source heterogeneous feature fusion*, *International journal of molecular sciences*, 21 (2020), p. 7271.
- [134] Z.-P. FENG AND C.-T. ZHANG, *Prediction of membrane protein types based on the hydrophobic index of amino acids*, *Journal of protein chemistry*, 19 (2000), pp. 269–275.
- [135] M. FEURER, A. KLEIN, K. EGGENSBERGER, J. SPRINGENBERG, M. BLUM, AND F. HUTTER, *Efficient and robust automated machine learning*, *Advances in neural information processing systems*, 28 (2015).
- [136] A. FIANNACA, M. LA ROSA, L. LA PAGLIA, R. RIZZO, AND A. URSO, *nrc: non-coding rna classifier based on structural features*, *BioData mining*, 10 (2017), p. 27.
- [137] A. FILONENKO, K. GUDKOV, A. LEBEDEV, I. ZAGAYNOV, AND N. ORLOV, *Fastext: Fast and small text extractor*, in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 4, IEEE, 2019, pp. 49–54.
- [138] G. FORMAN, *An extensive empirical study of feature selection metrics for text classification*, *Journal of machine learning research*, 3 (2003), pp. 1289–1305.
- [139] A. FRANKISH, M. DIEKHANS, A.-M. FERREIRA, R. JOHNSON, I. JUNGREIS, J. LOVELAND, J. M. MUDGE, C. SISU, J. WRIGHT, J. ARMSTRONG, ET AL., *Gencode reference annotation for the human and mouse genomes*, *Nucleic acids research*, 47 (2019), pp. D766–D773.
- [140] Y. FREUND AND R. E. SCHAPIRE, *A decision-theoretic generalization of on-line learning and an application to boosting*, *Journal of computer and system sciences*, 55 (1997), pp. 119–139.
- [141] L. FU, B. NIU, Z. ZHU, S. WU, AND W. LI, *Cd-hit: accelerated for clustering the next-generation sequencing data*, *Bioinformatics*, 28 (2012), pp. 3150–3152.

- [142] X. FU, W. ZHU, L. CAI, B. LIAO, L. PENG, Y. CHEN, AND J. YANG, *Improved pre-mirnas identification through mutual information of pre-mirna sequences and structures*, *Frontiers in genetics*, 10 (2019), p. 119.
- [143] Y. GAL AND Z. GHAHRAMANI, *A theoretically grounded application of dropout in recurrent neural networks*, in *Advances in neural information processing systems*, 2016, pp. 1019–1027.
- [144] F. GAO AND C.-T. ZHANG, *Comparison of various algorithms for recognizing short coding sequences of human genes*, *Bioinformatics*, 20 (2004), pp. 673–681.
- [145] S. GARCIA-HERRERO, B. SIMON, AND J. GARCIA-PLANELLAS, *The reproductive journey in the genomic era: From preconception to childhood*, *Genes*, 11 (2020), p. 1521.
- [146] A. GARG, N. SINGHAL, R. KUMAR, AND M. KUMAR, *mrnaloc: a novel machine-learning based in-silico tool to predict mrna subcellular localization*, *Nucleic acids research*, 48 (2020), pp. W239–W243.
- [147] A. GAVIN, M. BOSCHE, R. KRAUSE, P. GRANDI, M. MARZIOCH, A. BAUER, J. SCHULTZ, J. RICK, A. MICHON, C. CRUCIAT, ET AL., *Klein k hudak m dickson d*, Rudi T, Gnau V, Bauch A, Bastuck S, Huhse B, Leutwein C, Heurtier MA, Copley RR, Edelmann A, Querfurth E, Rybin V, Drewes G, Raida M, Bouwmeester T, Bork P, Seraphin B, Kuster B, Neubauer G, Superti-Furga G. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415 (2002), pp. 141–7.
- [148] A. A. GEORGES AND L. FRAPPIER, *Affinity purification–mass spectroscopy methods for identifying epstein–barr virus–host interactions*, (2017), pp. 79–92.
- [149] P. GEURTS, D. ERNST, AND L. WEHENKEL, *Extremely randomized trees*, *Machine learning*, 63 (2006), pp. 3–42.
- [150] X. GLOROT, A. BORDES, AND Y. BENGIO, *Deep sparse rectifier neural networks*, in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 315–323.
- [151] I. GOODFELLOW, Y. BENGIO, AND A. COURVILLE, *Deep learning*, MIT press, 2016.
- [152] S. GOODWIN, J. D. MCPHERSON, AND W. R. MCCOMBIE, *Coming of age: ten years of next-generation sequencing technologies*, *Nature Reviews Genetics*, 17 (2016), pp. 333–351.
- [153] P. GOYAL AND E. FERRARA, *Graph embedding techniques, applications, and performance: A survey*, *Knowledge-Based Systems*, 151 (2018), pp. 78–94.
- [154] P. GRABOWSKI AND J. RAPPILBER, *A primer on data analytics in functional genomics: how to move from data to insight?*, *Trends in biochemical sciences*, 44 (2019), pp. 21–32.

BIBLIOGRAPHY

- [155] R. GRANTHAM, *Amino acid difference formula to help explain protein evolution*, *science*, 185 (1974), pp. 862–864.
- [156] L. GROSENICK, B. KLINGENBERG, K. KATOVICH, B. KNUTSON, AND J. E. TAYLOR, *Interpretable whole-brain prediction analysis with graphnet*, *NeuroImage*, 72 (2013), pp. 304–321.
- [157] A. GROVER AND J. LESKOVEC, *node2vec: Scalable feature learning for networks*, in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
- [158] B. L. GUDENAS AND L. WANG, *Prediction of lncrna subcellular localization with deep learning from sequence features*, *Scientific reports*, 8 (2018), pp. 1–10.
- [159] T. GUIRIMAND, S. DELMOTTE, AND V. NAVRATIL, *Virhostnet 2.0: surfing on the web of virus/host molecular interactions data*, *Nucleic acids research*, 43 (2015), pp. D583–D587.
- [160] S.-H. GUO, E.-Z. DENG, L.-Q. XU, H. DING, H. LIN, W. CHEN, AND K.-C. CHOU, *inucpeknc: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition*, *Bioinformatics*, 30 (2014), pp. 1522–1529.
- [161] Y. GUO, L. YU, Z. WEN, AND M. LI, *Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences*, *Nucleic acids research*, 36 (2008), pp. 3025–3030.
- [162] S. GUPTA, J. DENNIS, R. E. THURMAN, R. KINGSTON, J. A. STAMATOYANNOPOULOS, AND W. S. NOBLE, *Predicting human nucleosome occupancy from primary sequence*, *PLoS computational biology*, 4 (2008), p. e1000134.
- [163] I. GUYON AND A. ELISSEEFF, *An introduction to variable and feature selection*, *Journal of machine learning research*, 3 (2003), pp. 1157–1182.
- [164] A. A. HAKIM, A. ERWIN, K. I. ENG, M. GALINIUM, AND W. MULIADY, *Automated document classification for news article in bahasa indonesia based on term frequency inverse document frequency (tf-idf) approach*, in *2014 6th international conference on information technology and electrical engineering (ICITEE)*, IEEE, 2014, pp. 1–4.
- [165] L. Y. HAN, C. Z. CAI, S. L. LO, M. C. CHUNG, AND Y. Z. CHEN, *Prediction of rna-binding proteins from primary sequence by a support vector machine approach*, *Rna*, 10 (2004), pp. 355–368.
- [166] P. C. HANSEN, *Truncated singular value decomposition solutions to discrete ill-posed problems with ill-determined numerical rank*, *SIAM Journal on Scientific and Statistical Computing*, 11 (1990), pp. 503–518.

- [167] R. A. HARATY, M. DIMISHKIEH, AND M. MASUD, *An enhanced k-means clustering algorithm for pattern discovery in healthcare data*, International Journal of distributed sensor networks, 11 (2015), p. 615740.
- [168] M. M. HASAN, S. BASITH, M. S. KHATUN, G. LEE, B. MANAVALAN, AND H. KURATA, *Meta-i6ma: an interspecies predictor for identifying dna n 6-methyladenine sites of plant genomes by exploiting informative features in an integrative machine-learning framework*, Briefings in Bioinformatics, 22 (2021), p. bbaa202.
- [169] M. M. HASAN, B. MANAVALAN, W. SHOOMBATONG, M. S. KHATUN, AND H. KURATA, *i4mc-mouse: Improved identification of dna n4-methylcytosine sites in the mouse genome using multiple encoding schemes*, Computational and structural biotechnology journal, 18 (2020), pp. 906–912.
- [170] S. HASHEMIFAR, B. NEYSHABUR, A. A. KHAN, AND J. XU, *Predicting protein–protein interactions through sequence-based deep learning*, Bioinformatics, 34 (2018), pp. i802–i810.
- [171] K. HE, X. ZHANG, S. REN, AND J. SUN, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
- [172] W. HE AND C. JIA, *Enhancerpred2. 0: predicting enhancers and their strength based on position-specific trinucleotide propensity and electron–ion interaction potential feature selection*, Molecular Biosystems, 13 (2017), pp. 767–774.
- [173] W. HE, C. JIA, AND Q. ZOU, *4mcpred: machine learning methods for dna n4-methylcytosine sites prediction*, Bioinformatics, 35 (2019), pp. 593–601.
- [174] J. W. HENCH, *Cells and tissues*, in Biomaterials, Artificial Organs and Tissue Engineering, Elsevier, 2005, pp. 59–70.
- [175] R. J. HENRY, *Next-generation sequencing for understanding and accelerating crop domestication*, Briefings in functional genomics, 11 (2012), pp. 51–56.
- [176] H. HERMJAKOB, L. MONTECCHI-PALAZZI, C. LEWINGTON, S. MUDALI, S. KERRIEN, S. ORCHARD, M. VINGRON, B. ROECHERT, P. ROEPSTORFF, A. VALENCIA, ET AL., *Intact: an open source molecular interaction database*, Nucleic acids research, 32 (2004), pp. D452–D455.
- [177] M. HIGASHIHARA, J. D. REBOLLEDO-MENDEZ, Y. YAMADA, AND K. SATOU, *Application of a feature selection method to nucleosome data: Accuracy improvement and comparison with other methods*, WSEAS Transactions on Biology and Biomedicine, 5 (2008), pp. 95–104.

BIBLIOGRAPHY

- [178] Y. HO, A. GRUHLER, A. HEILBUT, G. BADER, L. MOORE, S. ADAMS, A. MILLAR, P. TAYLOR, K. BENNETT, K. BOUTILIER, ET AL., *S[>] rensen bd, matthiesen j, hendrickson rc, gleeson f, pawson t, moran mf, durocher d, mann m, hogue cw, figeys d, tyers m (2002) systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry*, *Nature*, 415, pp. 180–183.
- [179] L. M. HOLDT, A. KOHLMAIER, AND D. TEUPSER, *Circular rnas as therapeutic agents and targets*, *Frontiers in physiology*, 9 (2018), p. 1262.
- [180] L. M. HOLDT, A. KOHLMAIER, AND D. TEUPSER, *Molecular roles and function of circular rnas in eukaryotic cells*, *Cellular and Molecular Life Sciences*, 75 (2018), pp. 1071–1098.
- [181] L. M. HOLDT, A. STAHRINGER, K. SASS, G. PICHLER, N. A. KULAK, W. WILFERT, A. KOHLMAIER, A. HERBST, B. H. NORTHOFF, A. NICOLAOU, ET AL., *Circular non-coding rna anril modulates ribosomal rna maturation and atherosclerosis in humans*, *Nature communications*, 7 (2016), pp. 1–14.
- [182] S. HOMBACH AND M. KRETZ, *Non-coding rnas: classification, biology and functioning*, *Non-coding RNAs in colorectal cancer*, (2016), pp. 3–17.
- [183] J. HONG, R. GAO, AND Y. YANG, *Crephan: Cross-species prediction of enhancers by using hierarchical attention networks*, *Bioinformatics*, (2021).
- [184] D. S. HORNE, *Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities*, *Biopolymers: Original Research on Biomolecules*, 27 (1988), pp. 451–477.
- [185] M. HOSSIN AND M. N. SULAIMAN, *A review on evaluation metrics for data classification evaluations*, *International journal of data mining & knowledge management process*, 5 (2015), p. 1.
- [186] R. HOSUR, J. PENG, A. VINAYAGAM, U. STELZL, J. XU, N. PERRIMON, J. BIENKOWSKA, AND B. BERGER, *A computational framework for boosting confidence in high-throughput protein-protein interaction datasets*, *Genome biology*, 13 (2012), pp. 1–14.
- [187] H. HOTELLING, *Analysis of a complex of statistical variables into principal components.*, *Journal of educational psychology*, 24 (1933), p. 417.
- [188] L. HU, Y. LIU, S. HAN, L. YANG, X. CUI, Y. GAO, Q. DAI, X. LU, X. KOU, Y. ZHAO, ET AL., *Jump-seq: genome-wide capture and amplification of 5-hydroxymethylcytosine sites*, *Journal of the American Chemical Society*, 141 (2019), pp. 8694–8697.
- [189] L. HU, X. WANG, Y. HUANG, P. HU, AND Z.-H. YOU, *A novel network-based algorithm for predicting protein-protein interactions using gene ontology*, *Frontiers in Microbiology*, (2021), p. 2441.

- [190] L. HU, J. ZHANG, X. PAN, H. YAN, AND Z.-H. YOU, *Hiscf: leveraging higher-order structures for clustering analysis in biological networks*, *Bioinformatics*, 37 (2021), pp. 542–550.
- [191] G. HUANG, Z. LIU, L. VAN DER MAATEN, AND K. Q. WEINBERGER, *Densely connected convolutional networks*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
- [192] Q. HUANG, W. ZHOU, F. GUO, L. XU, AND L. ZHANG, *6ma-pred: identifying dna n6-methyladenine sites based on deep learning*, *PeerJ*, 9 (2021), p. e10813.
- [193] Q.-Y. HUANG, Z.-H. YOU, S. LI, AND Z. ZHU, *Using chou’s amphiphilic pseudo-amino acid composition and extreme learning machine for prediction of protein-protein interactions*, in *2014 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2014, pp. 2952–2956.
- [194] S. HUANG, B. YANG, B. CHEN, N. BLIIM, U. UEERHAM, T. ARENDT, AND M. JANITZ, *The emerging role of circular rnas in transcriptome regulation*, *Genomics*, 109 (2017), pp. 401–407.
- [195] Y.-A. HUANG, Z.-H. YOU, X. CHEN, K. CHAN, AND X. LUO, *Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding*, *BMC bioinformatics*, 17 (2016), pp. 1–11.
- [196] Y.-A. HUANG, Z.-H. YOU, X. GAO, L. WONG, AND L. WANG, *Using weighted sparse representation model combined with discrete cosine transformation to predict protein-protein interactions from protein sequence*, *BioMed research international*, 2015 (2015).
- [197] Z. HUANG, W. XU, AND K. YU, *Bidirectional lstm-crf models for sequence tagging*, *arXiv preprint arXiv:1508.01991*, (2015).
- [198] F. HUBÉ AND C. FRANCASTEL, *Coding and non-coding rnas, the frontier has never been so blurred*, *Frontiers in Genetics*, 9 (2018), p. 140.
- [199] M. HUE, M. RIFFLE, J.-P. VERT, AND W. S. NOBLE, *Large-scale prediction of protein-protein interactions from structures*, *BMC bioinformatics*, 11 (2010), pp. 1–9.
- [200] M. A. IBRAHIM, M. U. G. KHAN, F. MEHMOOD, M. N. ASIM, AND W. MAHMOOD, *Ghs-net a generic hybridized shallow neural network for multi-label biomedical text classification*, *Journal of biomedical informatics*, 116 (2021), p. 103699.
- [201] I. IEREMIE, R. M. EWING, AND M. NIRANJAN, *Transformergo: predicting protein-protein interactions by modelling the attention between sets of gene ontology terms*, *Bioinformatics*, 38 (2022), pp. 2269–2277.

- [202] S. IOFFE AND C. SZEGEDY, *Batch normalization: Accelerating deep network training by reducing internal covariate shift*, (2015), pp. 448–456.
- [203] S. A. ISLAM, B. J. HEIL, C. M. KEARNEY, AND E. J. BAKER, *Protein classification using modified n-grams and skip-grams*, *Bioinformatics*, 34 (2018), pp. 1481–1487.
- [204] T. ITO, K. TASHIRO, S. MUTA, R. OZAWA, T. CHIBA, M. NISHIZAWA, K. YAMAMOTO, S. KUHARA, AND Y. SAKAKI, *Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins*, *Proceedings of the National Academy of Sciences*, 97 (2000), pp. 1143–1147.
- [205] A. IVANOV, S. MEMCZAK, E. WYLER, F. TORTI, H. T. PORATH, M. R. OREJUELA, M. PIECHOTTA, E. Y. LEVANON, M. LANDTHALER, C. DIETERICH, ET AL., *Analysis of intron sequences reveals hallmarks of circular rna biogenesis in animals*, *Cell reports*, 10 (2015), pp. 170–177.
- [206] G. JAMES, D. WITTEN, T. HASTIE, AND R. TIBSHIRANI, *An introduction to statistical learning*, vol. 112, Springer, 2013.
- [207] N. JAWORSKA AND A. CHUPETLOVSKA-ANASTASOVA, *A review of multidimensional scaling (mds) and its utility in various psychological domains*, *Tutorials in quantitative methods for psychology*, 5 (2009), pp. 1–10.
- [208] R. S. JAYANI, P. L. RAMANUJAM, AND S. GALANDE, *Studying histone modifications and their genomic functions by employing chromatin immunoprecipitation and immunoblotting*, *Methods in cell biology*, 98 (2010), pp. 35–56.
- [209] C. JIA AND W. HE, *Enhancerpred: a predictor for discovering enhancers based on the combination and selection of multiple features*, *Scientific reports*, 6 (2016), p. 38741.
- [210] J. JIA, Z. LIU, X. XIAO, B. LIU, AND K.-C. CHOU, *ippi-esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into pseaac*, *Journal of theoretical biology*, 377 (2015), pp. 47–56.
- [211] Z. JIN AND Y. LIU, *Dna methylation in human diseases*, *Genes & diseases*, 5 (2018), pp. 1–8.
- [212] P. A. JONES, *Functions of dna methylation: islands, start sites, gene bodies and beyond*, *Nature Reviews Genetics*, 13 (2012), pp. 484–492.
- [213] T. JOSHI, Y. CHEN, J. M. BECKER, N. ALEXANDROV, AND D. XU, *Genome-scale gene function prediction using multiple sources of high-throughput data in yeast *saccharomyces cerevisiae**, *Omics: a journal of integrative biology*, 8 (2004), pp. 322–333.

- [214] T. B. KALLEHAUGE, S. KOL, M. RØRDAM ANDERSEN, C. KROUN DAMGAARD, G. M. LEE, AND H. FAUSTRUP KILDEGAARD, *Endoplasmic reticulum-directed recombinant mrna displays subcellular localization equal to endogenous mrna during transient expression in cho cells*, *Biotechnology journal*, 11 (2016), pp. 1362–1367.
- [215] P. KAPRANOV, J. CHENG, S. DIKE, D. A. NIX, R. DUTTAGUPTA, A. T. WILLINGHAM, P. F. STADLER, J. HERTEL, J. HACKERMÜLLER, I. L. HOFACKER, ET AL., *Rna maps reveal new rna classes and a possible function for pervasive transcription*, *Science*, 316 (2007), pp. 1484–1488.
- [216] O. C. KARABULUT, B. A. KARPUZCU, E. TÜRK, A. H. IBRAHIM, AND B. E. SÜZEK, *ML-advinfect: a machine-learning based adenoviral infection predictor*, *Frontiers in Molecular Biosciences*, 8 (2021).
- [217] D. R. KELLEY, *Cross-species regulatory sequence activity prediction*, *PLoS computational biology*, 16 (2020), p. e1008050.
- [218] J. KHANAL, I. NAZARI, H. TAYARA, AND K. T. CHONG, *4mccnn: Identification of n4-methylcytosine sites in prokaryotes using convolutional neural network*, *IEEE Access*, 7 (2019), pp. 145455–145461.
- [219] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, *arXiv preprint arXiv:1412.6980*, (2014).
- [220] T. N. KIPF AND M. WELLING, *Variational graph auto-encoders*, *arXiv preprint arXiv:1611.07308*, (2016).
- [221] R. KOHAVI AND G. H. JOHN, *Wrappers for feature subset selection*, *Artificial intelligence*, 97 (1997), pp. 273–324.
- [222] L. KONG AND L. ZHANG, *i6ma-dncp: computational identification of dna n6-methyladenine sites in the rice genome using optimized dinucleotide-based features*, *Genes*, 10 (2019), p. 828.
- [223] M. KONG, Y. ZHANG, D. XU, W. CHEN, AND M. DEHMER, *Fctp-wsrc: protein–protein interactions prediction via weighted sparse representation based classification*, *Frontiers in genetics*, 11 (2020), p. 18.
- [224] W. KOPP, R. MONTI, A. TAMBURRINI, U. OHLER, AND A. AKALIN, *Deep learning for genomics using janggu*, *Nature communications*, 11 (2020), pp. 1–7.
- [225] T. KOUZARIDES, *Chromatin modifications and their function*, *Cell*, 128 (2007), pp. 693–705.

BIBLIOGRAPHY

- [226] I. A. KOVÁCS, K. LUCK, K. SPIROHN, Y. WANG, C. POLLIS, S. SCHLABACH, W. BIAN, D.-K. KIM, N. KISHORE, T. HAO, ET AL., *Network-based prediction of protein interactions*, Nature communications, 10 (2019), pp. 1–8.
- [227] L. KOZMA, *k nearest neighbors algorithm (knn)*, Helsinki University of Technology, 32 (2008).
- [228] R. KRISTELEIT, L. STIMSON, P. WORKMAN, AND W. AHERNE, *Histone modification enzymes: novel targets for cancer drugs*, Expert opinion on emerging drugs, 9 (2004), pp. 135–154.
- [229] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.
- [230] N. J. KROGAN, G. CAGNEY, H. YU, G. ZHONG, X. GUO, A. IGNATCHENKO, J. LI, S. PU, N. DATTA, A. P. TIKUISIS, ET AL., *Global landscape of protein complexes in the yeast saccharomyces cerevisiae*, Nature, 440 (2006), pp. 637–643.
- [231] N. KULMINSKAYA AND M. OBERER, *Protein-protein interactions regulate the activity of adipose triglyceride lipase in intracellular lipolysis*, Biochimie, 169 (2020), pp. 62–68.
- [232] J. T. KUNG, D. COLOGNORI, AND J. T. LEE, *Long noncoding rnas: past, present, and future*, Genetics, 193 (2013), pp. 651–669.
- [233] C. KUO-CHEN, *Prediction of protein cellular attributes using pseudo-amino acid composition*, Proteins: Structure, Function, and Bioinformatics, 43 (2001), pp. 246–255.
- [234] M.-H. T. LAI, *Common applications of next-generation sequencing technologies in genomic research*, (2013).
- [235] W. K. LAI AND B. F. PUGH, *Understanding nucleosome dynamics and their links to gene expression and dna replication*, Nature reviews Molecular cell biology, 18 (2017), pp. 548–562.
- [236] T. N. LAL, O. CHAPELLE, J. WESTON, AND A. ELISSEEFF, *Embedded methods*, in Feature extraction, Springer, 2006, pp. 137–165.
- [237] D. LALOVIĆ AND V. VELJKOVIĆ, *The global average dna base composition of coding regions may be determined by the electron-ion interaction potential*, Biosystems, 23 (1990), pp. 311–316.
- [238] J. LANCHANTIN, T. WEINGARTEN, A. SEKHON, C. MILLER, AND Y. QI, *Transfer learning for predicting virus-host protein interactions for novel virus sequences*, in Proceedings

- of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, 2021, pp. 1–10.
- [239] E. S. LANDER, L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY, J. BALDWIN, K. DEVON, K. DEWAR, M. DOYLE, W. FITZHUGH, ET AL., *Initial sequencing and analysis of the human genome*, (2001).
- [240] S. LAROCHELLE, *Dictating histone occupancy*, *Nature Structural & Molecular Biology*, 20 (2013), pp. 1145–1145.
- [241] E. LASDA AND R. PARKER, *Circular rnas: diversity of form and function*, *Rna*, 20 (2014), pp. 1829–1842.
- [242] N. Q. K. LE AND Q.-T. HO, *Deep transformers and convolutional neural network in identifying dna n6-methyladenine sites in cross-species genomes*, *Methods*, (2021).
- [243] 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*, 571 (2019), pp. 53–61.
- [244] Y. LECUN, L. BOTTOU, Y. BENGIO, P. HAFFNER, ET AL., *Gradient-based learning applied to document recognition*, *Proceedings of the IEEE*, 86 (1998), pp. 2278–2324.
- [245] D. LEE, R. KARCHIN, AND M. A. BEER, *Discriminative prediction of mammalian enhancers from dna sequence*, *Genome research*, 21 (2011), pp. 2167–2180.
- [246] E. C. S. LEE, S. A. M. ELHASSAN, G. P. L. LIM, W. H. KOK, S. W. TAN, E. N. LEONG, S. H. TAN, E. W. L. CHAN, S. K. BHATTAMISRA, R. RAJENDRAN, ET AL., *The roles of circular rnas in human development and diseases*, *Biomedicine & Pharmacotherapy*, 111 (2019), pp. 198–208.
- [247] T.-W. LEE, *Independent component analysis*, in *Independent component analysis*, Springer, 1998, pp. 27–66.
- [248] T.-Y. LEE, S.-A. CHEN, H.-Y. HUNG, AND Y.-Y. OU, *Incorporating distant sequence features and radial basis function networks to identify ubiquitin conjugation sites*, *PLoS one*, 6 (2011), p. e17331.
- [249] N. LEPAN ET AL., *Visualizing the history of pandemics*, *Visual Capitalist*, 14 (2020).
- [250] F. LI, J. CHEN, Z. GE, Y. WEN, Y. YUE, M. HAYASHIDA, A. BAGGAG, H. BENSMAIL, AND J. SONG, *Computational prediction and interpretation of both general and specific types of promoters in escherichia coli by exploiting a stacked ensemble-learning framework*, *Briefings in bioinformatics*, 22 (2021), pp. 2126–2140.

- [251] J. LI, J. ZHANG, L. ZUO, AND D. CHANG, *Reveal the cognitive process of deep learning during identifying nucleosome occupancy and histone modification*, in 2018 Chinese Automation Congress (CAC), IEEE, 2018, pp. 1856–1860.
- [252] J. LI, L. ZHANG, S. HE, F. GUO, AND Q. ZOU, *Sublocep: a novel ensemble predictor of subcellular localization of eukaryotic mrna based on machine learning*, *Briefings in Bioinformatics*, (2021).
- [253] P. LI, S. CHEN, H. CHEN, X. MO, T. LI, Y. SHAO, B. XIAO, AND J. GUO, *Using circular rna as a novel type of biomarker in the screening of gastric cancer*, *Clinica chimica acta*, 444 (2015), pp. 132–136.
- [254] P. LI, T. J. HASTIE, AND K. W. CHURCH, *Very sparse random projections*, in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 287–296.
- [255] X. LI, Z. ZHANG, X. LUO, J. SCHRIER, A. D. YANG, AND T. P. WU, *The exploration of n6-deoxyadenosine methylation in mammalian genomes*, *Protein & cell*, 12 (2021), pp. 756–768.
- [256] Y. LI, C. HUANG, L. DING, Z. LI, Y. PAN, AND X. GAO, *Deep learning in bioinformatics: Introduction, application, and perspective in the big data era*, *Methods*, 166 (2019), pp. 4–21.
- [257] Y. LI, Z. ZHAO, AND Z. TENG, *i4mc-el: Identifying dna n4-methylcytosine sites in the mouse genome using ensemble learning*, *BioMed Research International*, 2021 (2021).
- [258] Z. LI, C. HUANG, C. BAO, L. CHEN, M. LIN, X. WANG, G. ZHONG, B. YU, W. HU, L. DAI, ET AL., *Exon-intron circular rnas regulate transcription in the nucleus*, *Nature structural & molecular biology*, 22 (2015), pp. 256–264.
- [259] P. LIASHCHYNSKYI AND P. LIASHCHYNSKYI, *Grid search, random search, genetic algorithm: A big comparison for nas*, arXiv preprint arXiv:1912.06059, (2019).
- [260] L. LICATA, L. BRIGANTI, D. PELUSO, L. PERFETTO, M. IANNUCELLI, E. GALEOTA, F. SACCO, A. PALMA, A. P. NARDOZZA, E. SANTONICO, ET AL., *Mint, the molecular interaction database: 2012 update*, *Nucleic acids research*, 40 (2012), pp. D857–D861.
- [261] J. C. LIN, S. JEONG, G. LIANG, D. TAKAI, M. FATEMI, Y. C. TSAI, G. EGGER, E. N. GAL-YAM, AND P. A. JONES, *Role of nucleosomal occupancy in the epigenetic silencing of the mlh1 cpG island*, *Cancer cell*, 12 (2007), pp. 432–444.
- [262] K. LIN, A. C. MAY, AND W. R. TAYLOR, *Amino acid encoding schemes from protein structure alignments: Multi-dimensional vectors to describe residue types*, *Journal of theoretical biology*, 216 (2002), pp. 361–365.

- [263] Y. LIN, X. PAN, AND H.-B. SHEN, *Inclocator 2.0: a cell-line-specific subcellular localization predictor for long non-coding rnas with interpretable deep learning*, *Bioinformatics*, (2021).
- [264] Z. LIN AND X.-M. PAN, *Accurate prediction of protein secondary structural content*, *Journal of Protein Chemistry*, 20 (2001), pp. 217–220.
- [265] B. LIU, *ienhancer-psedeknc: Identification of enhancers and @articlebgroups based on pseudo degenerate kmer nucleotide composition*, *Neurocomputing*, 217 (2016), pp. 46–52.
- [266] B. LIU, L. FANG, R. LONG, X. LAN, AND K.-C. CHOU, *ienhancer-2l: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition*, 2016.
- [267] B. LIU, X. GAO, AND H. ZHANG, *Bioseq-analysis2. 0: an updated platform for analyzing dna, rna and protein sequences at sequence level and residue level based on machine learning approaches*, *Nucleic acids research*, 47 (2019), pp. e127–e127.
- [268] B. LIU, K. LI, D.-S. HUANG, AND K.-C. CHOU, *ienhancer-el: identifying enhancers and their strength with ensemble learning approach*, *Bioinformatics*, 34 (2018), pp. 3835–3842.
- [269] B. LIU, F. LIU, L. FANG, X. WANG, AND K.-C. CHOU, *repdna: a python package to generate various modes of feature vectors for dna sequences by incorporating user-defined physicochemical properties and sequence-order effects*, *Bioinformatics*, 31 (2015), pp. 1307–1309.
- [270] B. LIU, F. LIU, X. WANG, J. CHEN, L. FANG, AND K.-C. CHOU, *Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences*, *Nucleic acids research*, 43 (2015), pp. W65–W71.
- [271] B. LIU, Y. LIU, X. JIN, X. WANG, AND B. LIU, *irspot-dacc: a computational predictor for recombination hot/cold spots identification based on dinucleotide-based auto-cross covariance*, *Scientific reports*, 6 (2016), pp. 1–9.
- [272] B. LIU, J. XU, X. LAN, R. XU, J. ZHOU, X. WANG, AND K.-C. CHOU, *idna-prot | dis: identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition*, *PloS one*, 9 (2014), p. e106691.
- [273] Q. LIU, J. CHEN, Y. WANG, S. LI, C. JIA, J. SONG, AND F. LI, *Deeptorrent: a deep learning-based approach for predicting dna n4-methylcytosine sites*, *Briefings in bioinformatics*, 22 (2021), p. bbaa124.

- [274] Z. LIU, W. DONG, W. JIANG, AND Z. HE, *csdma: an improved bioinformatics tool for identifying dna 6 ma modifications via chou's 5-step rule*, Scientific reports, 9 (2019), pp. 1–9.
- [275] Z.-Y. LIU, J.-F. XING, W. CHEN, M.-W. LUAN, R. XIE, J. HUANG, S.-Q. XIE, AND C.-L. XIAO, *Mdr: an integrative dna n6-methyladenine and n4-methylcytosine modification database for rosaceae*, Horticulture research, 6 (2019).
- [276] W. LIU-WEI, Ş. KAFKAS, J. CHEN, N. J. DIMONACO, J. TEGNÉR, AND R. HOEHNDORF, *Deepviral: prediction of novel virus–host interactions from protein sequences and infectious disease phenotypes*, Bioinformatics, 37 (2021), pp. 2722–2729.
- [277] V. R. LIYANAGE, J. S. JARMASZ, N. MURUGESHAN, M. R. DEL BIGIO, M. RASTEGAR, AND J. R. DAVIE, *Dna modifications: function and applications in normal and disease states*, Biology, 3 (2014), pp. 670–723.
- [278] M. LLANO AND M. A. PEÑA-HERNANDEZ, *Defining pharmacological targets by analysis of virus–host protein interactions*, Advances in protein chemistry and structural biology, 111 (2018), pp. 223–242.
- [279] J. R. LLOYD, *Gefcom2012 hierarchical load forecasting: Gradient boosting machines and gaussian processes*, International Journal of Forecasting, 30 (2014), pp. 369–374.
- [280] J. LOUTEN, *Virus replication*, Essential Human Virology, (2016), p. 49.
- [281] K. LUGER, A. W. MÄDER, R. K. RICHMOND, D. F. SARGENT, AND T. J. RICHMOND, *Crystal structure of the nucleosome core particle at 2.8 Å resolution*, Nature, 389 (1997), pp. 251–260.
- [282] H. LV, F.-Y. DAO, D. ZHANG, Z.-X. GUAN, H. YANG, W. SU, M.-L. LIU, H. DING, W. CHEN, AND H. LIN, *idna-ms: an integrated computational tool for detecting dna modification sites in multiple genomes*, Iscience, 23 (2020), p. 100991.
- [283] Z. LV, D. WANG, H. DING, B. ZHONG, AND L. XU, *Escherichia coli dna n-4-methylcytosine site prediction accuracy improved by light gradient boosting machine feature selection technology*, Ieee Access, 8 (2020), pp. 14851–14859.
- [284] A. LYDIA AND S. FRANCIS, *A survey of optimization techniques for deep learning networks*, (2019), pp. 2454–9150.
- [285] Y. MA, X. ZHANG, Y.-Z. WANG, H. TIAN, AND S. XU, *Research progress of circular rnas in lung cancer*, Cancer biology & therapy, 20 (2019), pp. 123–129.

- [286] B. MANAVALAN, S. BASITH, T. H. SHIN, AND G. LEE, *Computational prediction of species-specific yeast dna replication origin via iterative feature representation*, Briefings in bioinformatics, 22 (2021), p. bbaa304.
- [287] B. MANAVALAN, S. BASITH, T. H. SHIN, L. WEI, AND G. LEE, *Meta-4mcpred: a sequence-based meta-predictor for accurate dna 4mc site prediction using effective feature representation*, Molecular Therapy-Nucleic Acids, 16 (2019), pp. 733–744.
- [288] M. MANDAL, A. MUKHOPADHYAY, AND U. MAULIK, *Prediction of protein subcellular localization by incorporating multiobjective pso-based feature subset selection into the general form of chou’s pseaac*, Medical & biological engineering & computing, 53 (2015), pp. 331–344.
- [289] K. D. MANDL AND A. K. MANRAI, *Potential excessive testing at scale: biomarkers, genomics, and machine learning*, Jama, 321 (2019), pp. 739–740.
- [290] S. MARTIN, D. ROE, AND J.-L. FAULON, *Predicting protein–protein interactions using signature products*, Bioinformatics, 21 (2005), pp. 218–226.
- [291] M. MATSUI AND D. R. COREY, *Non-coding rnas as drug targets*, Nature reviews Drug discovery, 16 (2017), p. 167.
- [292] J. S. MATTICK AND I. V. MAKUNIN, *Non-coding rna*, Human molecular genetics, 15 (2006), pp. R17–R29.
- [293] P. K. MEHER, A. RAI, AND A. R. RAO, *mloc-mrna: predicting multiple sub-cellular localization of mrnas using random forest algorithm coupled with feature selection via elastic net*, BMC bioinformatics, 22 (2021), pp. 1–24.
- [294] P. K. MEHER, S. SATPATHY, AND A. R. RAO, *mirnaloc: predicting mirna subcellular localizations based on principal component scores of physico-chemical properties and pseudo compositions of di-nucleotides*, Scientific reports, 10 (2020), pp. 1–12.
- [295] F. MEHMOOD, M. U. GHANI, M. N. ASIM, R. SHAHZADI, A. MEHMOOD, AND W. MAHMOOD, *Mpf-net: A computational multi-regional solar power forecasting framework*, Renewable and Sustainable Energy Reviews, 151 (2021), p. 111559.
- [296] S. MEMCZAK, M. JENS, A. ELEFSINIOTI, F. TORTI, J. KRUEGER, A. RYBAK, L. MAIER, S. D. MACKOWIAK, L. H. GREGERSEN, M. MUNSCHAUER, ET AL., *Circular rnas are a large class of animal rnas with regulatory potency*, Nature, 495 (2013), p. 333.
- [297] S. MERITY, N. S. KESKAR, AND R. SOCHER, *Regularizing and optimizing lstm language models*, arXiv preprint arXiv:1708.02182, (2017).

BIBLIOGRAPHY

- [298] R. MERRIS, *Laplacian matrices of graphs: a survey*, Linear algebra and its applications, 197 (1994), pp. 143–176.
- [299] J. G. MEYER, *Deep learning neural network tools for proteomics*, Cell Reports Methods, (2021), p. 100003.
- [300] T. MIKOLOV, K. CHEN, G. CORRADO, AND J. DEAN, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781, (2013).
- [301] L. D. MOORE, T. LE, AND G. FAN, *Dna methylation and its basic function*, Neuropsychopharmacology, 38 (2013), pp. 23–38.
- [302] K. V. MORRIS AND J. S. MATTICK, *The rise of regulatory rna*, Nature Reviews Genetics, 15 (2014), pp. 423–437.
- [303] R. MUHAMMAD, S. AHMED, D. MD FARID, S. SHATABDA, A. SHARMA, AND A. DEHZANGI, *Pyfeat: a python-based effective feature generation tool for dna, rna and protein sequences*, Bioinformatics, 35 (2019), pp. 3831–3833.
- [304] S. MUNIER, T. ROLLAND, C. DIOT, Y. JACOB, AND N. NAFFAKH, *Exploration of binary virus–host interactions using an infectious protein complementation assay*, Molecular & Cellular Proteomics, 12 (2013), pp. 2845–2855.
- [305] J. MYOUNG, *Two years of covid-19 pandemic: where are we now?*, 2022.
- [306] A. NABEEL, M. A. IBRAHIM, M. IMRAN MALIK, A. DENGEL, AND S. AHMED, *Circlocnet: A computational framework for circular rna sub-cellular localization prediction*, International Journal of Molecular Sciences, 23 (2022), p. 8221.
- [307] A. S. NAIR AND S. P. SREENADHAN, *A coding measure scheme employing electron-ion interaction pseudopotential (eiip)*, Bioinformation, 1 (2006), p. 197.
- [308] L. NANNI, *Fusion of classifiers for predicting protein–protein interactions*, Neurocomputing, 68 (2005), pp. 289–296.
- [309] L. NANNI AND A. LUMINI, *An ensemble of k -local hyperplanes for predicting protein–protein interactions*, Bioinformatics, 22 (2006), pp. 1207–1210.
- [310] G. J. NARLIKAR, H.-Y. FAN, AND R. E. KINGSTON, *Cooperation between complexes that regulate chromatin structure and transcription*, Cell, 108 (2002), pp. 475–487.
- [311] R. N. NAZAR, *The ribosomal 5.8 s rna: eukaryotic adaptation or processing variant?*, Canadian journal of biochemistry and cell biology, 62 (1984), pp. 311–320.

- [312] N. G. NGUYEN, V. A. TRAN, D. L. NGO, D. PHAN, F. R. LUMBANRAJA, M. R. FAISAL, B. ABAPIHI, M. KUBO, K. SATOU, ET AL., *Dna sequence classification by convolutional neural network*, Journal of Biomedical Science and Engineering, 9 (2016), p. 280.
- [313] W. S. NOBLE, S. KUEHN, R. THURMAN, M. YU, AND J. STAMATOYANNOPOULOS, *Predicting the in vivo signature of human gene regulatory sequences*, Bioinformatics, 21 (2005), pp. i338–i343.
- [314] I. M. NOOREN AND J. M. THORNTON, *Structural characterisation and functional significance of transient protein–protein interactions*, Journal of molecular biology, 325 (2003), pp. 991–1018.
- [315] T. C. NORTHEY, A. BAREŠIĆ, AND A. C. MARTIN, *Intpred: a structure-based predictor of protein–protein interaction sites*, Bioinformatics, 34 (2018), pp. 223–229.
- [316] A. NOWOSAD, Ü. TURANLI, AND K. LORENC, *The coronavirus sars-cov-2 and its impact on the world*, (2022).
- [317] I. NUSRAT AND S.-B. JANG, *A comparison of regularization techniques in deep neural networks*, Symmetry, 10 (2018), p. 648.
- [318] Y. L. ORLOV AND A. A. ANASHKINA, *Life: Computational genomics applications in life sciences*, 2021.
- [319] S. OROZCO-ARIAS, M. S. CANDAMIL-CORTÉS, P. A. JAIMES, J. S. PIÑA, R. TABARES-SOTO, R. GUYOT, AND G. ISAZA, *K-mer-based machine learning method to classify ltr-retrotransposons in plant genomes*, PeerJ, 9 (2021), p. e11456.
- [320] D. W. OTTER, J. R. MEDINA, AND J. K. KALITA, *A survey of the usages of deep learning for natural language processing*, IEEE Transactions on Neural Networks and Learning Systems, (2020).
- [321] Z. K. O'BROWN, K. BOULIAS, J. WANG, S. Y. WANG, N. M. O'BROWN, Z. HAO, H. SHIBUYA, P.-E. FADY, Y. SHI, C. HE, ET AL., *Sources of artifact in measurements of 6ma and 4mc abundance in eukaryotic genomic dna*, BMC genomics, 20 (2019), pp. 1–15.
- [322] H. O'GEEN, L. ECHIPARE, AND P. J. FARNHAM, *Using chip-seq technology to generate high-resolution profiles of histone modifications*, in Epigenetics Protocols, Springer, 2011, pp. 265–286.
- [323] A. PADRÒN, S. IWASAKI, AND N. T. INGOLIA, *Proximity rna labeling by apex-seq reveals the organization of translation initiation complexes and repressive rna granules*, Molecular cell, 75 (2019), pp. 875–887.

- [324] X. PAN AND K. XIONG, *Predcircrna: computational classification of circular rna from other long non-coding rna using hybrid features*, *Molecular Biosystems*, 11 (2015), pp. 2219–2226.
- [325] V. PAVET, M. PORTAL, J. MOULIN, R. HERBRECHT, AND H. GRONEMEYER, *Towards novel paradigms for cancer therapy*, *Oncogene*, 30 (2011), pp. 1–20.
- [326] T. PAWSON AND P. NASH, *Protein–protein interactions define specificity in signal transduction*, *Genes & development*, 14 (2000), pp. 1027–1047.
- [327] S. PERI, J. D. NAVARRO, T. Z. KRISTIANSEN, R. AMANCHY, V. SURENDRANATH, B. MUTHUSAMY, T. GANDHI, K. CHANDRIKA, N. DESHPANDE, S. SURESH, ET AL., *Human protein reference database as a discovery resource for proteomics*, *Nucleic acids research*, 32 (2004), pp. D497–D501.
- [328] B. PEROZZI, R. AL-RFOU, AND S. SKIENA, *Deepwalk: Online learning of social representations*, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.
- [329] L. PERRIN-COCON, O. DIAZ, C. JACQUEMIN, V. BARTHEL, E. OGIRE, C. RAMIÈRE, P. ANDRÉ, V. LOTTEAU, AND P.-O. VIDALAIN, *The current landscape of coronavirus-host protein–protein interactions*, *Journal of translational medicine*, 18 (2020), pp. 1–15.
- [330] I. PETTA, S. LIEVENS, C. LIBERT, J. TAVERNIER, AND K. DE BOSSCHER, *Modulation of protein–protein interactions for the development of novel therapeutics*, *Molecular Therapy*, 24 (2016), pp. 707–718.
- [331] T. H. PHAM, T. B. HO, D. H. TRAN, AND K. SATOU, *Prediction of histone modifications in dna sequences*, in *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, IEEE, 2007, pp. 959–966.
- [332] T. H. PHAML, D. H. TRAN, T. B. HO, K. SATOU, AND G. VALIENTE, *Qualitatively predicting acetylation and methylation areas in dna sequences*, *Genome Informatics*, 16 (2005), pp. 3–11.
- [333] C. PIAN, G. ZHANG, F. LI, AND X. FAN, *Mm-6mapred: identifying dna n6-methyladenine sites based on markov model*, *Bioinformatics*, 36 (2020), pp. 388–392.
- [334] Y. QI, J. KLEIN-SEETHARAMAN, AND Z. BAR-JOSEPH, *Random forest similarity for protein-protein interaction prediction from multiple sources*, in *Biocomputing 2005*, World Scientific, 2005, pp. 531–542.
- [335] C. R. RAHMAN, R. AMIN, S. SHATABDA, M. TOAHA, AND S. ISLAM, *A convolution based computational approach towards dna n6-methyladenine site identification and motif extraction in rice genome*, *Scientific Reports*, 11 (2021), pp. 1–13.

- [336] J. RAMOS ET AL., *Using tf-idf to determine word relevance in document queries*, in Proceedings of the first instructional conference on machine learning, vol. 242, Citeseer, 2003, pp. 29–48.
- [337] S. RAMPERSAD AND P. TENNANT, *Replication and expression strategies of viruses*, Viruses, (2018), p. 55.
- [338] G. RANJAN, A. K. VERMA, AND S. RADHIKA, *K-nearest neighbors and grid search cv based real time fault monitoring system for industries*, in 2019 IEEE 5th international conference for convergence in technology (I2CT), IEEE, 2019, pp. 1–5.
- [339] M. I. RAZZAK, M. IMRAN, AND G. XU, *Big data analytics for preventive medicine*, Neural Computing and Applications, 32 (2020), pp. 4417–4451.
- [340] J. READ, B. PFAHRINGER, G. HOLMES, AND E. FRANK, *Classifier chains for multi-label classification*, Machine Learning, 85 (2011), p. 333.
- [341] A. REHMAN, S. NAZ, AND I. RAZZAK, *Leveraging big data analytics in healthcare enhancement: Trends, challenges and opportunities*, arXiv preprint arXiv:2004.09010, (2020).
- [342] M. U. REHMAN, H. TAYARA, AND K. T. CHONG, *Denn-4mc: Densely connected neural network based n4-methylcytosine site prediction in multiple species*, Computational and structural biotechnology journal, 19 (2021), pp. 6009–6019.
- [343] J. RODRIGUEZ, L. LEE, B. LYNCH, AND T. TSUKIYAMA, *Nucleosome occupancy as a novel chromatin parameter for replication origin functions*, Genome research, 27 (2017), pp. 269–277.
- [344] S. ROSSET, *Robust boosting and its relation to bagging*, in Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 249–255.
- [345] E. ROSSI, F. MONTI, M. BRONSTEIN, AND P. LIÒ, *ncrna classification with graph convolutional networks*, arXiv preprint arXiv:1905.06515, (2019).
- [346] R. J. RUMMEL, *Applied factor analysis*, Northwestern University Press, 1988.
- [347] M. RUSNATI, P. CHIODELLI, A. BUGATTI, AND C. URBINATI, *Bridging the past and the future of virology: surface plasmon resonance as a powerful tool to investigate virus/host interactions*, Critical reviews in microbiology, 41 (2015), pp. 238–260.
- [348] L. SALWINSKI, C. S. MILLER, A. J. SMITH, F. K. PETTIT, J. U. BOWIE, AND D. EISENBERG, *The database of interacting proteins: 2004 update*, Nucleic acids research, 32 (2004), pp. D449–D451.

- [349] L. N. SANCHEZ-PINTO, L. R. VENABLE, J. FAHRENBACH, AND M. M. CHURPEK, *Comparison of variable selection methods for clinical predictive modeling*, International journal of medical informatics, 116 (2018), pp. 10–17.
- [350] V. SARAVANAN AND N. GAUTHAM, *Harnessing computational biology for exact linear b-cell epitope prediction: a novel amino acid composition-based feature descriptor*, Omics: a journal of integrative biology, 19 (2015), pp. 648–658.
- [351] K. SASIREKHA AND P. BABY, *Agglomerative hierarchical clustering algorithm-a*, International Journal of Scientific and Research Publications, 83 (2013), p. 83.
- [352] L. K. SAUL AND S. T. ROWEIS, *An introduction to locally linear embedding*, unpublished. Available at: <http://www.cs.toronto.edu/~roweis/lle/publications.html>, (2000).
- [353] A. F. SAVULESCU, E. BOUILHOL, N. BEAUME, AND M. NIKOLSKI, *Prediction of rna subcellular localization: learning from heterogeneous data sources*, Iscience, (2021), p. 103298.
- [354] A. F. SAVULESCU, R. BRACKIN, E. BOUILHOL, B. DARTIGUES, J. H. WARRELL, M. R. PIMENTEL, N. BEAUME, I. C. FORTUNATO, S. DALLONGEVILLE, M. BOULLE, ET AL., *Interrogating rna and protein spatial subcellular distribution in smfish data with dyfish*, Cell Reports Methods, 1 (2021), p. 100068.
- [355] R. E. SCHAPIRE AND Y. SINGER, *Boostexter: A boosting-based system for text categorization*, Machine learning, 39 (2000), pp. 135–168.
- [356] G. SCHNEIDER AND P. WREDE, *The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site*, Biophysical Journal, 66 (1994), pp. 335–344.
- [357] A. SCHOENROCK, B. SAMANFAR, S. PITRE, M. HOOSHYAR, K. JIN, C. A. PHILLIPS, H. WANG, S. PHANSE, K. OMIDI, Y. GUI, ET AL., *Efficient prediction of human protein-protein interactions at a global scale*, BMC bioinformatics, 15 (2014), p. 383.
- [358] B. SCHÖLKOPF, A. SMOLA, AND K.-R. MÜLLER, *Kernel principal component analysis*, in International conference on artificial neural networks, Springer, 1997, pp. 583–588.
- [359] S. SHAH AND A. ROSS, *Generating synthetic irises by feature agglomeration*, in 2006 international conference on image processing, IEEE, 2006, pp. 317–320.
- [360] K. SHAHBABIAN AND P. CHARTRAND, *Control of cytoplasmic mrna localization*, Cellular and Molecular Life Sciences, 69 (2012), pp. 535–552.

- [361] B. SHEKAR AND G. DAGNEW, *Grid search-based hyperparameter tuning and classification of microarray cancer data*, in 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), IEEE, 2019, pp. 1–8.
- [362] J. SHEN, J. ZHANG, X. LUO, W. ZHU, K. YU, K. CHEN, Y. LI, AND H. JIANG, *Predicting protein–protein interactions based only on sequences information*, Proceedings of the National Academy of Sciences, 104 (2007), pp. 4337–4341.
- [363] X. SHI, M. SUN, H. LIU, Y. YAO, AND Y. SONG, *Long non-coding rnas: a new frontier in the study of human diseases*, Cancer letters, 339 (2013), pp. 159–166.
- [364] S. SHOKRALLA, J. L. SPALL, J. F. GIBSON, AND M. HAJIBABAEI, *Next-generation sequencing technologies for environmental dna research*, Molecular ecology, 21 (2012), pp. 1794–1805.
- [365] K. SIMONYAN AND A. ZISSERMAN, *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv:1409.1556, (2014).
- [366] R. SINGH, D. PARK, J. XU, R. HOSUR, AND B. BERGER, *Struct2net: a web service to predict protein–protein interactions using a structure-based approach*, Nucleic acids research, 38 (2010), pp. W508–W515.
- [367] R. R. SOKAL AND B. A. THOMSON, *Population structure inferred by local spatial autocorrelation: an example from an amerindian tribal population*, American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists, 129 (2006), pp. 121–131.
- [368] M. S. SOROWER, *A literature survey on algorithms for multi-label learning*, Oregon State University, Corvallis, 18 (2010), pp. 1–25.
- [369] C. O. S. SORZANO, J. VARGAS, AND A. P. MONTANO, *A survey of dimensionality reduction techniques*, arXiv preprint arXiv:1403.2877, (2014).
- [370] E. SPYROMITROS, G. TSOUMAKAS, AND I. VLAHAVAS, *An empirical study of lazy multilabel classification algorithms*, in Hellenic conference on artificial intelligence, Springer, 2008, pp. 401–406.
- [371] D. STREIT, T. SHANMUGAM, A. GARBELYANSKI, S. SIMM, AND E. SCHLEIFF, *The existence and localization of nuclear snrnas in arabidopsis thaliana revisited*, Plants, 9 (2020), p. 1016.
- [372] M. STRICKER, M. N. ASIM, A. DENGEL, AND S. AHMED, *Circnet: An encoder–decoder-based convolution neural network (cnn) for circular rna identification*, Neural Computing and Applications, 34 (2022), pp. 11441–11452.

BIBLIOGRAPHY

- [373] Z.-D. SU, Y. HUANG, Z.-Y. ZHANG, Y.-W. ZHAO, D. WANG, W. CHEN, K.-C. CHOU, AND H. LIN, *iloc-lncrna: predict the subcellular location of lncrnas by incorporating octamer composition into general pseknrc*, *Bioinformatics*, 34 (2018), pp. 4196–4204.
- [374] T. C. SÜDHOF, *The synaptic vesicle cycle: a cascade of protein–protein interactions*, *Nature*, 375 (1995), pp. 645–653.
- [375] J.-N. SUN, H.-Y. YANG, J. YAO, H. DING, S.-G. HAN, C.-Y. WU, AND H. TANG, *Prediction of cyclin protein using two-step feature selection technique*, *IEEE Access*, 8 (2020), pp. 109535–109542.
- [376] J. L. SUSSMAN, D. LIN, J. JIANG, N. O. MANNING, J. PRILUSKY, O. RITTER, AND E. E. ABOLA, *Protein data bank (pdb): database of three-dimensional structural information of biological macromolecules*, *Acta Crystallographica Section D: Biological Crystallography*, 54 (1998), pp. 1078–1084.
- [377] M. TAHIR, M. HAYAT, I. ULLAH, AND K. T. CHONG, *A deep learning-based computational approach for discrimination of dna n6-methyladenosine sites by fusing heterogeneous features*, *Chemometrics and Intelligent Laboratory Systems*, 206 (2020), p. 104151.
- [378] M. TAHIR, H. TAYARA, M. HAYAT, AND K. T. CHONG, *Intelligent and robust computational prediction model for dna n4-methylcytosine sites via natural language processing*, *Chemometrics and Intelligent Laboratory Systems*, 217 (2021), p. 104391.
- [379] K. K. TAN, N. Q. K. LE, H.-Y. YEH, AND M. C. H. CHUA, *Ensemble of deep recurrent neural networks for identifying enhancers via dinucleotide physicochemical properties*, *Cells*, 8 (2019), p. 767.
- [380] P.-N. TAN, M. STEINBACH, AND V. KUMAR, *Introduction to data mining addison-wesley*, Reading, MA, USA, (2005).
- [381] J. TANG, C. DENG, AND G.-B. HUANG, *Extreme learning machine for multilayer perceptron*, *IEEE transactions on neural networks and learning systems*, 27 (2015), pp. 809–821.
- [382] J. TANG, M. QU, M. WANG, M. ZHANG, J. YAN, AND Q. MEI, *Line: Large-scale information network embedding*, in *Proceedings of the 24th international conference on world wide web*, 2015, pp. 1067–1077.
- [383] Q. TANG, F. NIE, J. KANG, AND W. CHEN, *mrnaloater: Enhance the prediction accuracy of eukaryotic mrna subcellular localization by using model fusion strategy*, *Molecular Therapy*, (2021).
- [384] A. TAREEN AND J. B. KINNEY, *Logomaker: beautiful sequence logos in python*, *Bioinformatics*, 36 (2020), pp. 2272–2274.

- [385] S. K. THAKER, J. CH'NG, AND H. R. CHRISTOFK, *Viral hijacking of cellular metabolism*, BMC biology, 17 (2019), pp. 1–15.
- [386] B. TIAN, X. WU, C. CHEN, W. QIU, Q. MA, AND B. YU, *Predicting protein–protein interactions by fusing various chou's pseudo components and using wavelet denoising approach*, Journal of Theoretical Biology, 462 (2019), pp. 329–346.
- [387] V. S. TIDAKE AND S. S. SANE, *Multi-label classification: a survey*, International Journal of Engineering and Technology, 7 (2018), pp. 1045–1054.
- [388] H. TILGNER, D. G. KNOWLES, R. JOHNSON, C. A. DAVIS, S. CHAKRABORTTY, S. DJEBALI, J. CURADO, M. SNYDER, T. R. GINGERAS, AND R. GUIGÓ, *Deep sequencing of subcellular rna fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncrnas*, Genome research, 22 (2012), pp. 1616–1625.
- [389] D. H. TRAN, T. H. PHAM, K. SATOU, AND T. B. HO, *Conditional random fields for predicting and analyzing histone occupancy, acetylation and methylation areas in dna sequences*, in Workshops on Applications of Evolutionary Computation, Springer, 2006, pp. 221–230.
- [390] G. TSOUMAKAS AND I. KATAKIS, *Multi-label classification: An overview*, International Journal of Data Warehousing and Mining (IJDWM), 3 (2007), pp. 1–13.
- [391] G. TSOUMAKAS AND I. VLAHAVAS, *Random k-labelsets: An ensemble method for multilabel classification*, in European conference on machine learning, Springer, 2007, pp. 406–417.
- [392] S. TSUKIYAMA, M. M. HASAN, H.-W. DENG, AND H. KURATA, *Bert6ma: prediction of dna n6-methyladenine site using deep learning-based approaches*, Briefings in Bioinformatics, (2022).
- [393] S. TSUKIYAMA, M. M. HASAN, S. FUJII, AND H. KURATA, *Lstm-phv: prediction of human-virus protein–protein interactions by lstm with word2vec*, Briefings in bioinformatics, 22 (2021), p. bbab228.
- [394] C.-W. TUNG AND S.-Y. HO, *Computational identification of ubiquitylation sites from protein sequences*, BMC bioinformatics, 9 (2008), pp. 1–15.
- [395] L. VAN DER MAATEN AND G. HINTON, *Visualizing data using t-sne.*, Journal of machine learning research, 9 (2008).
- [396] S. VASHISHTH, *Neural graph embedding methods for natural language processing*, arXiv preprint arXiv:1911.03042, (2019).
- [397] B. VENKATESH AND J. ANURADHA, *A review of feature selection and its methods*, Cybernetics and Information Technologies, 19 (2019), pp. 3–26.

BIBLIOGRAPHY

- [398] N. J. VICKERS, *Animal communication: when i'm calling you, will you answer too?*, *Current biology*, 27 (2017), pp. R713–R715.
- [399] L. P. VILLARREAL, *Are viruses alive?*, *Scientific American*, 291 (2004), pp. 100–105.
- [400] A. WAHAB, H. TAYARA, Z. XUAN, AND K. T. CHONG, *Dna sequences performs as natural language processing by exploiting deep learning algorithm for the identification of n4-methylcytosine*, *Scientific reports*, 11 (2021), pp. 1–9.
- [401] D. WANG, P. CUI, AND W. ZHU, *Structural deep network embedding*, in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 1225–1234.
- [402] D. WANG, Z. ZHANG, Y. JIANG, Z. MAO, D. WANG, H. LIN, AND D. XU, *Dm3loc: multi-label mrna subcellular localization prediction and analysis based on multi-head self-attention mechanism*, *Nucleic Acids Research*, 49 (2021), pp. e46–e46.
- [403] H. WANG, Y. DING, J. TANG, Q. ZOU, AND F. GUO, *Identify rna-associated subcellular localizations based on multi-label learning using chou's 5-steps rule*, *BMC genomics*, 22 (2021), pp. 1–14.
- [404] H.-T. WANG, F.-H. XIAO, G.-H. LI, AND Q.-P. KONG, *Identification of dna n6-methyladenine sites by integration of sequence features*, *Epigenetics & Chromatin*, 13 (2020), pp. 1–10.
- [405] J. WANG AND L. WANG, *Deep learning of the back-splicing code for circular rna formation*, *Bioinformatics*, 35 (2019), pp. 5235–5242.
- [406] J. T.-L. WANG, Q. MA, D. SHASHA, AND C. H. WU, *New techniques for extracting features from protein sequences*, *IBM Systems Journal*, 40 (2001), pp. 426–441.
- [407] L. WANG, H. J. PARK, S. DASARI, S. WANG, J.-P. KOCHER, AND W. LI, *Cpat: Coding-potential assessment tool using an alignment-free logistic regression model*, *Nucleic acids research*, 41 (2013), pp. e74–e74.
- [408] R.-S. WANG, Y. WANG, L.-Y. WU, X.-S. ZHANG, AND L. CHEN, *Analysis on multi-domain cooperation for predicting protein-protein interactions*, *BMC bioinformatics*, 8 (2007), p. 391.
- [409] W.-T. WANG, C. HAN, Y.-M. SUN, T.-Q. CHEN, AND Y.-Q. CHEN, *Noncoding rnas in cancer therapy resistance and targeted drug development*, *Journal of hematology & oncology*, 12 (2019), p. 55.

- [410] X. WANG, P. HU, AND L. HU, *A novel stochastic block model for network-based prediction of protein-protein interactions*, in International Conference on Intelligent Computing, Springer, 2020, pp. 621–632.
- [411] Y. WANG, H. YAO, AND S. ZHAO, *Auto-encoder based dimensionality reduction*, Neurocomputing, 184 (2016), pp. 232–242.
- [412] Z. WANG, X. LEI, AND F.-X. WU, *Identifying cancer-specific circrna-rbp binding sites based on deep learning*, Molecules, 24 (2019), p. 4035.
- [413] G. I. WEBB, E. KEOGH, AND R. MIIKKULAINEN, *Naïve bayes.*, Encyclopedia of machine learning, 15 (2010), pp. 713–714.
- [414] L. WEI, S. LUAN, L. A. E. NAGAI, R. SU, AND Q. ZOU, *Exploring sequence-based features for the improved prediction of dna n4-methylcytosine sites in multiple species*, Bioinformatics, 35 (2019), pp. 1326–1333.
- [415] L. WEI, R. SU, S. LUAN, Z. LIAO, B. MANAVALAN, Q. ZOU, AND X. SHI, *Iterative feature representations improve n4-methylcytosine site prediction*, Bioinformatics, 35 (2019), pp. 4930–4937.
- [416] L. WEI, C. ZHOU, H. CHEN, J. SONG, AND R. SU, *Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides*, Bioinformatics, 34 (2018), pp. 4007–4016.
- [417] B. WEN, W.-F. ZENG, Y. LIAO, Z. SHI, S. R. SAVAGE, W. JIANG, AND B. ZHANG, *Deep learning in proteomics*, Proteomics, 20 (2020), p. 1900335.
- [418] J. A. WEST, A. COOK, B. H. ALVER, M. STADTFELD, A. M. DEATON, K. HOCHEDLINGER, P. J. PARK, M. Y. TOLSTORUKOV, AND R. E. KINGSTON, *Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming*, Nature communications, 5 (2014), pp. 1–12.
- [419] J. S. WHISELL AND C. L. CLARKE, *Improving document clustering using okapi bm25 feature weighting*, Information retrieval, 14 (2011), pp. 466–487.
- [420] G. WHITE AND W. SEFFENS, *Using a neural network to backtranslate amino acid sequences*, Electronic Journal of Biotechnology, 1 (1998), pp. 17–18.
- [421] J. H. WILBERTZ, F. VOIGT, I. HORVATHOVA, G. ROTH, Y. ZHAN, AND J. A. CHAO, *Single-molecule imaging of mrna localization and regulation during the integrated stress response*, Molecular cell, 73 (2019), pp. 946–958.
- [422] R. E. WRIGHT, *Logistic regression.*, (1995).

- [423] X.-Z. WU AND Z.-H. ZHOU, *A unified view of multi-label performance measures*, in International Conference on Machine Learning, PMLR, 2017, pp. 3780–3788.
- [424] I. XENARIOS, L. SALWINSKI, X. J. DUAN, P. HIGNEY, S.-M. KIM, AND D. EISENBERG, *Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions*, Nucleic acids research, 30 (2002), pp. 303–305.
- [425] S. XIA, J. FENG, L. LEI, J. HU, L. XIA, J. WANG, Y. XIANG, L. LIU, S. ZHONG, L. HAN, ET AL., *Comprehensive characterization of tissue-specific circular rnas in the human and mouse genomes*, Briefings in bioinformatics, 18 (2017), pp. 984–992.
- [426] Y. XIAO, J. CAI, Y. YANG, H. ZHAO, AND H. SHEN, *Prediction of microrna subcellular localization by using a sequence-to-sequence model*, in 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018, pp. 1332–1337.
- [427] S. XING, N. WALLMEROOTH, K. W. BERENDZEN, AND C. GREFEN, *Techniques for the analysis of protein-protein interactions in vivo*, Plant physiology, 171 (2016), pp. 727–758.
- [428] H. XU, P. JIA, AND Z. ZHAO, *Deep4mc: systematic assessment and computational prediction for dna n4-methylcytosine sites by deep learning*, Briefings in bioinformatics, 22 (2021), p. bbaa099.
- [429] Z. YAN, E. LÉCUYER, AND M. BLANCHETTE, *Prediction of mrna subcellular localization using deep recurrent neural networks*, Bioinformatics, 35 (2019), pp. i333–i342.
- [430] A. YANG, W. ZHANG, J. WANG, K. YANG, Y. HAN, AND L. ZHANG, *Review on the application of machine learning algorithms in the sequence data mining of dna*, Frontiers in Bioengineering and Biotechnology, (2020), p. 1032.
- [431] J. YANG, K. LANG, G. ZHANG, X. FAN, Y. CHEN, AND C. PIAN, *Somm4mc: a second-order markov model for dna n4-methylcytosine site prediction in six species*, Bioinformatics, 36 (2020), pp. 4103–4105.
- [432] S. YANG, C. FU, X. LIAN, X. DONG, AND Z. ZHANG, *Understanding human-virus protein-protein interactions using a human protein complex-based analysis framework*, MSys-tems, 4 (2019), pp. e00303–18.
- [433] X. YANG, S. YANG, Q. LI, S. WUCHTY, AND Z. ZHANG, *Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method*, Computational and structural biotechnology journal, 18 (2020), pp. 153–161.
- [434] X. YANG, S. YANG, X. LIAN, S. WUCHTY, AND Z. ZHANG, *Transfer learning via multi-scale convolutional neural layers for human–virus protein–protein interaction prediction*, Bioinformatics, 37 (2021), pp. 4771–4778.

- [435] X. YANG, X. YE, X. LI, AND L. WEI, *idna-mt: Identification dna modification sites in multiple species by using multi-task learning based a neural network tool*, *Frontiers in genetics*, 12 (2021), p. 411.
- [436] X.-F. YANG, Y.-K. ZHOU, L. ZHANG, Y. GAO, AND P.-F. DU, *Predicting lncrna subcellular localization using unbalanced pseudo-k nucleotide compositions*, *Current Bioinformatics*, 15 (2020), pp. 554–562.
- [437] Y. YANG, X. FU, W. QU, Y. XIAO, AND H.-B. SHEN, *MirgoFs: a go-based functional similarity measurement for mirnas, with applications to the prediction of mirna subcellular localization and mirna–disease association*, *Bioinformatics*, 34 (2018), pp. 3547–3556.
- [438] D. YAO, L. ZHANG, M. ZHENG, X. SUN, Y. LU, AND P. LIU, *Circ2disease: a manually curated database of experimentally validated circrnas in human disease*, *Scientific reports*, 8 (2018), pp. 1–6.
- [439] Y. YAO, X. DU, Y. DIAO, AND H. ZHU, *An integration of deep learning with feature embedding for protein–protein interaction prediction*, *PeerJ*, 7 (2019), p. e7126.
- [440] C. YE AND B. P. TU, *Sink into the epigenome: histones as repositories that influence cellular metabolism*, *Trends in Endocrinology & Metabolism*, 29 (2018), pp. 626–637.
- [441] P. YE, Y. LUAN, K. CHEN, Y. LIU, C. XIAO, AND Z. XIE, *Methsmrt: an integrative database for dna n6-methyladenine and n4-methylcytosine generated by single-molecular real-time sequencing*, *Nucleic acids research*, (2016), p. gkw950.
- [442] H.-C. YI, Z.-H. YOU, D.-S. HUANG, AND C. K. KWOH, *Graph representation learning in bioinformatics: trends, methods and applications*, *Briefings in Bioinformatics*, (2021).
- [443] B. YIN, M. BALVERT, D. ZAMBRANO, A. SCHÖNHUTH, AND S. BOHTE, *An image representation based convolutional network for dna classification*, *arXiv preprint arXiv:1806.04931*, (2018).
- [444] Q. YIN, M. WU, Q. LIU, H. LV, AND R. JIANG, *DeepHistone: a deep learning approach to predicting histone modifications*, *BMC genomics*, 20 (2019), p. 193.
- [445] Z.-H. YOU, K. C. CHAN, AND P. HU, *Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest*, *PLoS one*, 10 (2015), p. e0125811.
- [446] Z.-H. YOU, K. C. CHAN, AND P. HU, *Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest*, *PLoS one*, 10 (2015), p. e0125811.

- [447] Z.-H. YOU, Y.-K. LEI, J. GUI, D.-S. HUANG, AND X. ZHOU, *Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data*, *Bioinformatics*, 26 (2010), pp. 2744–2751.
- [448] Z.-H. YOU, X. LI, AND K. C. CHAN, *An improved sequence-based prediction protocol for protein-protein interactions using amino acids substitution matrix and rotation forest ensemble classifiers*, *Neurocomputing*, 228 (2017), pp. 277–282.
- [449] Z.-H. YOU, L. ZHU, C.-H. ZHENG, H.-J. YU, S.-P. DENG, AND Z. JI, *Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set*, in *BMC bioinformatics*, vol. 15, Springer, 2014, pp. 1–9.
- [450] B. YU, C. CHEN, Z. YU, A. MA, B. LIU, AND Q. MA, *Prediction of protein-protein interactions based on elastic net and deep forest*, *bioRxiv*, (2020).
- [451] H. YU AND Z. DAI, *Snnrice6ma: a deep learning method for predicting dna n6-methyladenine sites in rice genome*, *Frontiers in genetics*, (2019), p. 1071.
- [452] N. YU, Z. YU, AND Y. PAN, *A deep learning method for lincrna detection using auto-encoder algorithm*, *BMC bioinformatics*, 18 (2017), p. 511.
- [453] A. ZAGHLOOL, A. AMEUR, C. WU, J. O. WESTHOLM, A. NIAZI, M. MANIVANNAN, K. BRAMLETT, M. NILSSON, AND L. FEUK, *Expression profiling and in situ screening of circular rnas in human tissues*, *Scientific reports*, 8 (2018), pp. 1–12.
- [454] A. ZAPPULO, D. VAN DEN BRUCK, C. C. MATTIOLI, V. FRANKE, K. IMAMI, E. MCSHANE, M. MORENO-ESTELLES, L. CALVIELLO, A. FILIPCHYK, E. PEGUERO-SANCHEZ, ET AL., *Rna localization is a key determinant of neurite-enriched proteome*, *Nature communications*, 8 (2017), pp. 1–13.
- [455] F. ZENG, G. FANG, AND L. YAO, *A deep neural network for identifying dna n4-methylcytosine sites*, *Frontiers in genetics*, 11 (2020), p. 209.
- [456] M. ZENG, Y. WU, C. LU, F. ZHANG, F.-X. WU, AND M. LI, *DeepInCloc: a deep learning framework for long non-coding rna subcellular localization prediction based on subsequence embedding*, *bioRxiv*, (2021).
- [457] R. ZENG, S. CHENG, AND M. LIAO, *4mcpred-mtl: Accurate identification of dna 4mc sites in multiple species using multi-task deep learning based on multi-head attention mechanism*, *Frontiers in Cell and Developmental Biology*, 9 (2021), p. 819.
- [458] R. ZENG AND M. LIAO, *6mapred-msff: A deep learning model for predicting dna n6-methyladenine sites across species based on a multi-scale feature fusion mechanism*, *Applied Sciences*, 11 (2021), p. 7731.

- [459] C.-T. ZHANG, R. ZHANG, AND H.-Y. OU, *The z curve database: a graphic representation of genome sequences*, *Bioinformatics*, 19 (2003), pp. 593–599.
- [460] L. ZHANG, X. XIAO, AND Z.-C. XU, *ipromoter-5mc: A novel fusion decision predictor for the identification of 5-methylcytosine sites in genome-wide dna promoters*, *Frontiers in Cell and Developmental Biology*, 8 (2020), p. 614.
- [461] M.-L. ZHANG AND Z.-H. ZHOU, *ML-knn: A lazy learning approach to multi-label learning*, *Pattern recognition*, 40 (2007), pp. 2038–2048.
- [462] Q. C. ZHANG, D. PETREY, L. DENG, L. QIANG, Y. SHI, C. A. THU, B. BISIKIRSKA, C. LEFEBVRE, D. ACCILI, T. HUNTER, ET AL., *Structure-based prediction of protein–protein interactions on a genome-wide scale*, *Nature*, 490 (2012), pp. 556–560.
- [463] S. ZHANG AND H. QIAO, *Kd-klnmf: Identification of lncrnas subcellular localization with multiple features and nonnegative matrix factorization*, *Analytical Biochemistry*, 610 (2020), p. 113995.
- [464] T. ZHANG, P. TAN, L. WANG, N. JIN, Y. LI, L. ZHANG, H. YANG, Z. HU, L. ZHANG, C. HU, ET AL., *Rnallocate: a resource for rna subcellular localizations*, *Nucleic acids research*, 45 (2017), pp. D135–D138.
- [465] W. ZHANG, B. CHENG, AND B. XU, *Application of next-generation sequencing technology in forensic science*, *Chinese Journal of Forensic Medicine*, (2017), pp. 40–43.
- [466] X. ZHANG, J. WANG, J. LI, W. CHEN, AND C. LIU, *Crlncrc: a machine learning-based method for cancer-related long noncoding rna identification using integrated features*, *BMC medical genomics*, 11 (2018), pp. 99–112.
- [467] Z. ZHANG, T. YANG, AND J. XIAO, *Circular rnas: promising biomarkers for human diseases*, *EBioMedicine*, 34 (2018), pp. 267–274.
- [468] Z.-Y. ZHANG, Y.-H. YANG, H. DING, D. WANG, W. CHEN, AND H. LIN, *Design powerful predictor for mrna subcellular location prediction in homo sapiens*, *Briefings in Bioinformatics*, 22 (2021), pp. 526–535.
- [469] J. ZHAO, *Phospholipase d and phosphatidic acid in plant defence response: from protein–protein and lipid–protein interactions to hormone signalling*, *Journal of Experimental Botany*, 66 (2015), pp. 1721–1736.
- [470] Z. ZHAO, X. ZHANG, F. CHEN, L. FANG, AND J. LI, *Accurate prediction of dna n4-methylcytosine sites via boost-learning various types of sequence features*, *BMC genomics*, 21 (2020), pp. 1–11.

- [471] Y. ZHENG, H. WANG, Y. ZHANG, X. GAO, E. P. XING, AND M. XU, *Poly (a)-dg: A deep-learning-based domain generalization method to identify cross-species poly (a) signal without prior knowledge from target species*, PLoS computational biology, 16 (2020), p. e1008297.
- [472] C. ZHOU, C. WANG, H. LIU, Q. ZHOU, Q. LIU, Y. GUO, T. PENG, J. SONG, J. ZHANG, L. CHEN, ET AL., *Identification and analysis of adenine n6-methylation sites in the rice genome*, Nature plants, 4 (2018), pp. 554–563.
- [473] X. ZHOU, B. PARK, D. CHOI, AND K. HAN, *A generalized approach to predicting protein-protein interactions between virus and host*, BMC genomics, 19 (2018), pp. 69–77.
- [474] Y. Z. ZHOU, Y. GAO, AND Y. Y. ZHENG, *Prediction of protein-protein interactions using local description of amino acid sequence*, in Advances in computer science and education applications, Springer, 2011, pp. 254–262.
- [475] Z.-H. ZHOU AND J. FENG, *Deep forest*, arXiv preprint arXiv:1702.08835, (2017).
- [476] D. ZHU, P. CUI, Z. ZHANG, J. PEI, AND W. ZHU, *High-order proximity preserved embedding for dynamic networks*, IEEE Transactions on Knowledge and Data Engineering, 30 (2018), pp. 2134–2144.
- [477] H. ZHU, *Snyder m*, Protein chip technology. Curr Opin Chem Biol, 7 (2003), pp. 55–63.
- [478] M.-F. ZHU, J. DONG, AND D.-S. CAO, *rdnase: R package for generating various numerical representation schemes of dna sequences*, 2016.
- [479] A. ZIRKEL AND A. PAPANTONIS, *Detecting circular rnas by rna fluorescence in situ hybridization*, in Circular RNAs, Springer, 2018, pp. 69–75.
- [480] H. ZOU AND T. HASTIE, *Regularization and variable selection via the elastic net*, Journal of the royal statistical society: series B (statistical methodology), 67 (2005), pp. 301–320.
- [481] J. ZOU, M. HUSS, A. ABID, P. MOHAMMADI, A. TORKAMANI, AND A. TELENTI, *A primer on deep learning in genomics*, Nature genetics, 51 (2019), pp. 12–18.
- [482] H. ZULFIQAR, Z.-J. SUN, Q.-L. HUANG, S.-S. YUAN, H. LV, F.-Y. DAO, H. LIN, AND Y.-W. LI, *Deep-4mcw2v: a sequence-based predictor to identify n4-methylcytosine sites in escherichia coli*, Methods, (2021).
- [483] Y. ZUO, Y. LI, Y. CHEN, G. LI, Z. YAN, AND L. YANG, *Psekraac: a flexible web server for generating pseudo k-tuple reduced amino acids composition*, Bioinformatics, 33 (2017), pp. 122–124.

APPENDIX

The proposed framework is developed on top of following APIs

Pandas	https://pandas.pydata.org
scikit-learn	https://scikit-learn.org
Scipy	https://scipy.org
Itertools	https://docs.python.org/3/library/itertools.html
Numpy	https://numpy.org/
mlxtend	http://rasbt.github.io/mlxtend/
Torch	https://pytorch.org/
Tensorflow	https://www.tensorflow.org/
Matplotlib	https://matplotlib.org/
Gensim	https://pypi.org/project/gensim/
Igraph	https://igraph.org/
Keras	https://keras.io/
Nltk	https://www.nltk.org/
Networkx	https://networkx.org/
Node2vec	https://snap.stanford.edu/node2vec/
math	https://docs.python.org/3/library/math.html

PERSONAL INFORMATION

Muhammad Nabeel Asim



📍 German Research Center for Artificial Intelligence

☎ +49152227786647

✉ Muhammad_Nabeel.Asim@dfki.de

🗣 [Google Scholar Profile](#)

CAREER SUMMARY



Earned PhD in Bioinformatic with Summa cum laude Distinction, Six years experience in utilizing and designing Artificial Intelligence methods to solve real world problems related to Bio-informatics, Natural Language Processing and Energy. During this research span published 7 Conference and 27 Journal papers with 150⁺ Journal impact factor and 480⁺ Citations.

WORK EXPERIENCE

April 2019 to present

German Research Center for Artificial Intelligence

Research Assistant

- Developed several DNA, RNA and protein sequence analysis applications:
 - **DNA:** Histone occupancy detector, Histone Acetylation and Methylation level predictor, DNA modification predictor, Enhancer identification and strength predictor
 - **RNA:** Non-coding RNA classification, RNA subcellular location prediction, Non-coding RNA interaction prediction
 - **Protein:** Protein-virus interaction prediction, Protein-Protein interaction prediction, Non-coding RNA protein interaction prediction
- Supervised master's Thesis and Seminars
- Wrote research proposals
- In collaboration with Sartorius developed antibody performance prediction application
- In collaboration with Daimler developed close domain question answering system

SARTORIUS



July 2018 to March 2019

Al-Khawarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore

Assistant Manager

- Wrote research proposals and won 21 million PKR funding
- Led diverse research projects related to Natural Language Processing and Energy Prediction
 - **NLP:** Police complaint filtering system, Criminal case log summarization application, Hate speech detector
 - **Energy:** Renewable energy forecasting framework, Energy generation and demand forecasting system



July 2017 to June 2018



Al-Khwarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore

Senior Research Associate

- Developed diverse types of Natural Language Processing applications: Biomedical information retrieval framework, Biomedical question answering system, Medical code and category assignment framework, Document image and text classification system, News categorization application, Social media brand monitoring application
- Supervised bachelors student's Final year projects and masters student's Thesis

January 2016 to jun 2017



Al-Khwarizmi Institute of Computer Science (KICS), University of Engineering and Technology (UET), Lahore

Research Associate

- Worked on Text classification, Information retrieval, Query expansion for biomedical document retrieval
- Designed novel feature selection methods, Developed multi-label classification frameworks for English and Urdu languages

September 2012 to October 2014



University of Management and Technology (UMT), Lahore

Teacher Assistant

- Performed students evaluation through quizzes and assignments
- Supervised semester projects

June 2013 to August 2013



MITCHELL'S FRUIT FARMS LIMITED

Internee Engineer

Monitored and Analyzed Sensors data

EDUCATION AND TRAINING

May 2019 to December 2022



Ph.D Computer Science (**BiInformatics**)

Technische Universität Kaiserslautern, Kaiserslautern, Germany

- **Thesis Topic:** An Efficient Automated Machine Learning Framework for Genomics and Proteomics Sequence Analysis
- Developed computational framework that contains all existing sequence encoding methods and traditional feature engineering methods as well as classifiers along with following novelties
 - Incorporated novel encoders competent in transforming raw DNA/RNA and protein sequences to numerical representations
 - Developed deep learning based predictors with explainable decisions
 - Utilizing proposed framework developed several web applications that are publically available at: https://sds_genetic_analysis.opendfki.de/

June 2015 to March 2017

M.Sc Electrical Engineering (Machine Learning)

University of Engineering and Technology (UET) , Lahore, Pakistan

Courses of interest: Machine Learning, Pattern Recognition, Engineering Statistics, Simulation Modeling and Analysis, Design and Analysis of Computer Algorithms



Thesis Topic: Selection of maximally relevant , minimally redundant features for classification applications

January 2010 to February 2014

B.Sc Electrical Engineering



University of Management and Technology, Lahore, Pakistan

Courses of interest: Computer Organization and Architecture, Digital Logic Design, Modern Microprocessors Systems, Programing Fundamentals, Object Oriented Programing, Data Structure, Signals and Systems, Digital Signal Processing

PUBLICATIONS



PATENT

1. **Muhammad Nabeel Asim**, Christoph Zehe, Olivier Cloarec, Johan Trygg, and Sheraz Ahmed. "METHOD, COMPUTER PROGRAM PRODUCT AND SYSTEM FOR OPTIMIZING PROTEIN EXPRESSION"; Application under review in European patent; MB&P Ref: S15547EU - hb / yma

REFERRED JOURNAL PUBLICATIONS

1. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Ahtisham Fazeel, Andreas Dengel, and Sheraz Ahmed. "DNA-MP: a generalized DNA modifications predictor for multiple species based on powerful sequence encoding method." *Briefings in Bioinformatics* (2022). DOI: [10.1093/bib/bbac546](https://doi.org/10.1093/bib/bbac546), Impact factor: 13.994
2. **Muhammad Nabeel Asim**, Ahtisham Fazeel, Muhammad Ali Ibrahim, Andreas Dengel, and Sheraz Ahmed. "MP-VHPPI: Meta predictor for viral host protein-protein interaction prediction in multiple hosts and viruses." *Frontiers in Medicine* 9 (2022) DOI: [10.3389/fmed.2022.1025887](https://doi.org/10.3389/fmed.2022.1025887), Impact factor: 5.058
3. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Imran Razzak, Andreas Dengel and Sheraz Ahmed. "Histone-Net: A Multi-Paradigm Computational Framework for Histone Occupancy and Modification Prediction", *Complex & Intelligent Systems* pp.1-21 (2022) DOI: [10.1007/s40747-022-00802-w](https://doi.org/10.1007/s40747-022-00802-w), Impact factor: 6.7
4. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "EL-RMLocNet: An Explainable LSTM Network for RNA-Associated Multi-Compartment Localization Prediction", *Computational and Structural Biotechnology Journal* (2022) DOI: [10.1016/j.csbj.2022.07.031](https://doi.org/10.1016/j.csbj.2022.07.031), Impact factor: 6.155
5. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "Circ-LocNet: A Computational Framework for Circular RNA Sub-Cellular Localization Prediction", *International Journal of Molecular Sciences* 23(15), p.8221 (2022) DOI: [10.3390/ijms2315822](https://doi.org/10.3390/ijms2315822), Impact factor: 6.208
6. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "CONR-NET: A Collection of Neural Refinements for Protein Protein Interaction Prediction", *iScience* (2022), Impact factor: 6.107
7. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Christoph Zehe, Johan Trygg, Andreas Dengel and Sheraz Ahmed. "BoT-Net: A Lightweight Bag of Tricks based Neural Network for Efficient lncRNA-miRNA Interaction Prediction", *Interdisciplinary Sciences: Computational Life Sciences* pp.1-22 (2022) DOI: [10.1007/s12539-022-00535-x](https://doi.org/10.1007/s12539-022-00535-x), Impact factor: 3.492
8. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "LGCA-VHPPI: A Local-Global Residue Context Aware Viral-Host Protein-Protein Interaction Predictor", *PLOS ONE* 17(7), p.e0270275 (2022) DOI: [10.1371/journal.pone.0270275](https://doi.org/10.1371/journal.pone.0270275), Impact factor: 3.75





9. Faiza Mehmood, Muhammad Usman Ghani, Hina Ghafoor, Rehab Shahzadi, **Muhammad Nabeel Asim** and Waqar Mahmood. "EGD-SNet: A computational search engine for predicting an end-to-end machine learning pipeline for Energy Generation & Demand Forecasting." *Applied Energy* 324 (2022): 119754. DOI: [10.1016/j.apenergy.2022.119754](https://doi.org/10.1016/j.apenergy.2022.119754), Impact factor: 11.446
10. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "Advances in Computational Methodologies for Classification and Sub-Cellular Locality Prediction of Non-Coding RNAs", *International Journal of Molecular Sciences* 22(16), (2021) p.8719 DOI: [10.3390/ijms22168719](https://doi.org/10.3390/ijms22168719), Impact factor: 6.208
11. Marco Stricker, **Muhammad Nabeel Asim**, Andreas Dengel and Sheraz Ahmed. "CircNet: an encoder–decoder-based convolution neural network (CNN) for circular RNA identification", *Neural Computing and Applications* pp.1-12 (2021) DOI: [10.1007/s00521-020-05673-1](https://doi.org/10.1007/s00521-020-05673-1), Impact factor: 5.102
12. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Christoph Zehe, Johan Trygg, Andreas Dengel and Sheraz Ahmed. "MirLocPredictor: A ConvNet-Based Multi-Label MicroRNA Subcellular Localization Predictor by Incorporating k-Mer Positional Information", *Genes* 11(12), (2020) p.147 DOI: [10.3390/genes11121475](https://doi.org/10.3390/genes11121475), Impact factor: 4.141
13. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Christoph Zehe, Johan Trygg, Andreas Dengel and Sheraz Ahmed. "A Robust and Precise ConvNet for small non-coding RNA classification (RPC-snRC)." , *IEEE Access* 9 (2020): pp. 19379-19390 DOI: [10.1109/ACCESS.2020.3037642](https://doi.org/10.1109/ACCESS.2020.3037642), Impact factor: 3.476
14. Faiza Mehmood, Muhammad Usman Ghani, **Muhammad Nabeel Asim**, Rehab Shahzadi, Aamir Mehmood and Waqar Mahmood. "MPF-Net: A computational multi-regional solar power forecasting framework." *Renewable and Sustainable Energy Reviews* 151 (2021): 111559. DOI: [10.1016/j.rser.2021.111559](https://doi.org/10.1016/j.rser.2021.111559), Impact factor: 16.779
15. Muhammad Ali Ibrahim, Muhammad Usman Ghani Khan, Faiza Mehmood, **Muhammad Nabeel Asim**, and Waqar Mahmood. "GHS-NET a generic hybridized shallow neural network for multi-label biomedical text classification." *Journal of biomedical informatics* 116 (2021): 103699. DOI: [10.1016/j.jbi.2021.103699](https://doi.org/10.1016/j.jbi.2021.103699), Impact factor: 8.00
16. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Usman Ghani Khan, Waqar Mahmood, Andreas Dengel and Sheraz Ahmed. "Benchmarking performance of machine and deep learning-based methodologies for Urdu text document classification", *Neural Computing and Applications* 33(11), (2021) pp.5437-5469 DOI: [10.1007/s00521-020-05321-8](https://doi.org/10.1007/s00521-020-05321-8), Impact factor: 5.102
17. Faiza Mehmood, M. U. Ghani, M. A. Ibrahim, R. Shahzadi, W. Mahmood and **Muhammad Nabeel Asim**. "A Precisely Xtreme-Multi Channel Hybrid Approach for Roman Urdu Sentiment Analysis," in *IEEE Access*, vol. 8, pp. 192740-192759, 2020. DOI: [10.1109/ACCESS.2020.3030885](https://doi.org/10.1109/ACCESS.2020.3030885), Impact factor: 3.476
18. Muhammad Wasim, **Muhammad Nabeel Asim** and Usman Ghani Khan. "Multilabel Biomedical Question Classification for Lexical Answer Type Prediction", *Journal of biomedical informatics* (2019): 103143. DOI: [10.1016/j.jbi.2019.103143](https://doi.org/10.1016/j.jbi.2019.103143), Impact factor: 8.00
19. **Muhammad Nabeel Asim**, Muhammad Wasim and Usman Ghani Khan. "A Survey of Ontology Learning Techniques and Applications", *DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION*. DOI: [10.1093/database/bay101](https://doi.org/10.1093/database/bay101), Impact factor: 4.462
20. Muhammad Wasim, Waqar Mahmood, **Muhammad Nabeel Asim** and Usman Ghani Khan "Multi-Label Question Classification for Factoid and List Type Questions in Biomedical Question Answering", *IEEE Access* 7 (2019): 3882-3896. Impact DOI: [10.1109/ACCESS.2018.2887165](https://doi.org/10.1109/ACCESS.2018.2887165), Impact factor: 3.476
21. **Muhammad Nabeel Asim**, Muhammad Wasim, Usman Ghani Khan and Waqar Mahmood "Improved Biomedical Term Selection in Pseudo Relevance Feedback", *DATABASE-THE JOURNAL OF BIOLOGICAL DATABASES AND CURATION* vol(2018). DOI: [10.1093/database/bay056](https://doi.org/10.1093/database/bay056), Impact factor: 4.462



ELSEVIER



Springer



IJCSNS



THE SCIENCE AND INFORMATION ORGANIZATION



IJCSNS

22. **Muhammad Nabeel Asim**, Muhammad Wasim, Usman Ghani Khan and Waqar Mahmood. "The Use of Ontology in Retrieval: A Study on Textual, Multilingual, and Multimedia Retrieval", IEEE Access 7 (2019): 21662-21686. DOI: [10.1109/ACCESS.2019.2897849](https://doi.org/10.1109/ACCESS.2019.2897849), Impact factor: 3.476
23. Abdur Rehman, Kashif Javaid, Haroon Babri and **Muhammad Nabeel Asim**, "Selection of the Most Relevant Terms Based on a Max-Min Ratio metric for Text Classification" 2017. accepted in Expert Systems With Applications. DOI: [10.1016/j.eswa.2018.07.028](https://doi.org/10.1016/j.eswa.2018.07.028), Impact factor: 8.665
24. Muhammad Wasim, **Muhammad Nabeel Asim**, Usman Ghani Khan, Zahoor ur Rehman and Seungmin Rho. "Lexical Paraphrasing and Pseudo Relevance Feedback for Biomedical Document Retrieval" Multimedia Tools and Applications (2018) 1-32. DOI: [10.1007/s11042-018-6060-z](https://doi.org/10.1007/s11042-018-6060-z), Impact factor: 2.577
25. **Muhammad Nabeel Asim**, Abdur Rehman and Muhammad Idrees. "Effect of Pruning on Feature Ranking metrics in Highly Skewed Datasets in Text Classification" International Journal of Computer Science and Network Security(IJCSNS), 17(10),2017 DOI: [201710/20171018](https://doi.org/201710/20171018)
26. **Muhammad Nabeel Asim**, Abdur Rehman and Umar Shoaib. "Accuracy Based Feature Ranking Metric for Multi-Label Text Classification" International Journal of Advanced Computer Science and Applications (IJACSA), 8(10), 2017. DOI: [10.14569/IJACSA.2017.081048](https://doi.org/10.14569/IJACSA.2017.081048), Impact factor: 1.092
27. **Muhammad Nabeel Asim**, Muhammad Salman Khalid, Muhammad Idrees and Abdur Rehman. "Efficient Utilization of Cryptographic Resources in Embedded Computing Systems" International Journal of Computer Science and Network Security (IJCSNS),17(10), 2017 DOI: [201710/20171011](https://doi.org/201710/20171011)

CONFERENCE PUBLICATIONS





1. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Christoph Zehe., Cloarec, O Sjogren, R Johan Trygg, Andreas Dengel and Sheraz Ahmed. "L2S-MirLoc: A Lightweight Two Stage MiRNA Sub-Cellular Localization Prediction Framework", *International Joint Conference on Neural Networks, (IJCNN)* (pp. 1-8) (2021). IEEE DOI: [10.1109/IJCNN52387.2021.9534015](https://doi.org/10.1109/IJCNN52387.2021.9534015)
2. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "ChrSLoc-Net: Machine Learning-Based Prediction of Channelrhodopsins Proteins within Plasma Membrane", *IEEE EMBS International Conference on Biomedical and Health Informatics, (BHI)* (pp. 1-4) (2021). IEEE DOI: [10.1109/BHI50953.2021.9508615](https://doi.org/10.1109/BHI50953.2021.9508615)
3. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "K-mer Neural Embedding Performance Analysis Using Amino Acid Codons", *International Joint Conference on Neural Networks, (IJCNN)* (pp. 1-8). (2020) IEEE DOI: [10.1109/IJCNN48605.2020.9206892](https://doi.org/10.1109/IJCNN48605.2020.9206892)
4. **Muhammad Nabeel Asim**, Muhammad Ali Ibrahim, Muhammad Imran Malik, Andreas Dengel and Sheraz Ahmed. "Enhancer-DSNet: A Supervisedly Prepared Enriched Sequence Representation for the Identification of Enhancers and Their Strength.", *27th International Conference on Neural Information Processing, (ICONIP-2020)* (pp. 38-48). Springer, Cham DOI: [10.1007/978-3-030-63836-8_4](https://doi.org/10.1007/978-3-030-63836-8_4)
5. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Muhammad Usman Ghani Khan, Andreas Dengel and Sheraz Ahmed. "A robust hybrid approach for textual document classification.", *International conference on document analysis and recognition (ICDAR)* (2019) (pp. 1390-1396). IEEE DOI: [10.1109/ICDAR.2019.00224](https://doi.org/10.1109/ICDAR.2019.00224)
6. **Muhammad Nabeel Asim**, Muhammad Imran Malik, Muhammad Usman Ghani Khan, Andreas Dengel and Sheraz Ahmed. "Two stream deep network for document image classification", *International conference on document analysis and recognition (ICDAR)* (2019) (pp. 1410-1416). IEEE DOI: [10.1109/ICDAR.2019.00227](https://doi.org/10.1109/ICDAR.2019.00227)
7. **Muhammad Nabeel Asim**, Muhammad Wasim, Sajid Ali and Abdur Rehman. "Comparison of Feature Selection Methods in Text Classification on highly skewed Datasets", *Electrical Engineering and Computing Technologies (INTELLECT)*, 2017 First International Conference on Latest trends in. IEEE, 2017. DOI: [10.1109/INTELLECT.2017.8277634](https://doi.org/10.1109/INTELLECT.2017.8277634)

SKILLS AND INTERESTS




IT SKILLS

	Python		MATLAB & Simulink
	Tensorflow & Keras		LATEX
	Pytorch		Ubuntu
	C/C++		Windows

LANGUAGES

	Punjabi	Native Speaker
	URDU	Native Speaker
	English	Fluent Speaker

INTERESTS

	Literature & History
	Football & Rugby
	Music & Movies

ACHIEVEMENTS & AWARDS

	Ph.D Summa cum laude Distinction	December 2022
	Technische Universität Kaiserslautern, Kaiserslautern, Germany	
	Ph.D Scholarship	2019 to 2022
	Technische Universität Kaiserslautern, Kaiserslautern, Germany	
	BS Scholarship	2010 to 2014
	University of Management and Technology, Lahore, Pakistan	
	Best Masters Thesis Award	2017
	Department of Electrical engineering, UET, Lahore	
	Best Researcher Award	2018
	Al-Khawarizmi Institute of Computer Science (KICS), UET, Lahore	
	UMT Sports Department Vice President	2011 to 2014
	UMT Sports based Scholarships Candidates Skills Evaluator	2011 to 2014